Research Article

Wolfgang Karl Härdle and Elena Silyakova*

Implied basket correlation dynamics

DOI: 10.1515/strm-2014-1176
Received December 25, 2014; revised June 14, 2016; accepted June 30, 2016

Abstract: Equity basket correlation can be estimated both using the physical measure from stock prices, and also using the risk neutral measure from option prices. The difference between the two estimates motivates a so-called "dispersion strategy". We study the performance of this strategy on the German market and propose several profitability improvement schemes based on implied correlation (IC) forecasts. Modelling IC conceals several challenges. Firstly the number of correlation coefficients would grow with the size of the basket. Secondly, IC is not constant over maturities and strikes. Finally, IC changes over time. We reduce the dimensionality of the problem by assuming equicorrelation. The IC surface (ICS) is then approximated from the implied volatilities of stocks and the implied volatility of the basket. To analyze the dynamics of the ICS we employ a dynamic semiparametric factor model.

Keywords: Correlation risk, dimension reduction, dispersion strategy, dynamic factor models

MSC 2010: 62H25, 62H15, 62H20

1 Introduction

Equity basket correlation is an important risk factor. It characterizes the strength of linear dependence between assets and thus measures the degree of portfolio diversification. It is an input for many pricing models, plays a key role in portfolio optimization and risk management. The concept of a time-varying correlation is frequently used in studies that describe the joint dynamics of assets, see [8, 21]. However, the idea of considering the correlation as an asset, on its own, is relatively new and has recently gained popularity together with the emergence of such derivative instruments as variance, volatility, correlation swaps and trading strategies with them, see [9, 19]. In this context being able to predict correlation patterns might help to reveal profitable trading opportunities. One of the most common ways of obtaining a correlation exposure is to replicate it with variance swaps. In this paper we study the behavior of a particular vehicle for trading correlation known as a "dispersion strategy", in which one sells a stock index volatility and buys individual volatilities, see [1]. We propose several ways of improving the profitability of the strategy by extracting information from a dynamic model of implied correlation.

Unlike asset prices, correlations are not directly observed in the market and need to be estimated in the context of a particular model. Obtaining a well-conditioned and invertible estimate of an empirical correlation matrix is often a complicated task, in particular when the dimensionality of basket elements N is higher than the time series length T. Here some work has been done in the field of random matrix theory (RMT), in which the case "large N, small T" is studied in an asymptotic setting, see [2, 31, 36]. A further segment of research has moved in the direction of developing various regularization methods for sample covariance and correlation matrices, such as a shrinkage technique proposed in [32], regularization via thresholding in [5], bending in [6], factor models in [23] and many others. There are some studies that propose a dynamic

Wolfgang Karl Härdle: Ladislaus von Bortkiewicz Chair of Statistics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany; and Sim Kee Boon Institute for Financial Economics, Singapore Management University, Administration Building, 81 Victoria Street, 188065 Singapore, e-mail: haerdle@hu-berlin.de
***Corresponding author: Elena Silyakova:** Ladislaus von Bortkiewicz Chair of Statistics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany, e-mail: silyakova@gmail.com

model for return correlation such as a DCC model (see [21]) and in a high-dimensional setting (see [22]). The common feature of all these studies is that the empirical correlation matrix is estimated under the physical measure from the time series of asset returns. Alternatively, instead of relying on historical data, one can infer correlation from the current snapshot of the option market. Option prices reflect the expectations of market participants about the future price (volatility) and disclose their perceptions of market risk, see [3, 12]. Some recent studies have shown that the implied volatility (IV), that equates the model option price and the one taken from the market, contains incremental information beyond the historical estimate and outperforms it in forecasting future volatility, see [7, 16, 25]. Yet only a few papers have studied the predictive content of the correlation, implied by option prices. Some work has been done for foreign exchange (FX) options (see [13, 33]), which showed that correlations implied from FX options are useful for forecasting future currency correlations. Skintzi and Refenes [38] investigated the average correlation implied by equity options and introduced the Implied Correlation index (ICX). They showed that ICX, computed from current option prices, is a useful proxy for the future realized correlation. Driessen et al. [20] investigated the power of options-implied correlation to explain the future realized correlation and conclude that its predictive power is quite high.

Here we model the implied correlation (IC), which is an object of very high dimensionality. Similarly to the IV, every day one recovers an IC surface. We model the IC with a dynamic semiparametric factor model (DSFM) (see [24, 35, 39]) and find that it yields a low-dimensional representation as a linear combination of a small number of time-invariant basis functions (surfaces), whose time evolution is driven by a series of coefficients; technical aspects are also described in [40]. We produce an IC forecast and use it in several hedging schemes for a dispersion strategy. For the empirical analysis we chose the German market represented by the DAX portfolio over the 2-years sample period from 20100802 to 20120801 (dates are written as YYYYMMDD). Backtesting shows that the hedge allows to the reduction of potential losses and increases the average profitability of the strategy.

The paper is structured as follows. In Section 2 we introduce the notions of realized, model-implied and model-free-implied volatility and correlation and describe the basic setup of a dispersion strategy with variance swaps. The DSFM model for IC is introduced in Section 3 starting with general description in Section 3.1, followed by the description of the functional principal component analysis (FPCA) approach to find the basis functions in Section 3.2 and the estimation procedure for both factors and factor loadings in Section 3.3. Section 4 presents the dataset taken for the empirical study, followed by a description of the estimation results in Section 5. First, in Section 5.1 we interpret the obtained factors and factor loadings and propose a time series model for low-dimensional factors. Finally, in Section 5.2 we propose and compare alternative dispersion strategy setups: a no hedge, a naïve approach and an advanced hedge. Section 6 concludes.

2 Correlation trading

2.1 Average basket correlation

In a basket of N assets, correlation $\rho_{i,j}$ measures linear dependence between the i-th and the j-th asset return, $i, j \in \{1, \dots, N\}$. Standard statistical analysis yields that the basket variance σ_B^2 can be decomposed as

$$\sigma_B^2 = \sum_i w_i^2 \sigma_i^2 + \sum_i \sum_{j \neq i} w_i w_j \sigma_i \sigma_j \rho_{ij}, \tag{2.1}$$

where σ_i^2 denotes the variance of the i-th asset return and w_i its weight in the basket. Now, assuming that ρ_{ij} is constant for every pair (i, j), one can imply the equicorrelation ρ from (2.1):

$$\rho = \frac{\sigma_B^2 - \sum_i w_i^2 \sigma_i^2}{\sum_i \sum_{j \neq i} w_i w_j \sigma_i \sigma_j}. \tag{2.2}$$

Later we call ρ a basket correlation or simply a correlation. The corresponding correlation matrix has all the off-diagonal elements equal to ρ and thus offers several advantages. Firstly, plugging $\rho_{i,j} = \rho$ into (2.1)

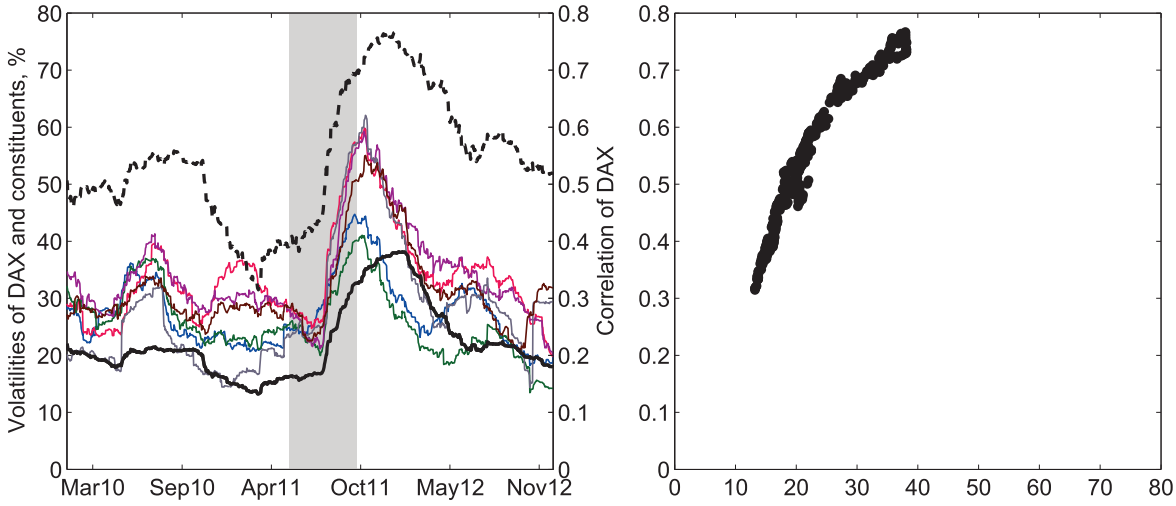


Figure 1. Left panel: DAX correlation (2.2) (dashed), DAX volatility (2.8) (solid black), volatility of DAX constituents Adidas, BMW, Siemens, Daimler, E.ON, Lufthansa volatilities (2.8) (color), the stock market fall 2011 (shaded area). Right panel: Scatter plot DAX volatility vs. correlation. Estimation period: from 20100104 till 20121228; estimation window: 3 months.

reproduces the basket variance σ_B^2. Secondly, if $-\frac{1}{N-1} < \rho < 1$ then the correlation matrix is positive semi-definite, see [29]. This property becomes particularly important if N is large. A closer look also reveals that (2.2) is in fact a nonlinear weighted average over all $\rho_{i,j}$ in the basket:

$$\rho = \sum_i \sum_{j \neq i} c_{i,j} \rho_{i,j}$$

with weights $c_{i,j}$ defined by

$$c_{i,j} = \frac{w_i w_j \sigma_i \sigma_j}{\sum_i \sum_{j \neq i} w_i w_j \sigma_i \sigma_j}.$$

Bourgoin [10] showed that if a correlation matrix is positive semi-definite, for sufficiently large baskets it holds that $0 \leq \rho \leq 1$. Using this property, minimum and maximum variances of a basket, $\sigma^2_{B,\min}$ and $\sigma^2_{B,\max}$ respectively, are defined as follows:

$$\sigma^2_{B,\min} = \sum_i w_i^2 \sigma_i^2 \quad \text{and} \quad \sigma^2_{B,\max} = \sum_i w_i^2 \sigma_i^2 + \sum_i \sum_{j \neq i} w_i w_j \sigma_i \sigma_j. \tag{2.3}$$

The minimum variance $\sigma^2_{B,\min}$ is achieved when $\rho = 0$ that is when the assets in a basket are fully diversified. In the case of no diversification, one observes the maximal possible basket variance $\sigma^2_{B,\max}$ corresponding to $\rho = 1$.

Further we can rewrite ρ by substituting (2.3) to (2.2):

$$\rho = \frac{\sigma_B^2 - \sigma^2_{B,\min}}{\sigma^2_{B,\max} - \sigma^2_{B,\min}} \tag{2.4}$$

and obtain an additional interpretation as a measure for the degree of diversification, see [38]. In fact, (2.4) shows how far σ_B^2 is from its minimal value $\sigma^2_{B,\min}$ relative to the possible value range $\sigma^2_{B,\max} - \sigma^2_{B,\min}$, or in other words, how far the basket is from the perfect diversification. A high value of ρ is the sign of a poorly diversified portfolio, which is typical for the market downturn, when asset prices simultaneously drop driving σ_B^2 up. It means diversification benefits disappear in times when they are needed most. To hedge against correlation risk investors look for derivative securities that offer higher payoffs (premia) when the correlation decreases.

If a basket is constructed from the constituents of an equity index with weights equal to index weights, then the corresponding basket correlation would serve as a benchmark for a sector, an industry or a whole

market average correlation. Figure 1 shows an example of the DAX correlation together with the volatility of DAX and some of its components. Firstly, we see that the correlation and the volatility vary over time. Secondly, the volatility of the basket (DAX) is smaller than almost any individual volatility of its constituents, which illustrates the impact of the diversification effect on the portfolio risk. Finally, there is a clear linear dependence of the correlation of the basket and its volatility. However the strength of this dependence changes when the volatility exceeds a certain threshold. We investigate this phenomenon and propose a dataset correction scheme in Section 4.

2.2 Implied versus realized correlation

Based on (2.2) we conclude that the exposure to the basket correlation ρ can be achieved by exposures to the variances of a basket σ_B^2 and its constituents, σ_i^2. Such trades can be realized via a combination of variance swaps. A variance swap is an over-the-counter contract opened at t, which at $t+\tau$ pays the difference between the variance cumulated over the life time of the swap $\sigma_{t+\tau}^2$ and the fixed pre-defined strike $\tilde{\sigma}_t^2(\tau)$:

$$\{\sigma_{t+\tau}^2 - \tilde{\sigma}_t^2(\tau)\}N_{\text{var}}, \tag{2.5}$$

where N_{var} is the notional amount. Here and later t and τ are given in fractions of a year.

The strike of a variance swap is the risk-neutral expectation at t of the integrated variance from t to $t+\tau$. It is also known as the model-free-implied variance (MFIV), where "model-free" indicates that the expectation does not depend on the specification of the underlying price process, see [12]. MFIV can be approximated by a function of current option prices (see [11, 12, 14]) which has the following form:

$$\tilde{\sigma}_t^2(\tau) = \mathsf{E}_t^Q\left[\int_t^{t+\tau} \sigma^2(s)ds\right] = \frac{2e^{r\tau}}{\tau}\left\{\int_0^{S_t} \frac{P_t(K,\tau)dK}{K^2} + \int_{S_t}^{\infty} \frac{C_t(K,\tau)dK}{K^2}\right\}, \tag{2.6}$$

where E_t^Q is the expected value at t under the risk-neutral measure Q, $P_t(K,\tau)$ $\{C_t(K,\tau)\}$ is the price at t of put {call} with exercise price K and time to maturity τ, S_t is the price of the asset in t, and r is the annualized continuously compounded risk-free interest rate.

MFIV can be opposed to the implied variance $\hat{\sigma}_t^2(\kappa,\tau)$, the square of the implied volatility (IV), which is obtained by solving

$$V_t(\hat{\sigma},\kappa,\tau) - \check{V}_t(\kappa,\tau) = 0, \tag{2.7}$$

where V_t is the theoretical (model) option price, $\check{V}_t$ is the option price taken from the market, and $\kappa = K/(S_t e^{r\tau})$ is the moneyness of the option. IV, in comparison to MFIV, is a function of both κ and τ, meaning that at every t one recovers a cloud of points, which can be approximated by a surface, see [17, 24].

The floating leg of the variance swap, the realized variance (RV) of an asset from t to $t+\tau$, can be computed from the time series of daily asset returns in different ways, depending on the contract specification. Here we use the most common form

$$\sigma_{t+\tau}^2 = \tau^{-1}\sum_{i=252t}^{252(t+\tau)}\left(\log\frac{S_i}{S_{i-1}}\right)^2. \tag{2.8}$$

In [15], $\sigma_{t+\tau}^2 - \tilde{\sigma}_t^2(\tau)$ is referred to as the variance risk premium (VRP), which is shown to be strongly negative for major US stock indexes over the sample period from January 1996 to December 2003. The negative sign indicates that investors are willing to pay extra to hedge themselves against possible future market turmoil. Bakshi et al. [4], who investigated the S&P100 index and its largest constituents from 1991 to 1995, also found significant negative difference between realized and option-implied volatilities for the average of 25 stocks and stressed that this difference is less pronounced than for the index. Driessen et al. [20] studied each S&P100 constituent individually. Their t-test for H_0, that the sample means of RV and MFIV are equal, was not rejected for the majority of stocks in the sample from January 1996 to December 2003.

We check the same hypothesis on the German market for the sample period from 20100104 to 20121228 using the dataset described in Section 4. Table 1 summarizes the results of a t-test for the null hypothesis that RV and MFIV are on average equal against the alternative RV < MFIV. H_0 is strongly rejected for the DAX index. For the DAX constituents the rejection rate decreases with the options' maturity τ: with $\tau = 0.25$ (3 months) and $\tau = 0.5$ (6 months) H_0 cannot be rejected at a 5% significance level for 8 out of 30 DAX constituents, with $\tau = 1$ (1 year) for 13 constituents. Table 1 reports the t-test results for these 13 stocks. In addition, Table 2 reports sample averages of RV and MFIV and their differences for all 30 DAX constituents. The latter are found to be negative for most of the stocks and for the DAX index.

Driessen et al. [20] interpreted their t-test results as indirect evidence that a negative correlation risk premium (CRP) exists. To identify the existence of CRP in the DAX dataset we compute the model-free-implied correlation (MFIC) $\tilde{\rho}_t(\tau)$ from the MFIVs of DAX and its constituents and the realized correlation (RC) $\rho_{t+\tau}$ from the corresponding RV by applying (2.2):

$$\tilde{\rho}_t(\tau) = \frac{\tilde{\sigma}^2_{t,\mathrm{DAX}}(\tau) - \sum_i w_i^2 \tilde{\sigma}^2_{t,i}(\tau)}{\sum_i \sum_{j\neq i} w_i w_j \tilde{\sigma}_{t,i}(\tau)\tilde{\sigma}_{t,j}(\tau)}, \tag{2.9}$$

$$\rho_{t+\tau} = \frac{\sigma^2_{t+\tau,\mathrm{DAX}} - \sum_i w_i^2 \sigma^2_{t+\tau,i}}{\sum_i \sum_{j\neq i} w_i w_j \sigma_{t+\tau,i}\sigma_{t+\tau,j}}. \tag{2.10}$$

Figure 2 plots the MFIC and the RC of DAX computed over the 3-month window and with 3 month maturity respectively ($\tau = 0.25$). The hypothesis H_0: RC = MFIC of the t-test is strongly rejected. Using this finding and taking into account results in the literature, we would expect $\rho_{t+\tau} - \tilde{\rho}_t(\tau)$ (CRP) to be negative most of the time. One of the ways of exploiting this observation is to make a bet on the market correlation by entering a dispersion strategy.

2.3 Dispersion strategy with variance swaps

We study one of the variations of the dispersion strategy, which consists of selling the variance of the basket (DAX) and buying variances of basket constituents.

The dispersion strategy can be implemented by taking a short position in the variance swap (2.5) on an index and long positions in variance swaps on its constituents with notional amounts proportional to index weights. The payoff of a dispersion strategy at $t + \tau$ is then defined by

$$D_{t+\tau} = -\{\sigma^2_{t+\tau,B} - \tilde{\sigma}^2_{t,B}(\tau)\} + \sum_{i=1}^{N} w_i^2 \{\sigma^2_{t+\tau,i} - \tilde{\sigma}^2_{t,i}(\tau)\}. \tag{2.11}$$

Then we apply (2.2) and rewrite (2.11) in the following form:

$$D_{t+\tau} = \tilde{\rho}_t(\tau) \sum_i \sum_{j\neq i} w_i w_j \tilde{\sigma}_{t,i}(\tau)\tilde{\sigma}_{t,j}(\tau) - \rho_{t+\tau} \sum_i \sum_{j\neq i} w_i w_j \sigma_{t+\tau,i}\sigma_{t+\tau,j}. \tag{2.12}$$

Based on empirical findings described in Section 2.2 we assume $\tilde{\sigma}_{t,i}(\tau) \approx \sigma_{t+\tau,i}$ for each constituent stock and simplify the payoff (2.12) as follows:

$$D_{t+\tau} \approx \sum_i \sum_{j\neq i} w_i w_j \tilde{\sigma}_{t,i}(\tau)\tilde{\sigma}_{t,j}(\tau)\{\tilde{\rho}_t(\tau) - \rho_{t+\tau}\},$$

which illustrates that by entering the dispersion strategy one obtains exposure to $\rho_{t+\tau} - \tilde{\rho}_t(\tau)$, where the floating leg $\rho_{t+\tau}$ is computed with (2.8) and (2.2) at expiry, and the fixed leg $\tilde{\rho}_t(\tau)$ is a function of variance swap strikes (2.6). The test results described in Section 2.2 suggest that we should, on average, expect $\rho_{t+\tau} - \tilde{\rho}_t(\tau) < 0$. It also means the dispersion strategy with payoff $D_{t+\tau}$ on average would have a profit. However, as one can see in Figure 2, there might be days when $\rho_{t+\tau} - \tilde{\rho}_t(\tau) \geq 0$. In order to hedge against these potential losses one needs a forecast of the floating leg of the dispersion strategy.

Another possible modification of the dispersion trading strategy does not involve trading on the OTC market and can be implemented with standardized market instruments, puts and calls. The strategy consists

	$\tau = 0.25$	$\tau = 0.5$	$\tau = 1$
BASF	0.00000	0.00000	0.15298∗
Commerzbank	0.00000	0.00005	0.66073∗
Continental AG	0.99998∗	0.99998∗	0.99998∗
Deutsche Bank	0.00001	0.00000	0.06985∗
Deutsche Börse	0.95674∗	0.95064∗	0.96357∗
Fresenius Medical Care	0.45640∗	0.27716∗	0.81540∗
Henkel	0.90404∗	0.99997∗	0.99680∗
K+S	0.00000	0.00000	0.99981∗
Lanxess	0.27625∗	0.05776∗	0.99989∗
Linde	0.76162∗	0.87214∗	1.00000∗
RWE	0.40725∗	0.21673∗	0.05305∗
ThyssenKrupp	0.01733	0.00272	0.69073∗
Volkswagen	0.99998∗	0.99998∗	0.99998∗

Table 1. The results of t-test for the equality of sample averages with H_0: RV = MFIV against the alternative RV < MFIV of DAX constituents for which H_0 is not rejected at 5% significance level at least for one τ. Test results which cannot reject H_0 are marked with $*$. The test was performed for the volatilities of the DAX index and its 30 constituent stocks computed over the time period 20100104–20121228 for three different maturities/estimation windows: $\tau = 0.25, 0.5, 1$. The test results are presented for the subsample of 13 DAX constituents, for which H_0 cannot be rejected at least for one τ.

	$\tau = 0.25$			$\tau = 0.5$			$\tau = 1$		
	σ	$\tilde{\sigma}$	$\sigma - \tilde{\sigma}$	σ	$\tilde{\sigma}$	$\sigma - \tilde{\sigma}$	σ	$\tilde{\sigma}$	$\sigma - \tilde{\sigma}$
Adidas	27.56	30.06	−2.50	28.35	31.07	−2.71	29.72	31.45	−1.73
Allianz	29.38	31.93	−2.55	30.44	33.03	−2.58	32.72	33.94	−1.22
BASF	28.95	31.08	−2.13	29.73	31.62	−1.89	31.32	31.58	−0.26
Bayer	27.39	30.74	−3.35	27.93	31.23	−3.31	28.85	31.17	−2.31
Beiersdorf	33.90	37.09	−3.20	34.53	37.96	−3.43	35.95	37.94	−1.99
BMW	19.57	23.95	−4.38	19.79	24.17	−4.38	20.26	23.90	−3.64
Commerzbank	46.47	52.77	−6.30	47.40	52.54	−5.14	51.02	51.70	−0.68
Continental AG	41.45	39.82	1.63∗	43.84	40.55	3.29∗	48.28	41.60	6.69∗
Daimler	34.18	37.55	−3.36	34.93	38.45	−3.52	36.93	38.85	−1.93
Deutsche Bank	39.26	43.56	−4.30	39.76	43.93	−4.18	42.52	43.51	−1.00
Deutsche Börse	30.03	30.40	−0.37	31.09	31.10	0.00∗	32.90	31.48	1.43∗
Deutsche Post	31.61	33.63	−2.02	32.05	34.18	−2.13	33.13	34.67	−1.54
Deutsche Telekom	25.50	27.54	−2.04	26.20	28.26	−2.06	27.66	29.12	−1.46
E.ON	23.62	26.26	−2.64	24.11	26.53	−2.41	24.67	27.48	−2.81
Fresenius Medical Care	28.12	29.21	−1.09	28.67	29.68	−1.01	30.19	30.04	0.16∗
Fresenius SE	19.01	22.44	−3.43	19.21	23.17	−3.96	19.89	23.62	−3.73
HeidelbergCement	21.16	23.56	−2.40	21.61	23.69	−2.08	22.77	24.08	−1.31
Henkel	39.04	39.00	0.05∗	40.96	39.49	1.46∗	45.03	39.99	5.04∗
Infineon	23.15	26.08	−2.94	23.57	26.70	−3.13	24.30	27.05	−2.75
K+S	40.78	43.75	−2.98	41.75	45.03	−3.27	45.88	45.08	0.80∗
Lanxess	29.99	30.84	−0.85	30.86	31.78	−0.92	33.51	32.52	1.00∗
Linde	39.90	39.92	−0.02	41.02	40.55	0.47∗	42.83	41.07	1.76∗
Lufthansa	22.18	25.18	−2.99	22.72	26.32	−3.61	23.47	26.77	−3.30
Merck	23.83	26.06	−2.23	24.43	26.40	−1.97	25.51	26.39	−0.89
Munich Re	23.35	26.54	−3.19	23.88	27.91	−4.03	25.28	29.10	−3.82
RWE	27.43	29.15	−1.72	27.93	29.71	−1.78	28.75	30.38	−1.63
SAP	21.45	23.67	−2.22	21.73	25.14	−3.42	22.00	26.82	−4.82
Siemens	25.94	28.53	−2.59	26.82	29.78	−2.95	28.63	30.35	−1.72
ThyssenKrupp	36.32	38.37	−2.04	36.78	38.71	−1.93	38.80	38.90	−0.10
Volkswagen	37.91	36.31	1.61∗	39.48	36.90	2.59∗	41.94	36.91	5.03∗
DAX Index	21.72	25.38	−3.66	22.30	26.67	−4.37	23.40	27.81	−4.41

Table 2. Mean of $\sqrt{\text{RV}}$ (σ) and $\sqrt{\text{MFIV}}$ ($\tilde{\sigma}$) and their difference $\sqrt{\text{RV}} - \sqrt{\text{MFIV}}$ ($\sigma - \tilde{\sigma}$), for DAX index and its 30 constituent stocks computed over the time period 20100104–20121228 for three different maturities/estimation windows: $\tau = 0.25, 0.5, 1$. $\sigma - \tilde{\sigma} \geq 0$ are marked with $*$.

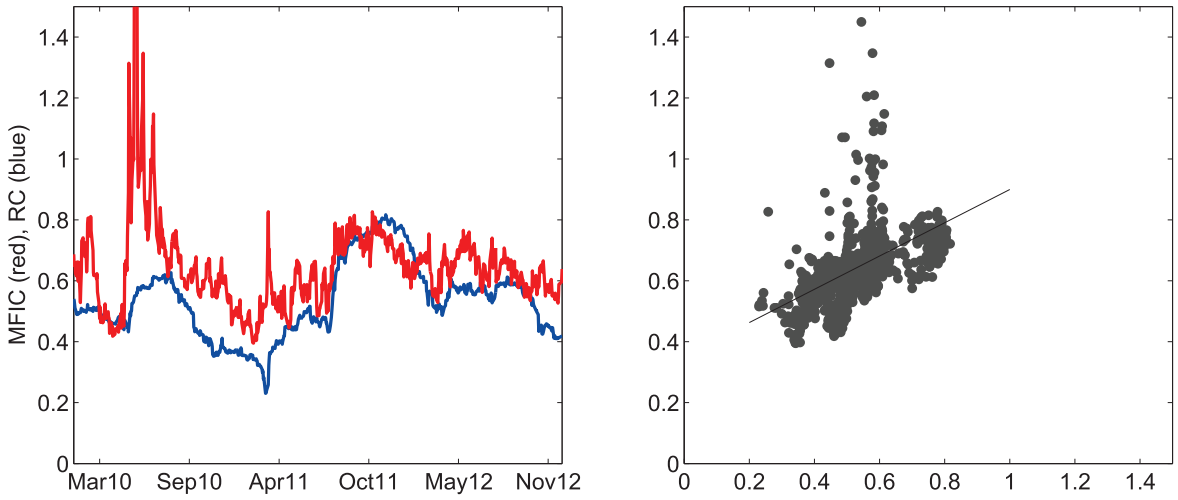


Figure 2. Left panel: DAX $\rho_{t,\tau}$ (blue), $\tilde{\rho}_t(\tau)$ (red). Right panel: Scatter plot of DAX $\rho_{t,\tau}$ (horizontal axis) vs. $\tilde{\rho}_t(\tau)$ (vertical axis), for $t + 0.25$ from 20100802 till 20120801.

in selling index option straddles and purchasing straddles in options on index components. The forecast of the implied correlation surface can provide the insight into the relative cost of index options compared to the price of options on individual stocks that comprise the index. In comparison to the single historical or implied volatility forecast, usually used for this purpose, the correlation surface can provide information for trading options on the whole maturity spectrum. Which means one can buy straddles with different strikes, depending on the implied correlation forecast.

3 Modeling and forecasting correlation dynamics

To determine the amount of hedge for $D_{t+\tau}$ we model the implied correlation (IC) and use the forecast to approximate the floating leg of the dispersion strategy $\rho_{t+\tau}$. By applying (2.2) to IV of a basket $\widehat{\sigma}_{t,B}(\kappa, \tau)$ and its N constituents $\widehat{\sigma}_{t,i}(\kappa, \tau)$, $i \in \{1, \ldots, N\}$, for every t we obtain the IC surface (ICS)

$$\widehat{\rho}_t(\kappa, \tau) = \frac{\widehat{\sigma}^2_{t,B}(\kappa, \tau) - \sum_i w_i^2 \widehat{\sigma}^2_{t,i}(\kappa, \tau)}{\sum_i \sum_{j \neq i} w_i w_j \widehat{\sigma}_{t,i}(\kappa, \tau) \widehat{\sigma}_{t,j}(\kappa, \tau)}. \tag{3.1}$$

Figure 3 displays $\widehat{\rho}_t(\kappa, \tau)$ in different trading days: 20111209, 20120710. Due to the specific option data structure, every day one observes a "cloud of strings" that visually resembles a surface and can be recovered by applying nonparametric smoothing. One can clearly see that surfaces have shape similarities, but vary in levels, slopes and curvatures. Thus they may be treated as daily realizations of a random function. In addition one can observe that the strings do not have fixed spacial locations. In order to model the dynamics of such a complicated multi-dimensional object we apply the DSFM that reduces the dimensionality of the problem and allows the ICS to be studied in a conventional time-series context.

3.1 Model characterization

At every day t one observes ICs $\widehat{\rho}(\kappa_{t,j}, \tau_{t,j})$, $t = 1, \ldots, T$, $j = 1, \ldots, J_t$, where j is the index of observations and J_t the total number of observations at day t. Prior to introducing the model we exclude the case of a fully undiversified basket, with $\widehat{\rho} = 1$, from the analysis and apply a variance stabilizing transformation. Fisher's Z-transformation [29] gives

$$T(u) := \frac{1}{2} \log \frac{1+u}{1-u}$$

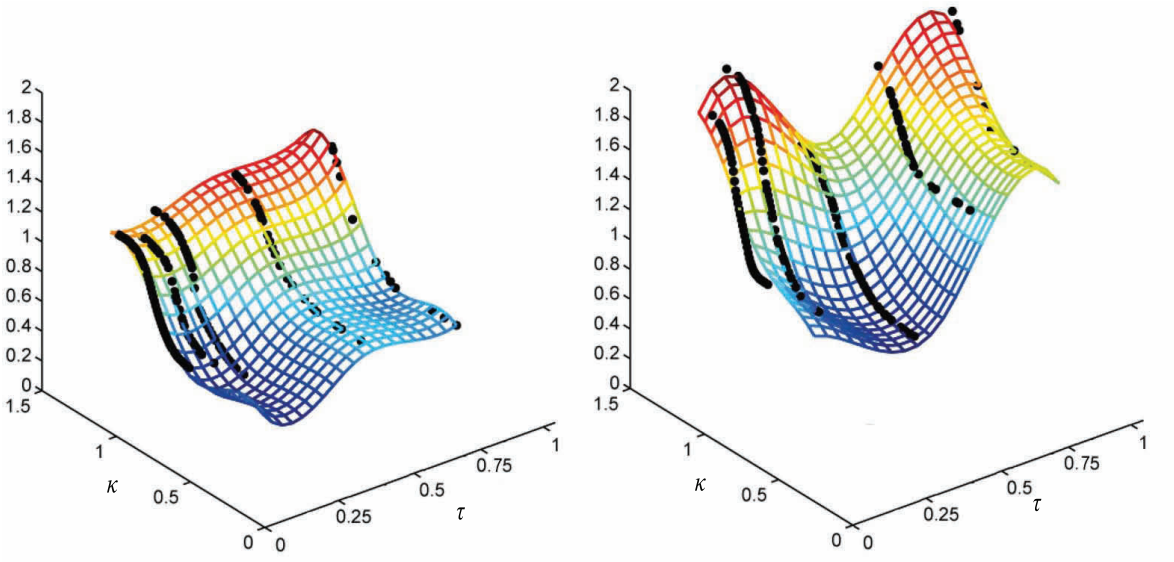


Figure 3. ICS implied by prices of DAX options traded on 20111209 (left) and 20120710 (right); surfaces recovered by the Nadaraya–Watson smoothing.

with $Y_{t,j} := T\{\widehat{\rho}(\kappa_{t,j}, \tau_{t,j})\}$.

Our aim is to model the dynamics of $\{(Y_{t,j}, X_{t,j}),\ 1 \le t \le T,\ 1 \le j \le J_t\}$, where $X_{t,j} = (\kappa_{t,j}, \tau_{t,j})$. The technique we employ allows us to reduce the dimensionality and to simultaneously study the dynamics of Y_t by approximation through an L-dimensional object with $L \ll J$. The DSFM, first introduced by Fengler et al. [24] in an application to IV surface dynamics, and then extended by Park et al. [35] and Song et al. [39], has these desired properties.

The basic idea is to approximate $\mathsf{E}(Y_t|X_t)$ by the sum of $L+1$ smooth basis functions $m := \{m_0, \dots, m_L\}^\top$ (factor loadings) weighted by time-dependent coefficients $Z_t := (1, Z_{t,1}, \dots, Z_{t,L})^\top$ (factors):

$$Y_{t,j} = m_0(X_{t,j}) + \sum_{l=1}^{L} Z_{t,l} m_l(X_{t,j}) + \varepsilon_{t,j}. \tag{3.2}$$

In representation (3.2), m are chosen data-driven and do not have a particular (parametric) form.

Here two important remarks are appropriate. Firstly, the unknown basis functions m must be estimated. Fengler et al. [24] estimated both m and Z_t iteratively using kernel smoothing techniques, Park et al. [35] approximated m by tensor B-splines basis functions weighted by a coefficients matrix. Here we employ a functional principal component analysis (FPCA) approach that will be described in Section 3.2. The nonparametric estimation procedure that we use is introduced in Section 3.3; the basics of this technique can be found in [39].

The second issue is the estimation of the latent factors Z_t. Having the data-driven basis $\widehat{m}_l$ in hand, we can estimate daily factors by the ordinary least squares (OLS) method. Afterwards one fits the econometric model to $\widehat{Z}_t$, as it was done by Cont and Da Fonseca [17] and Hafner [26], who fitted AR(1) to every $Z_{t,l}$, $l \in \{1, \dots, L\}$, or by Fengler et al. [24] who considered a multivariate VAR(2) process.

3.2 Correlation surface with FPCA

We approximate the ICS by the sum of orthogonal functions. By doing so we involve the FPCA theory by looking at the ICS as a stationary random function $f : \mathbb{R}^2 \to \mathbb{R}$.

Let $\mathcal{J} = [\kappa_{\min}, \kappa_{\max}] \times [\tau_{\min}, \tau_{\max}]$ be the range of possible values of $\kappa_{t,j}$ and $\tau_{t,j}$. We introduce (ρ_t), $t \in \{1, \dots, T\}$, the sample of i.i.d. smooth random functions (surfaces). Every ρ_t is a smooth map $\rho_t : \mathcal{J} \to \mathbb{R}$ and satisfies $\int_{\mathcal{J}} \mathsf{E}(\rho_t^2) < \infty$. Also for every ρ_t we assume a well-defined mean function $\mu(u) = \mathsf{E}\{\rho_t(u)\}$ and the

existence of a covariance function $\psi(u,v) = \mathsf{E}[\{\rho_t(u) - \mu(u)\}\{\rho_t(v) - \mu(v)\}]$. With $\phi(u,v) = \mathsf{E}\{\rho_t(u)\rho_t(v)\}$ the covariance function can be expressed as

$$\psi(u,v) = \phi(u,v) - \mu(u)\mu(v), \tag{3.3}$$

which can be also interpreted as a covariance coefficient of two points on the surface with coordinates u and $v \in \mathcal{J}$. Since (3.3) is a symmetric positive definite function, we can use it as a nucleus of the integral transform, performed by the linear operator. Define the covariance operator Γ that transforms f into (Γf):

$$(\Gamma f)(u) = \int_{\mathcal{J}} \psi(u,v) f(v) dv.$$

Γ is a symmetric positive operator with orthonormal eigenfunctions $\{\gamma_j\}_{j=1}^{\infty}$, $\gamma_j : \mathcal{J} \to \mathbb{R}$, and associated eigenvalues $\{\lambda_j\}_{j=1}^{\infty}$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. Now we can express (3.3) in terms of eigenfunctions and eigenvalues of the covariance operator Γ by applying Mercer's theorem (see, e.g., [30]):

$$\psi(u,v) = \sum_{j=1}^{\infty} \lambda_j \gamma_j(u) \gamma_j(v).$$

Taking eigenfunctions $\{\gamma_j\}_{j=1}^{\infty}$ as a basis, we represent $\rho_t(u) - \mu(u)$ as a generalized Fourier series with coefficients given by

$$\zeta_{tj} = \int_{\mathcal{J}} \{\rho_t(u) - \mu(u)\} \gamma_j(u) du,$$

called the j-th principal component score with $\mathsf{E}(\zeta_{tj}) = 0$, $\mathsf{E}(\zeta_{tj}^2) = \lambda_j$ and $\mathsf{E}(\zeta_{tj}\zeta_{ik}) = 0$ for $j \neq k$, see [37]. Thus one may rewrite $\rho_t(u) - \mu(u)$ in the Karhunen–Loève form:

$$\rho_t(u) - \mu(u) = \sum_{j=1}^{\infty} \zeta_{tj} \gamma_j(u). \tag{3.4}$$

Here ζ_{tj} indicates how strong the influence of the j-th basis function on the shape of the t-th surface is. The higher the score, the closer the shape of ρ_t resembles the shape of the j-th eigenfunction.

In practice one needs to take L eigenfunctions to replace the infinite sum in (3.4) by the finite sum of L basis functions, corresponding to the highest eigenvalues. One calls $\{\gamma_j\}_{j=1}^{L}$ the empirical orthonormal basis, see [37]. In the next subsection we discuss the estimation procedure for $\{\gamma_j\}_{j=1}^{L}$ as well as criteria for the L selection.

3.3 Estimation algorithm

In model (3.2) both Z_t and m must be estimated. We do that in two steps.

At the *first step*, we estimate the covariance operator introduced in Section 3.2 and take $\widehat{\mu}$ as $\widehat{m}_0$ and $\widehat{\gamma}_l$ as $\widehat{m}_l$, $l \in \{1, \ldots, L\}$.

The covariance function (3.3) is estimated as described in [27, 41]. The procedure consists in least-squares fitting of two local linear models, for $\widehat{\mu}$ and $\widehat{\psi}$.

Given $u \in \mathcal{J}$ we choose $(\widehat{a}_\mu, \widehat{b}_\mu) = (a_\mu, b_\mu)$ to minimize

$$\sum_{t=1}^{T} \sum_{j=1}^{J_t} \{Y_{t,j} - a_\mu - b_\mu(u - X_{t,j})\}^2 \mathcal{K}_{h_\mu}(X_{t,j} - u), \tag{3.5}$$

and take $\widehat{\mu}(u) = \widehat{a}_\mu$. Then, given $u, v \in \mathcal{J}$ we choose $(\widehat{a}_\phi, \widehat{b}_{\phi,1}, \widehat{b}_{\phi,2}) = (a_\phi, b_{\phi,1}, b_{\phi,2})$ to minimize

$$\sum_{t=1}^{T} \sum_{j,k:1 \leq j \neq k \leq J_t} \{Y_{t,j}Y_{t,k} - a_\phi - b_{\phi,1}(u - X_{t,j}) - b_{\phi,2}(v - X_{t,k})\}^2 \mathcal{K}_{h_\phi}(X_{t,j} - u)\mathcal{K}_{h_\phi}(X_{t,k} - v), \tag{3.6}$$

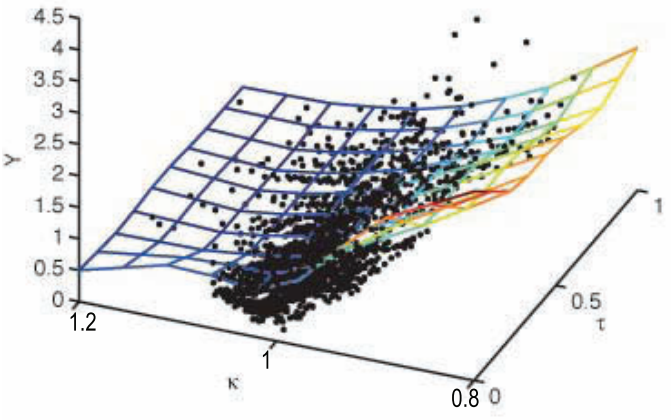


Figure 4. Mean function $\widehat{\mu}(u)$ of the DAX ICS with corresponding data points, estimated from 20100802 till 20110801 with $h_\mu = (h_{\mu,1}, h_{\mu,2})^\top = (0.12, 0.17)^\top$.

and take $\widehat{\phi}(u, v) = \widehat{a}_\phi$. Here $\mathcal{K}_h$ denotes the two-dimensional product kernel, $\mathcal{K}_h(\bar{q}) = k_{h_1}(\bar{q}_1) \times k_{h_2}(\bar{q}_2)$, $h = (h_1, h_2)^\top$, based on one-dimensional $k_h(\bar{q}) = h^{-1}k(h^{-1}\bar{q})$. For our application we selected the quartic kernel, where $k(\bar{q}) = 15/16(1 - \bar{q}^2)^2$ for $|\bar{q}| < 1$ and 0 otherwise. For both (3.5) and (3.6) kernel bandwidths $h_\mu = (h_{\mu,1}, h_{\mu,2})^\top$ and $h_\phi = (h_{\phi,1}, h_{\phi,2})^\top$ are to be selected. The procedure is described in Appendix B. Figure 4 shows an example of $\widehat{\mu}(u)$ estimated using the dataset described in Section 4 for a sub-sample from 20100802 to 20110801.

Finally, having estimates $\widehat{\mu}(u)$ and $\widehat{\phi}(u, v)$, we compute $\widehat{\psi}(u, v)$ using (3.3) and take its L eigenfunctions corresponding to the largest eigenvalues as $\widehat{m}_l$, $l \in \{1, \ldots, L\}$. Parameter L is chosen in such a way that the selected eigenfunctions explain the large share of variability in the original data. It is also necessary to mention that $\widehat{\psi}(u, v)$ is a matrix of a very large dimensionality. To obtain its consistent estimator, suitable for further spectral decomposition, various matrix regularization techniques can be used, e.g. banding as in [6], thresholding as in [5], eigenvalues cleaning as in [31] and factor models described in [23]. We use the latter in this step.

In the *second step*, using $\widehat{m}$, we obtain the estimates $\widehat{Z}_t = (1, \widehat{Z}_{t,1}, \ldots, \widehat{Z}_{t,L})^\top$ as minimizers of the following least squares criterion:

$$\widehat{Z}_t = \arg\min_{Z_t} \sum_{t=1}^{T} \sum_{j=1}^{J_t} \{Y_{t,j} - Z_t^\top \widehat{m}(X_{t,j})\}^2.$$

4 Data

We study the dispersion strategy over the 2-year sample period from 20100802 to 20120801 on the German market represented by the DAX basket. The basket is composed of 23 stocks, constituents of DAX, with the most liquidly traded options and weights proportional to the current market capitalization. To model the dynamics of the IC and to construct the dispersion trade we operate with three main variables representing different correlation estimates: MFIC, RC, and IC. The datasets are described in Table 3.

The *MFIC dataset* contains daily series of MFICs with maturities 0.083, 0.25, 0.5 and 1 years computed via (2.9) from variance swap rates given by Bloomberg as a discrete approximation of (2.6).

The *RC dataset* contains daily series of RCs computed with (2.8) and (2.10) from the Bloomberg end-of-day stock prices over estimation windows 0.083, 0.25, 0.5 and 1 years.

The *IC dataset* is constructed using out-of-the-money (OTM) DAX and single stock options from the EUREX database. To estimate the DSFM model and produce forecasts for the sample period the dataset covers one additional year from 20090803 to 20100730. The dataset is transaction-based, meaning every trade is registered with the date it occurred, expiry date, underlying ticker, exercise price (strike) and settlement

		Min.	Max.	Mean	Median	Stdd.	Skew.	Kurt.
IC	κ	0.8000	1.2000	0.9825	0.9825	0.0986	0.0690	2.0661
	τ	0.0274	0.9671	0.2442	0.1753	0.1979	1.3717	4.3941
	$\widehat{\rho}_t(\kappa, \tau)$	0.0587	0.9998	0.6150	0.6290	0.1566	−0.2739	2.6115
MFIC	$\tilde{\rho}_t(0.083)$	0.3895	0.4860	0.6061	0.6193	0.0834	0.0696	0.1957
	$\tilde{\rho}_t(0.25)$	0.4446	0.9795	0.6549	0.6573	0.0850	0.0613	0.1631
	$\tilde{\rho}_t(0.5)$	0.4997	1.4730	0.7037	0.6953	0.0866	1.8188	0.1305
	$\tilde{\rho}_t(1)$	0.5611	1.0851	0.7496	0.7422	0.0905	0.7764	0.6788
RC	$\rho_{t+0.083}$	0.1754	0.8955	0.5373	0.5013	0.1331	0.5221	−0.2154
	$\rho_{t+0.25}$	0.2774	0.8149	0.5566	0.5363	0.1192	0.2489	−0.8083
	$\rho_{t+0.5}$	0.3794	0.7343	0.5759	0.5713	0.1053	−0.0243	−1.4012
	ρ_{t+1}	0.4312	0.6581	0.5924	0.6050	0.0522	−1.2443	0.9875

Table 3. Summary statistics: IC data computed from the DAX index and constituents options over the period from 20090803 till 20120801 including the 1 year estimation period (3 years, 770 trading days, 135 obs./day). MFIC computed from daily variance swaps rates. RC computed from daily stock returns from 20100802 till 20120801 (2 years, 515 trading days). The figures are given after filtering and data preparation.

price. To obtain IV from option prices via (2.7) we distinguish between index and single stock options. For index options, which have the European type of option payoff, the Black–Sholes (BS) model is used. To account for dividends and early execution in options on single stocks (American payoff) we use binomial trees (see [18]) and the bisection algorithm. Other necessary model parameters, such as stock prices, index levels, dividend amounts for constituent stocks, interest rates and stock market capitalization, are taken from the Bloomberg database. As a risk-free rate proxy we take daily values of EURIBOR (Euro Interbank Offered Rate) with 1 week up to 1 year maturities and use linear interpolation to obtain values for required option τ. We use the most liquid segment of data with κ ranging from 0.8 to 1.2 and τ from 10 days to 1 year. Options outside of this range are excluded from the data set due to the poor data quality, which does not allow to recover implied volatility surfaces for the DAX and all constituents and to compute implied correlation on a daily basis. Figure 3 shows an example of the ICS plotted using the entire available option data, including options outside of the τ-range from 0.8 to 1.2, for two selected "rich with data" days (20111209 and 2012071). As one can see, some correlations observed in Figure 3 are more extreme in comparison to the values in Table 3. The plots show the nature of the implied correlation estimate, which is not necessarily observed in a range from 0 to 1. Those days reveal the possibility of a so-called "volatility arbitrage". Having in mind the empirical findings described in Section 2.2, stating that the VRP of an index is much more pronounced than of constituents, one might take a short position in a too expensive delta-hedged index option, when the implied correlation is considerably higher than 1.

Options from the original EUREX dataset are not given on a regular (κ, τ)-grid, required in (3.1). In the τ-dimension, maturities are standardized by market regulation, so for every t one can find several τ_t, similarly for the index and for all constituents. However, in κ-dimension one needs to interpolate. At every t we use the original (κ_t, τ_t) grid of the index and linearly interpolate IVs of all constituents to obtain values corresponding to this grid. To avoid computational problems with a highly skewed empirical distribution of (κ_t, τ_t), we transform the initial space $[0.8, 1.2] \times [0.03, 1]$ to $[0, 1]^2$ using an empirical distribution function. Also, we remove options with extremely high IVs (larger than 50%) considering them to be the misprints in trade registration and finally use (3.1) to obtain IC, which produces, on average, 135 observations per day.

Figure 1 shows a linear dependence between basket correlation and volatility. We check this finding in the RC dataset for different estimation windows and in IC dataset for different maturities. The RC data allows for the identification of a breakpoint, a threshold, after which the strength of the dependence changes, see Figures 5 and 6. This phenomenon is persistent over different estimation windows. The IC dataset does not show any clear change in correlation/volatility dependence. Since the IC is used to obtain a forecast of a floating leg of the dispersion strategy, which is RC, we propose making a regime dependent correction of the IC dataset as described in Appendix A.

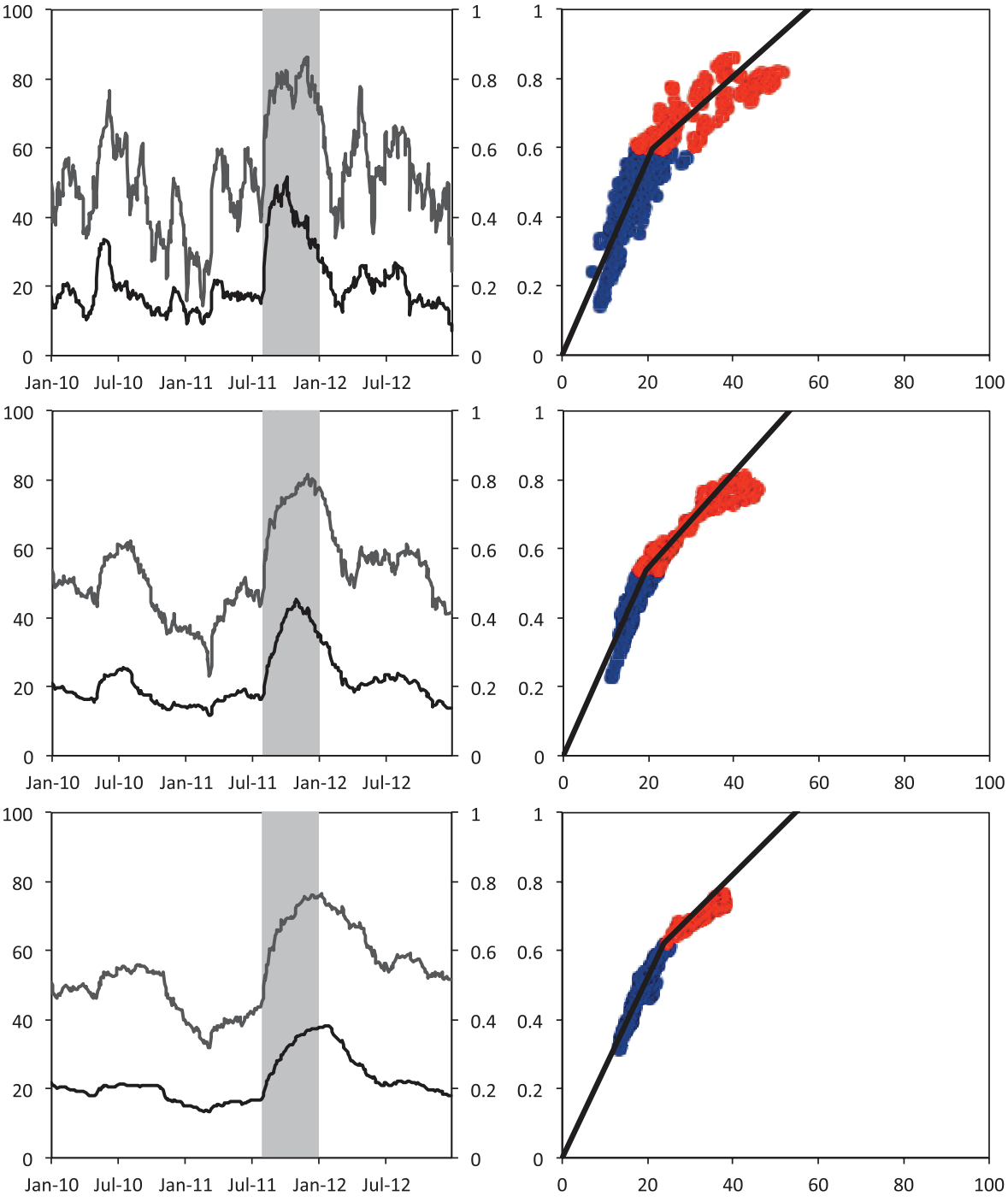


Figure 5. DAX $\sigma_{B,t,\tau}$ (solid line) vs. $\rho_{t,\tau}$ (dashed line), scatter plot $\sigma_{B,t,\tau}$ vs. $\rho_{t,\tau}$, for t, τ from 20100104 till 20121228, estimated with (2.8) and (2.2) over 1 month ($\tau = 0.083$), 3 months ($\tau = 0.25$) and 6 months ($\tau = 0.5$) window. Shaded area: Aug 2011 market fall. The switch point for two regression lines is defined as described in Section 4.

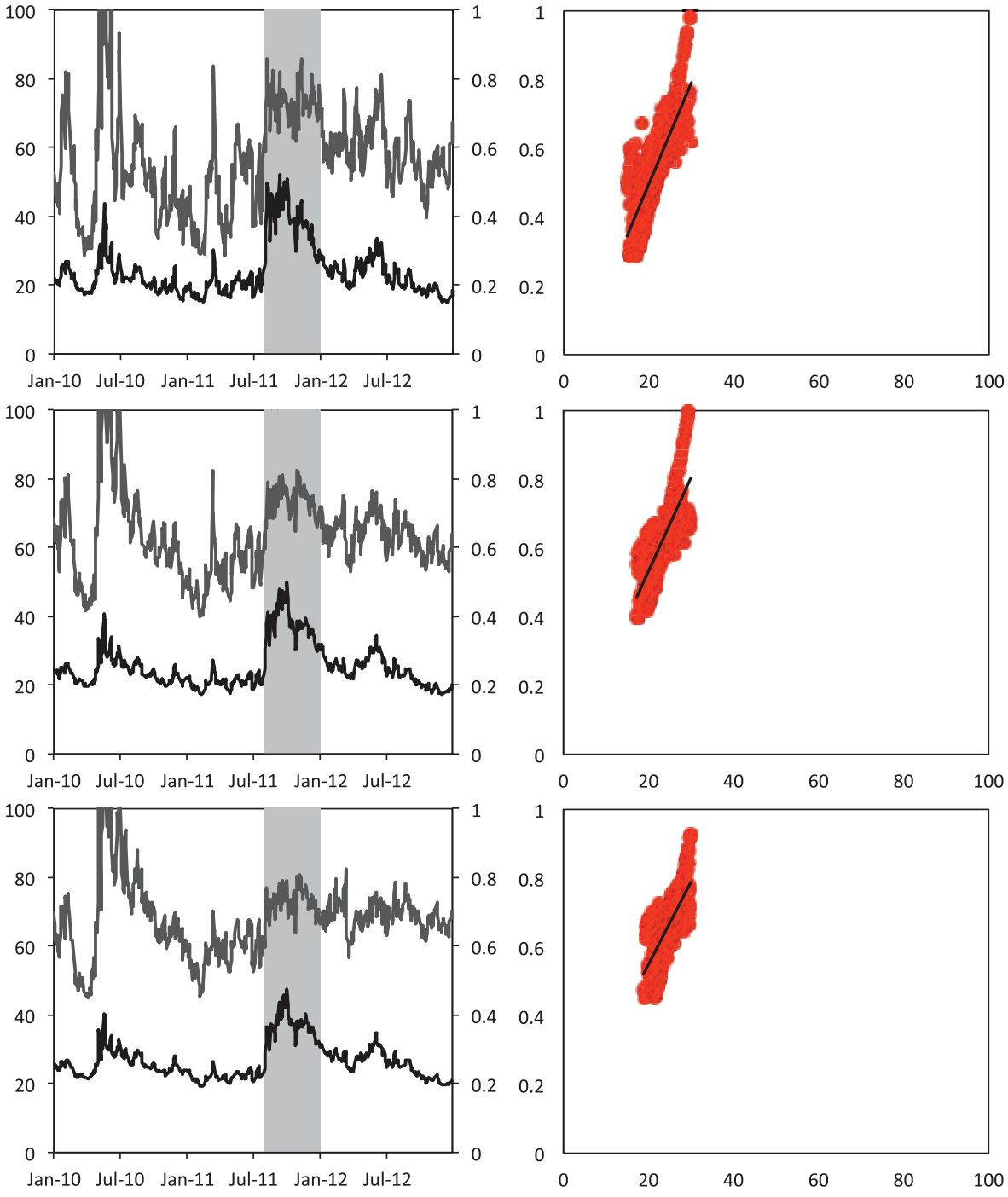


Figure 6. DAX $\widehat{\sigma}_{t,B}(1,\tau)$ (solid line) vs. $\widehat{\rho}_t(1,\tau)$ (dashed line), scatter plot $\widehat{\sigma}_{t,B}(1,\tau)$ and $\widehat{\rho}_t(1,\tau)$, for t,τ from 20100104 till 20121228, estimated from IVs with (3.1) for option with 1 month ($\tau = 0.083$), 3 months ($\tau = 0.25$) and 6 months ($\tau = 0.5$) maturity. Shaded area: Aug 2011 market fall.

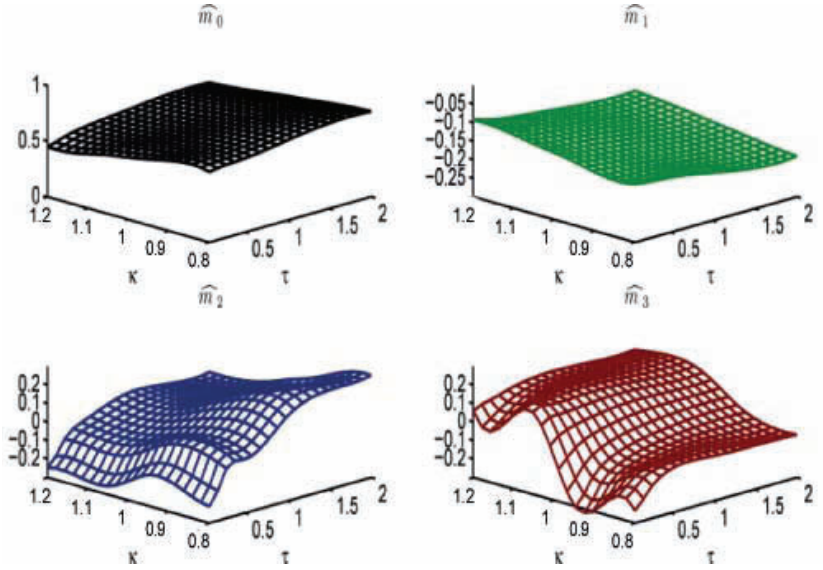


Figure 7. Factor loadings $\widehat{m}_0, \widehat{m}_1, \widehat{m}_2, \widehat{m}_3$ estimated from 20090803 till 20100730.

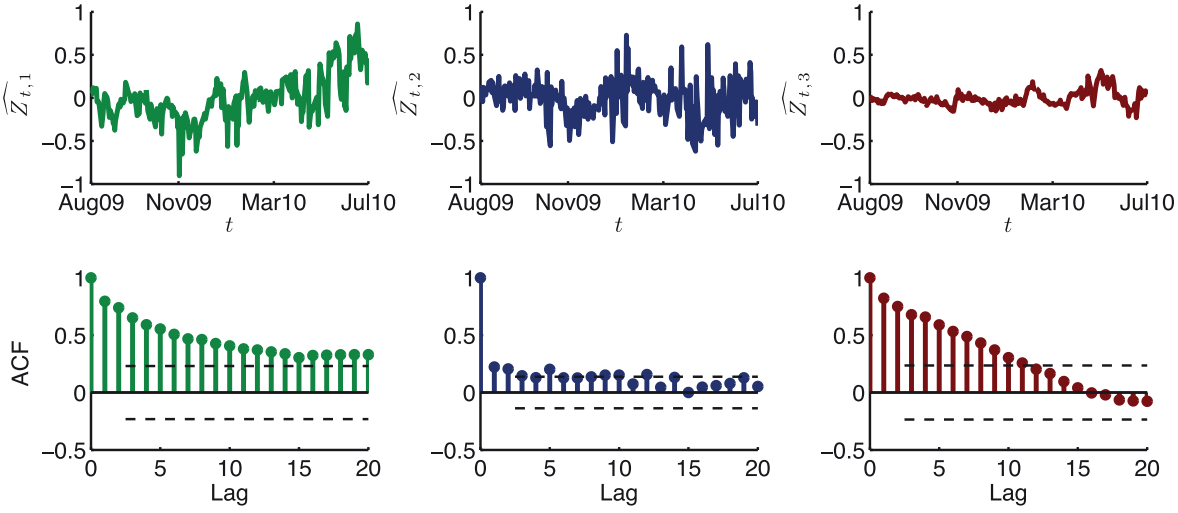


Figure 8. Driving factors of the DAX ICS $\widehat{Z}_{t,1}, \widehat{Z}_{t,2}, \widehat{Z}_{t,3}$ and ACF up to the 20th lag from 20090803 till 20100730.

5 Empirical results

5.1 Estimation results and factor modeling

Using the IC dataset described in Section 4, we estimate the DSFM model for three non-overlapping sub-samples 20090803–20100730 (the 1st year), 20100802–20110729 (the 2nd year), 20110802–20120801 (the 3rd year), and for the entire sample 20090803–20120801. All sub-samples include particularly volatile periods caused by the stock market falls in May 2010, "Flash Crash 2010", and a more pronounced drop in August 2011.

An example of an estimation over the 1st sample year common factor loadings $\widehat{m}_0, \widehat{m}_1, \widehat{m}_2, \widehat{m}_3$ and the daily time series of factors $\widehat{Z}_{t,1}, \widehat{Z}_{t,2}, \widehat{Z}_{t,3}$ is given in Figures 7 and 8. Now the modeling task is simplified to the low-dimensional analysis of factor series. We fit the VAR model of order p for $\widehat{Z}_{t,1}, \widehat{Z}_{t,2}, \widehat{Z}_{t,3}$. Before proposing a proper VAR specification, we check if $\widehat{Z}_t$ has characteristics that violate assumptions for linear multiple time series models. We perform the augmented Dickey–Fuller (ADF) test to check each $\widehat{Z}_{t,1}, \widehat{Z}_{t,2}, \widehat{Z}_{t,3}$

	$\widehat{Z}_{t,1}$	$\widehat{Z}_{t,2}$	$\widehat{Z}_{t,3}$
20090803–20100730 (the 1st year)	−2.991 (1)	−6.982 (1)	−5.710 (3)
20100802–20110729 (the 2nd year)	−1.666∗ (3)	−3.090 (2)	−4.480 (1)
20110802–20120801 (the 3rd year)	−3.511 (2)	−3.796 (3)	−3.480 (2)
20090803–20120801 (the entire sample)	−4.025 (1)	−6.912 (3)	−8.979 (1)

Table 4. Augmented Dickey–Fuller (ADF) test on $\widehat{Z}_{t,1}, \widehat{Z}_{t,2}, \widehat{Z}_{t,3}$. The number of lags included in the ADF regression (in brackets) is chosen by starting with three lags and subsequently deleting lag terms, until the last one is significant at 5% level. Test statistics that do not reject the hypothesis of a unit root at 5% level are denoted by ∗.

		AIC	HQIC	SBIC
20090803–20100730 (the 1st year)	1	1.923	2.061	2.162
	2	1.839∗	1.975∗	2.152∗
	3	1.856	2.052	2.304
	4	1.882	2.060	2.389
20100802–20110729 (the 2nd year)	1	−2.868	−2.800	−2.699
	2	−3.075∗	−2.932∗	−2.755∗
	3	−3.068	−2.898	−2.645
	4	−3.051	−2.854	−2.525
20110802–20120801 (the 3rd year)	1	−0.118	−0.051	0.048
	2	−0.355	−0.238∗	−0.064∗
	3	0.361∗	−0.193	0.055
	4	−0.360	−0.144	0.179
20090803–20120801 (the entire sample)	1	0.745	0.773	0.818
	2	0.384∗	0.461∗	0.539∗
	3	0.397	0.467	0.579
	4	0.412	0.475	0.621

Table 5. Akaike's information criterion (AIC), Schwarz' Bayesian information criterion (SBIC), and the Hannan and Quinn information criterion (HQIC) for defining the optimal lag order p of a VAR model for DAX and S&P100 ICS factors $\widehat{Z}_{t,1}, \widehat{Z}_{t,2}, \widehat{Z}_{t,3}$. The symbol ∗ appearing next to the test statistics indicates the optimal lag at 5% significance level.

for stationarity, see Table 4. For $\widehat{Z}_{t,2}$ in sub-sample 20100802–20110729 we cannot reject the hypothesis of a unit root, so we use its first differences instead. Then we define the appropriate number of lags, or order p, by computing Akaike's information criterion (AIC), Schwarz' Bayesian information criterion (SBIC), and the Hannan and Quinn information criterion (HQIC) values, see Table 5. The symbol ∗ appearing next to the test statistics indicates the optimal lag. Except for the sub-sample 20110802–20120801, the test statistics suggest $p = 2$, so we make a choice in favor of this specification. The estimation results are summarized in Table 6. We also conducted a portmanteau (Q) test for the null hypothesis that a series of residuals exhibits no autocorrelation. The test does not indicate the presence of a serial correlation.

Based on the results, we can distinguish the influence of each factor on the time evolution of the ICS. The first factor can be interpreted as level, the second as maturity and the third as a moneyness effect. The relative sizes of the largest eigenvalues of (3.3) suggests that $\widehat{m}_1$ is capable of capturing the biggest share of the surface variability. The variation captured by the second $\widehat{m}_2$ has a smaller influence, since it is only responsible for the surface shape transformation in the τ-dimension. Finally, since the variation of the ICS in the κ-dimension is relatively small, the $\widehat{m}_3$ has a smaller impact, which is also reflected in the $\widehat{Z}_{t,3}$ series.

The forecast of $\widehat{Z}_{t,1}, \widehat{Z}_{t,2}, \widehat{Z}_{t,3}$ modeled with VAR(2) together with estimated fixed $\widehat{m}_0, \widehat{m}_1, \widehat{m}_2, \widehat{m}_3$ gives a forecast of the ICS.

	20090803–20120801 (the entire sample)			20090803–20100730 (the 1st year)		
	$Z_{1,t}$	$Z_{2,t}$	$Z_{3,t}$	$Z_{1,t}$	$Z_{2,t}$	$Z_{3,t}$
$Z_{1,t-1}$	0.645∗	−0.012	−0.019	0.630∗	−0.032	0.029
$Z_{1,t-2}$	0.310∗	0.008	0.029	0.276∗	0.013	−0.060∗
$Z_{2,t-1}$	−0.104∗	0.259∗	0.156	−0.036	0.047	0.036
$Z_{2,t-2}$	0.057∗	0.406∗	−0.014	−0.039	0.339	−0.104∗
$Z_{3,t-1}$	−0.07	0.140∗	0.471∗	−0.091	−0.494∗	0.525∗
$Z_{3,t-2}$	0.149∗	0.118∗	0.251∗	0.046	0.181	−0.208∗
c	0.004	0.006	−0.003	−0.004	−0.001	0.001

	20100802–20110729 (the 2nd year)			20110802–20120801 (the 3rd year)		
	$Z_{1,t}$	$Z_{2,t}$	$Z_{3,t}$	$Z_{1,t}$	$Z_{2,t}$	$Z_{3,t}$
$Z_{1,t-1}$	0.809∗	0.048	−0.202∗	0.339∗	0.191∗	−0.001
$Z_{1,t-2}$	0.254∗	−0.029	0.112∗	0.351∗	0.041	−0.036
$Z_{2,t-1}$	0.188∗	0.687∗	−0.223∗	0.355∗	0.264∗	0.132∗
$Z_{2,t-2}$	0.091	0.262∗	0.018	0.084	0.302∗	−0.045
$Z_{3,t-1}$	0.453∗	0.051	0.162∗	−0.197	−0.008	0.623∗
$Z_{3,t-2}$	0.118	0.118	0.275∗	0.044	0.240∗	0.204∗
c	−0.004	0.001	−0.002	0.003	−0.001	−0.002

Table 6. The estimated parameters for the VAR(2) model for DAX ICS factors. Estimates which are not significant at 5% level are marked with ∗.

τ	Min.	Max.	Mean	Median	Stdd.	Skew.	Kurt.
0.083	−108.04	72.30	−1.14	−0.71	8.00	−6.61	100.49
0.25	−255.48	49.53	−1.20	−0.41	11.49	−17.58	372.33
0.5	−216.04	32.78	−0.74	−0.30	9.37	−18.66	425.86
1	−64.84	76.59	−0.01	−0.38	7.47	2.74	46.85

Table 7. Performance of naïve hedge, summary statistics for $\varepsilon^h_{t+\tau}$ from 20100101 till 20120801.

5.2 Backtesting the dispersion strategy

Here we show that using the correlation forecast, one can improve the original dispersion strategy (2.11) and test it empirically over the 2-years sample period 20100801–20120802. We compare the payoff of the strategy without hedging with the *naïve hedging strategy* and propose its improvement, the *advanced strategy*.

To obtain the value of the naïve hedge position to be held over Δt days from $t + \tau - \Delta t$ till $t + \tau$, we make a Δt-days ahead DSFM forecast $\widehat{\rho}_{t+\tau}(1, t + \tau)$ and use it as $\rho_{t+\tau}$ in (2.11). Thus the size of the position is defined by

$$D^h_{t+\tau} = \sum_i \sum_{j \neq i} w_i w_j \tilde{\sigma}_{t,i}(\tau)\tilde{\sigma}_{t,j}(\tau)\{\tilde{\rho}_t(\tau) - \widehat{\rho}_{t+\tau}(1, t + \tau)\}. \tag{5.1}$$

The corresponding relative hedging error is given by

$$\varepsilon^h_{t+\tau} = \frac{D^h_{t+\tau} - D_{t+\tau}}{D_{t+\tau}} = -\frac{\widehat{\rho}_{t+\tau}(1, t + \tau) - \rho_{t+\tau}}{\tilde{\rho}_t(\tau) - \rho_{t+\tau}}, \tag{5.2}$$

where $\varepsilon^h_{t+\tau} < 0$ (> 0) means that the hedge (5.1) underestimates (overestimates) the actual position (2.11). Table 7 gives summary statistics for (5.2) over the studied sample period for three trades with four different maturities: 0.083, 0.25, 0.5 and 1 years. The statistic includes 515 trades originated every day and expired over the given 2-year sample period, Δt is one day.

Strategy	τ	Min.	Max.	Mean	Stdd.
$D_{t+\tau}$	0.083	−1502.58	**1080.23**	87.09	356.94
(no hedge)	0.25	−1531.94	**1282.31**	101.92	440.54
	0.5	−1270.90	**1301.28**	136.91	456.75
	1	−872.76	**760.92**	**134.26**	299.01
$D_{t+\tau} - D^h_{t+\tau}$	0.083	−3237.72	617.40	15.35	**203.09**
(naïve hedge)	0.25	−1726.53	413.28	35.90	**110.14**
	0.5	−1301.47	344.91	41.13	**91.91**
	1	−914.27	327.03	79.62	**93.14**
$D^{\text{adv}}_{t+\tau}$	0.083	**−1375.99**	1011.38	**100.93**	256.50
(advanced hedge)	0.25	**−1137.79**	**1282.31**	**195.09**	248.41
	0.5	**−760.85**	**1301.28**	**231.35**	281.66
	1	**−367.89**	623.38	123.04	190.80

Table 8. Summary statistics for payoffs $D_{t+\tau}$ (no hedge), $D_{t+\tau} - D^h_{t+\tau}$ (naïve hedge), $D^{\text{adv}}_{t+\tau}$ (advanced hedge) from 20100101 till 20120801. Δt is one day. Best results (highest min, max, mean and smallest stdd.) are given in bold.

The improved version of the strategy uses the DSFM forecast $\widehat{\rho}_{t+\tau}(1, t+\tau)$ as a trigger which defines whether one should hedge or not. If $\widehat{\rho}_{t+\tau}(1, t+\tau) \geq \tilde{\rho}_t(\tau)$ (DSFM predicts loss in dispersion strategy), an investor takes an offsetting (with negative sign) position in (5.1); if $\widehat{\rho}_{t+\tau}(1, t+\tau) < \tilde{\rho}_t(\tau)$ (DSFM predicts gain in dispersion strategy), then no hedge is necessary. Thus we can write the payoff of the advanced strategy at $t+\tau$ as follows:

$$D^{\text{adv}}_{t+\tau} = \begin{cases} D_{t+\tau} - D^h_{t+\tau} & \text{if } \widehat{\rho}_{t+\tau}(1, t+\tau) \geq \tilde{\rho}_t(\tau), \\ D_{t+\tau} & \text{if } \widehat{\rho}_{t+\tau}(1, t+\tau) < \tilde{\rho}_t(\tau). \end{cases}$$

Since a variance swap contract costs nothing to initiate (we ignore transactions costs), the presented series of daily payoffs corresponds to daily P&L of the hypothetical trade where swaps expire daily over the whole period from 20100801 to 20120802. We compare the cash flows from three strategies. As one can see in Table 8, the advanced strategy outperforms the other two by having the smallest maximal losses, highest maximal gains ($\tau = 0.25, 0.5$) and the highest (second highest for $\tau = 1$) average payoff over the studied sample period.

In this simplified setting the financing costs are not taken into account for both strategies.

6 Conclusions

In this study we investigated the implied correlation (IC) of the DAX index basket and introduced a hedging approach for the dispersion trading strategy using the IC forecast. We applied the dynamic semiparametric factor model (DSFM) to the IC dataset from January 2010 to August 2012, recovered four basis functions and three time series of factors and used them to forecast the IC. The advanced dispersion strategy we employed using the IC forecast shows the smallest maximal losses, the highest maximal gains and the highest average payoff over the studied sample period in comparison to the alternative strategies. So, we conclude that our modeling approach can be of potential use in equity dispersion trading.

The choice of DSFM as a model for the IC surface (ICS) dynamics is motivated by the degenerated dataset design, which has to be modeled nonparametrically. On the other hand we were driven by the necessity to reduce the dimensionality of the problem and facilitate the forecasting. DSFM satisfies both requirements. It captures the form of the ICS by its nonparametric part well, and allows a simple parametric model for dynamics to be used. At the later modeling stage we fit the three-dimensional VAR(2) model, which is a good choice to carry out the forecasting exercise. In addition, we found that it is possible to separate the influence of each recovered basis function on the ICS shape. The functions allow their interpretation as level, moneyness and maturity effects. The strength of these effects is defined by the time series of corresponding factors, which can be characterized as drivers of the correlation risk. An interesting task would be to study the presence,

τ	$\sigma_{B,t+\tau}$	$\rho_{t+\tau}$	Slope 1	Slope 2
0.083	20.24	0.5917	0.0361	0.0085
0.25	20.34	0.5728	0.0336	0.0093
0.5	22.42	0.6008	0.0286	0.0094
Average	21.00	0.5884	0.0328	0.0091

Table 9. Segmented linear regression of $\rho_{t+\tau}$ on $\sigma_{B,t+\tau}$ with one break point, $\tau = 0083, 0.25, 0.5$ for $t + \tau$, from 20100104 till 20120801.

size and magnitude of the correlation risk premia, captured by these factors. We consider this findings to be important topics for further research.

A Switch point selection for correlation regimes

The dependence of ρ and σ_B observed in RV and RC is not pronounced in case of ATM IV and IC, see Figures 5 and 6. Therefore we propose a market regime correction scheme for the IC dataset. The idea is to find a breakpoint between two segments of a piecewise linear regression of $\rho_{t+\tau}$ on $\sigma_{B,t+\tau}$. Using the procedure described in [34], we fit a segmented linear regression with one break point. Based on results summarized in Table 9 we make the following state-dependent correction: if $\widehat{\sigma}_{B,t}(1,\tau) > 21$ (high volatility regime), then $\widehat{\rho}_t(\kappa,\tau) = 0.0091\widehat{\sigma}_{B,t}(\kappa,\tau)$.

B Smoothing parameters selection

For both (3.5) and (3.6) kernel bandwidths $h_\mu = (h_{\mu,1}, h_{\mu,2})^\top$ and $h_\phi = (h_{\phi,1}, h_{\phi,2})^\top$ are to be selected. As suggested in [28] we use the penalizing function approach to select optimal h_μ^{opt}, minimizing the mean integrated squared error (MISE):

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{J_t}\sum_{j=1}^{J_t}\left\{Y_{t,j} - \sum_{l=1}^{L}\widehat{Z}_{t,l}\widehat{m}_l(X_{t,j})\right\}^2 w_{h^*,t}(X_{t,j})\Xi_{\text{AIC}}\left\{\frac{W_{h^*,t,j}(X_{t,j})}{TJ_t}\right\}, \tag{B.1}$$

with the Akaike (1970) information criterion (AIC) as penalizing function $\Xi_{\text{AIC}}(q) = \exp(2q)$ and $W_{h^*,t,j}(X_{t,j})$ defined by

$$W_{h^*,t,j}(X_{t,j}) = \frac{\mathcal{K}_h(0)}{J_t^{-1}\sum_{k=1}^{J_t}\mathcal{K}_h(X_{t,k} - X_{t,j})}$$

for every $X_{t,j}$, $1 \le t \le T$, $1 \le j \le J_t$.

Since the distribution of the observations is very uneven, we use the weighted version of the criterion with weights $w_{h^*,t}(\bar{u}) := p_{h^*,t}^{-1}(\bar{u})$, where $p_{h^*,t}(\bar{u})$ is the average design density. For every $X_{t,j}$, $1 \le t \le T$, $1 \le j \le J_t$ it is defined by

$$p_{h^*,t}(X_{t,j}) = J_t^{-1}\sum_{k=1}^{J_t}\mathcal{K}_h(X_{t,k} - X_{t,j}).$$

The bandwidth $h_{\mu\text{AIC}}^{\text{opt}} = (h_{\mu 1}, h_{\mu 2})^\top$ corresponding to the minimal criterion (B.1) is taken as optimal. The bandwidth h^* of the weighting function is constant and does not depend on the choice of h_μ.

Acknowledgment: Thanks go to L. Udvarhelyi for editorial assistance.

Funding: The authors gratefully acknowledge financial support from the Deutsche Forschungsgemeinschaft through CRC 649 "Economic Risk".

References

[1] P. Allen and N. Granger, Correlation vehicles. Techniques for trading equity correlation, Technical Report, JPMorgan, London, 2005.

[2] Z. D. Bai, Methodologies in spectral analysis of large dimensional random matrices, *Statist. Sinica* **9** (1999), 611–677.

[3] G. Bakshi, C. Cao and Z. Chen, Do call and underlying prices always move in the same direction?, *Rev. Financial Stud.* **13** (2000), no. 3, 549–584.

[4] G. Bakshi, N. Kapadia and D. Madan, Stock return characteristics, skew laws, and differential pricing of individual equity options, *Rev. Financial Stud.* **16** (2003), 101–143.

[5] P. J. Bickel and E. Levina, Covariance regularization by thresholding, *Ann. Statist.* **36** (2008), no. 6, 2577–2604.

[6] P. J. Bickel and E. Levina, Regularized estimation of large covariance matrices, *Ann. Statist.* **36** (2008), no. 1, 199–227.

[7] B. J. Blair, S.-H. Poon and S. J. Taylor, Forecasting S&P100 volatility: The incremental information content of implied volatilities and high-frequency index returns, *J. Econom.* **105** (2001), no. 1, 5–26.

[8] T. Bollerslev, R. F. Engle and J. M. Wooldridge, A capital asset pricing model with time-varying covariances, *J. Political Econom.* **96** (1988), no. 1, 116–31.

[9] S. Bossu, S. Strasser and R. Guichard, Just what you need to know about variance swaps, Technical Report, JPMorgan, London, 2005.

[10] F. Bourgoin, Stress-testing correlations: An application to portfolio risk management, in: *Developments in Forecast Combination and Portfolio Choice*, Wiley, New York, (2001).

[11] D. T. Breeden and R. H. Litzenberger, Prices of state-contingent claims implicit in option prices, *J. Business* **51** (1978), no. 4, 621–51.

[12] M. Britten-Jones and A. J. Neuberger, Option prices, implied price processes, and stochastic volatility, *J. Finance* **55** (2000), no. 2, 839–866.

[13] J. M. Campa and P. Chang, The forecasting ability of correlations implied in foreign exchange options, *J. Int. Money Finance* **17** (1998), no. 6, 855–880.

[14] P. Carr and D. Madan, Towards a theory of volatility trading, in: *Reprinted in Option Pricing, Interest Rates, and Risk Management*, Cambridge University Press, Cambridge (1998), 417–427.

[15] P. Carr and L. Wu, Variance risk premiums, *Rev. Financial Stud.* **22** (2009), no. 3, 1311–1341.

[16] B. Christensen and N. Prabhala, The relation between implied and realized volatility, *J. Financial Econom.* **50** (1998), no. 2, 125–150.

[17] R. Cont and J. Da Fonseca, Dynamics of implied volatility surfaces, *Quant. Finance* **2** (2002), 45–60.

[18] J. C. Cox, S. A. Ross and M. Rubinstein, Option pricing: A simplified approach, *J. Financial Econom.* **7** (1979), 229–263.

[19] K. Demeterfi, E. Derman, M. Kamal and J. Zou, More than you ever wanted to know about volatility swaps, Technical Report, Goldman Sachs, New York, 1999.

[20] J. Driessen, P. J. Maenhout and G. Vilkov, The price of correlation risk: Evidence from equity options, *J. Finance* **9** (2009), 1377–1406.

[21] R. Engle, Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models, *J. Bus. Econom. Statist.* **20** (2002), no. 3, 339–350.

[22] R. F. Engle, N. Shephard and K. Sheppard, Fitting vast dimensional time-varying covariance models, Economics Series Working Papers 403, University of Oxford, Oxford, 2008.

[23] J. Fan, Y. Fan and J. Lv, High dimensional covariance matrix estimation using a factor model, *J. Econom.* **147** (2008), 186–197.

[24] M. R. Fengler, W. K. Härdle and E. Mammen, A semiparametric factor model for implied volatility surface dynamics, *J. Financial Econom.* **5** (2007), no. 2, 189–218.

[25] J. Fleming, The quality of market volatility forecasts implied by S&P100 index option prices, *J. Empirical Finance* **5** (1998), no. 4, 317–345.

[26] R. Hafner, *Stochastic Implied Volatility*, Springer, Heidelberg, 2004.

[27] P. Hall, H.-G. Müller and J.-L. Wang, Properties of principal component methods for functional and longitudinal data analysis, *Ann. Statist.* **34** (2006), no. 3, 1493–1517.

[28] W. Härdle, M. Müller, S. Sperlich and A. Werwatz, *Nonparametric and Semiparametric Models*, Springer, Heidelberg, 2004.

[29] W. Härdle and L. Simar, *Applied Multivariate Statistical Analysis*, 4rd ed., Springer, Heidelberg, 2012.

[30] J. Indritz, *Methods in Analysis*, Macmillan, New York, 1963.

[31] L. Laloux, P. Cizeau, J.-P. Bouchaud and M. Potters, Noise dressing of financial correlation matrices, *Phys. Rev. Lett.* **83** (1999), 1467–1470.

[32] O. Ledoit and M. Wolf, Improved estimation of the covariance matrix of stock returns with an application to portfolio selection, *J. Empirical Finance* **10** (2003), no. 5, 603–621.

[33] J. Lopez and C. Walter, Is implied correlation worth calculating? Evidence from foreign exchange options and historical data, *J. Derivatives* **7** (2000), no. 3, 65–81.

[34] V. M. R. Muggeo, Estimating regression models with unknown break-points, *Stat. Med.* **22** (2003), no. 19, 3055–3071.
[35] B. Park, E. Mammen, W. Härdle and S. Borak, Dynamic semiparametric factor models, *J. Amer. Statist. Assoc.* **104** (2009), 284–298.
[36] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, T. Guhr and H. E. Stanley, Random matrix approach to cross correlations in financial data, *Phys. Rev. E* **65** (2002), Article ID 066126.
[37] J. Ramsay and B. W. Silverman, *Functional Data Analysis*, 2nd ed., Springer Ser. Statist., Springer, Heidelberg, 2010.
[38] V. Skintzi and A. Refenes, Implied correlation index: A new measure of diversification, *J. Futures Markets* **25** (2005), 171–197.
[39] S. Song, W. K. Härdle and Y. Ritov, High dimensional nonstationary time series modelling with generalized dynamic semiparametric factor model, *Econom. J.* **17** (2014), 1–32.
[40] S. Sperlich, O. B. Linton and W. K. Härdle, Integration and backfitting methods in additive models-finite sample properties and comparison, *Test* **8** (1999), 419–458.
[41] F. Yao, H.-G. Müller and J.-L. Wang, Functional data analysis for sparse longitudinal data, *J. Amer. Statist. Assoc.* **100** (2005), 577–590.

Journal of Econometrics 192 (2016) 499–513

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

TENET: Tail-Event driven NETwork risk☆

Wolfgang Karl Härdle [a], Weining Wang [a,b], Lining Yu [c,*]

[a] C.A.S.E. - Center for Applied Statistics and Econometrics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

[b] Department of Mathematics, King's College London, United Kingdom

[c] C.A.S.E. - Center for Applied Statistics and Econometrics, IRTG 1792, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

ARTICLE INFO

Article history:
Available online 11 February 2016

JEL classification:
G01
G18
G32
G38
C21
C51
C63

Keywords:
Systemic risk
Systemic risk network
Generalized quantile
Quantile single-index regression
Value at risk
CoVaR
Lasso

ABSTRACT

CoVaR is a measure for systemic risk of the networked financial system conditional on institutions being under distress. The analysis of systemic risk is the focus of recent econometric analyses and uses tail event and network based techniques. Here, in this paper we bring tail event and network dynamics together into one context. In order to pursue such joint efforts, we propose a semiparametric measure to estimate systemic interconnectedness across financial institutions based on tail-driven spillover effects in a high dimensional framework. The systemically important institutions are identified conditional to their interconnectedness structure. Methodologically, a variable selection technique in a time series setting is applied in the context of a single-index model for a generalized quantile regression framework. We could thus include more financial institutions into the analysis to measure their tail event interdependencies and, at the same time, be sensitive to non-linear relationships between them. Network analysis, its behaviour and dynamics, allows us to characterize the role of each financial industry group in 2007–2012: the depositories received and transmitted more risk among other groups, the insurers were less affected by the financial crisis. The proposed TENET - Tail Event driven NETwork technique allows us to rank the Systemic Risk Receivers and Systemic Risk Emitters in the US financial market.

1. Introduction

Systemic risk endangers the stability of the financial market, the failure of one institution may harm the whole financial system. The sources of risk are complex, as both exogenous and endogenous factors are involved. This calls for a study on a financial network which accounts for interaction between the agents in the financial market. Although the notion *systemic risk* is not novel in academic literature (see, e.g., Minsky (1977)), it had been neglected both in academia and in the financial risk industry until the outbreak of the financial crisis in 2008. Some financial institutions collapsed, even some major ones like Lehman Brothers, Federal Home Loan Mortgage Corporation (Freddie Mac), and Federal National Mortgage Association (Fannie Mae). The magnitude of repercussions caused by this financial crisis and its complexity revealed a significant flaw in financial regulations. As in the past, regulations had been focused primarily on stability of a single financial institution. The detailed actions involved the establishment of Financial Stability Board (FSB) after G-20 London summit in 2009, integration of systemic risk agenda into Basel III in 2010 prior to the G-20 meeting in Seoul, and enacting the Dodd Frank Wall Street Reform and Consumer Protection Act ('Dodd Frank Act') in the US in 2010, which is said to have bought the most radical changes into the US financial system since the Great Depression.

In this context, the focus is on *systemically important financial institutions* (SIFIs) whose failure may not only impair the functioning of the financial system but also have adverse effects on the real sectors of the economy. Therefore, we face several challenges such as identifying SIFIs, studying the propagation mechanism of a shock in a system, or in a network formed by financial institutions, investigating the response of a system to a shock as a whole network and establishing a theoretical framework for systemic risk.

Although systemic risk is a relatively straightforward concept aimed at measuring risk stemming from interaction between the agents, the variety of systemic risk measures and diversity of

☆ Financial support from the Deutsche Forschungsgemeinschaft (DFG) via SFB 649 "Ökonomisches Risiko", IRTG 1792 "High-Dimensional Non-Stationary Times Series" and Sim Kee Boon Institute for Financial Economics, Singapore Management University are gratefully acknowledged.

* Corresponding author.
E-mail addresses: haerdle@wiwi.hu-berlin.de (W.K. Härdle), wangwein@wiwi.hu-berlin.de (W. Wang), yulining@wiwi.hu-berlin.de (L. Yu).

http://dx.doi.org/10.1016/j.jeconom.2016.02.013
0304-4076/

methods to model interaction effects lead to the fact that the literature on this topic is highly heterogeneous. The relevant literature in this field can be broadly divided into two groups: economic modelling of systemic risk and financial intermediation including microeconomic (e.g. Beale et al. (2011), Eisenberg and Noe (2001)) and macroeconomic approaches (e.g. Gertler and Kiyotaki (2010)) with the emphasis on theoretical, structural frameworks, and quantitative modelling with the emphasis on empirical analysis. The quantitative literature can be further classified by statistical methodology into quantile regression based modelling such as linear bivariate model by Adrian and Brunnermeier (2011), Acharya et al. (2012), Brownlees and Engle (2015), high-dimensional linear model by Hautsch et al. (2015) and Betz et al. (2016), partial quantile regression by Giglio et al. (2012) and by Chao et al. (2015). Further approaches include principal-component-based analysis, e.g. by Bisias et al. (2012), Rodriguez-Moreno and Peña (2013) and others; statistical modelling based on default probabilities by Lehar (2005), Huang et al. (2009), and others; graph theory and network topology, e.g. Boss et al. (2006), Chan-Lau et al. (2009), and Diebold and Yilmaz (2014).

Our paper belongs to the quantitative group of the aforementioned literature, namely, modelling the tail event driven network risk based on quantile regressions augmented with non-linearity and variable selection in a high dimensional time series setting. Our method is in nature different from Acharya et al. (2012) and Brownlees and Engle (2015)'s method. Acharya et al. (2012) has measured the systemic risk relevance without capturing the network effects of liquidity exposure, and Brownlees and Engle (2015) analyse the risk of a specific asset given the distress of the whole system, which is a reverse of our system at institutional analysis, and their method would capture little spillover effects. Therefore, we believe, that our method is a good addition to the literature of systemic risk measures. Also compared to Diebold and Yilmaz (2014), we focus more on the tail event driven interconnectedness, which cannot be captured by conditional correlation. As a starting point of our research we take co-Value-at-Risk, or CoVaR, modeled by Adrian and Brunnermeier (2011) (from here on abbreviation as AB), where 'co-' stands for 'conditional', 'contagion', 'comovement'. To capture the tail interconnectedness between the financial institutions in the system AB evaluate bivariate linear quantile regressions for publicly traded financial companies in the US.

Whereas AB focus on bivariate measurement of tail risk, we aim at assessing the systemic risk contribution of each institution conditional on its tail interconnectedness with the relevant institutions. Thus, the primary challenge is selecting the set of relevant risk drivers for each financial institution. Statistically we address this issue by employing a variable selection method in the context of single-index model (SIM) for generalized quantile regressions, i.e. for quantiles and expectiles. We further extend it to a time series variable selection context in high dimensions. The semi-parametric framework due to the SIM allows us to investigate possible non-linearities in tail interconnectedness. Based on identified relevant risk drivers we construct a financial network consisting of spillover effects across financial institutions. Further we define two indices: Systemic Risk Receiver and Systemic Risk Emitter, which combine network structure and market capitalization to identify the systemically important financial institutions.

The assumption of non-linear relationship between returns of financial companies is motivated by previous work by Chao et al. (2015), who find that the dependency between any pair of financial assets is often non-linear, especially in periods of economic downturn. Moreover, non-linearity assumption is more flexible especially in a high dimensional setting where the system becomes too complex to support the belief of linear relationships. From the 2012 US financial company list from NASDAQ, we select 100 financial institutions consisting of the top 25 financial institutions from each industry group: Depositories, Insurance companies, Broker–Dealers and Others. These four groups are divided by Standard Industrial Classification (SIC) codes. Our model is evaluated, based on weekly log returns of these 100 publicly traded US financial institutions. Firm specific characteristics from balance sheet information such as leverage, maturity mismatch, market to book and size are added into the model as well. Furthermore, the macro state variables are also involved. The time period from 5 January, 2007 to 4 January, 2013 covers one recession (from December 2007 to June 2009) and several documented financial crises (2008 and 2011). Dividing companies by industry groups and including several market perturbations allows not only to select the key players for each time period, but also additionally to highlight the connections between financial industries, which can in turn provide additional information on the nature of market dislocations. In application we find out that there are more interconnectedness between 2008 and 2010. While the bank sector plays a major role in the financial crisis, the insurance companies, however, play more passive roles in terms of risk transmission and risk reception. The most connected financial institutions with respect to incoming and outgoing links are ranked based on our network analysis. The new insight of our finding is that the non-linear relationships between financial firms are stronger during the financial crisis than the stable periods. In addition, to identify the systemically important firms, we weight the connections by firms' market capitalization. The empirical findings suggest that our method can effectively identify the systemic risk relevant firms similar to the literature. Moreover, we can discover the asymmetry and non linearity of the firms' dependency structure, which leads to more accurate measures in terms of backtesting performance. All the R programs for this paper can be found on www.quantlet.de or www.quantlet.de/d3/ia/.

The rest of the paper is organized as follows. In Section 2 our approach of systemic risk modelling is outlined. Section 3 illustrates the empirical application. Section 4 concludes. Appendix A presents the statistical methodology and the related theorems. Appendix B contains proofs.

2. Systemic risk modelling

In this section, we lay down the background and the basic setup of our systemic risk analysis, which can be divided into three steps.

2.1. Basic concepts

Traditional measures assessing riskiness of a financial institution such as Value at Risk (VaR), or expected shortfall (ES) are based either on company characteristics or integrated macro state variables which account for the general state of the economy. Thus, for example, the VaR of a financial institution i at $\tau \in (0, 1)$ is defined as:

$$\mathrm{P}(X_{i,t} \leq VaR_{i,t,\tau}) \stackrel{\mathrm{def}}{=} \tau, \tag{1}$$

where τ is the quantile level, $X_{i,t}$ represents the log return of financial institution i at time t. AB propose, the risk measure CoVaR (Conditional Value at Risk) which takes spillover effects and the macro state of the economy into account. The CoVaR of a financial institution j given X_i at level $\tau \in (0, 1)$ at time t is defined as:

$$\mathrm{P}\left\{X_{j,t} \leq CoVaR_{j|i,t,\tau} | R_{i,t}\right\} \stackrel{\mathrm{def}}{=} \tau, \tag{2}$$

where $R_{i,t}$ denotes the information set which includes the event of $X_{i,t} = VaR_{i,t,\tau}$ and M_{t-1}. Note that M_{t-1} is a vector of macro state

variables reflecting the general state of the economy (see Section 3 for details of macro state variables).

We start with the concept of CoVaR, which is estimated in two steps of linear quantile regression:

$$X_{i,t} = \alpha_i + \gamma_i M_{t-1} + \varepsilon_{i,t}, \tag{3}$$

$$X_{j,t} = \alpha_{j|i} + \gamma_{j|i} M_{t-1} + \beta_{j|i} X_{i,t} + \varepsilon_{j|i,t}, \tag{4}$$

$F^{-1}_{\varepsilon_{i,t}}(\tau|M_{t-1}) = 0$ and $F^{-1}_{\varepsilon_{j|i,t}}(\tau|M_{t-1}, X_{i,t}) = 0$ are assumed. AB propose, in the first step, to determine VaR of an institution i by applying quantile (tail event) regression of log return of company i on macro state variables. The $\beta_{j|i}$ in (4) has standard linear regression interpretation, i.e. it determines the sensitivity of log return of an institution j to changes in tail event log return of an institution i. In the second step the CoVaR is calculated by plugging in VaR of company i at level τ estimated in (5) into the equation (6):

$$\widehat{\mathrm{VaR}}_{i,t,\tau} = \hat{\alpha}_i + \hat{\gamma}_i M_{t-1}, \tag{5}$$

$$\widehat{\mathrm{CoVaR}}^{AB}_{j|i,t,\tau} = \hat{\alpha}_{j|i} + \hat{\gamma}_{j|i} M_{t-1} + \hat{\beta}_{j|i} \widehat{\mathrm{VaR}}_{i,t,\tau}. \tag{6}$$

Thus, the risk of a financial institution j is calculated via a macro state and a VaR of an institution i. Here the coefficient $\widehat{\beta}_{j|i}$ of (6) reflects the degree of interconnectedness. By setting j to be the return of a system, e.g. value-weighted average return on a financial index, and i to be the return of a financial company i, we obtain the *contribution* CoVaR which characterizes how a company i influences the rest of the financial system. By doing the reverse, i.e. by setting j equal to a financial institution and i to a financial system, one obtains *exposure* CoVaR, i.e. the extent to which a single institution is exposed to the overall risk of a system.

This approach allows us to identify the key elements of systemic risk, namely, network effects, a single institution's contribution to systemic risk and a single institution's exposure to systemic risk. In our models, we expand AB's method in three aspects. First of all, AB perform pairwise quantile regression, since two companies are not interacting in an isolated environment, all other interaction effects need to be considered. This motivates us to extend this bivariate model to a (ultra)high dimensional setting by including more variables into the analysis, hence a variable selection should be carried out. Secondly, a linear relationship between system return and a single institution's return is assumed by AB. Hautsch et al. (2015) apply a linear LASSO based variable selection to select variables to estimate the VaR of the system. We enhance their methodologies by employing the nonlinear models because of the complexity of the financial system. The flexible SIM will be implemented to allow the nonlinear relationship in this case. Thirdly, AB use average market valued asset returns weighted by lagged market valued total assets to calculate the system return, as they point out it may create mechanical correlation between a single financial institution and the value-weighted financial index. Instead of the regression on system return, we proposed two market capitalization weighted indices which combine the connectedness structure of the companies: the index of Systemic Risk Receiver and the index of Systemic Risk Emitter, to measure the systemic risk contributions, and further to identify the systemically important financial institutions.

2.2. Step 1 VaR estimation

TENET can be illustrated by three steps. In the first step we estimate VaR for each financial institution by using linear quantile regression as in AB:

$$X_{i,t} = \alpha_i + \gamma_i M_{t-1} + \varepsilon_{i,t}, \tag{7}$$

$$\widehat{\mathrm{VaR}}_{i,t,\tau} = \hat{\alpha}_i + \hat{\gamma}_i M_{t-1}, \tag{8}$$

$X_{i,t}$ and M_{t-1} are defined as in Section 2.1. Note that the VaR is estimated by the linear quantile regression (7) of log return of an institution i on macro state variables. This is justified by the analysis of Chao et al. (2015), who found evidence of linear effects in regressing $X_{i,t}$ on M_{t-1}.

2.3. Step 2 network analysis

2.3.1. Connectedness analysis

In this step, TENET builds up a risk interdependence network based on SIM for quantile regression with variable selection. Note that our model can be easily extended to the case of expectiles, which provide coherent risk measures. First the basic element of the network: CoVaR calculation has to be determined. As in Eq. (2), X_j represents a single institution, and the CoVaR of institution j is estimated by conditioning on its information set. This information set will not only include the asset returns of other firms estimated and the macro variables used in the previous step, but also uses control variables on internal factors of institution j, i.e. the company specific characteristics such as leverage, maturity mismatch, market-to-book and size. This setting will allow us to model the risk spillover channels among institutions mostly caused by liquidity or risk exposure. Our choice of information set is more comprehensive than AB, and a similar motivation can be found in Hautsch et al. (2015). Further, a systemic risk network is built motivated by Diebold and Yilmaz (2014). TENET captures nonlinear dependency as it is based on a SIM quantile variable selection technique. See Appendix A for more details of the statistical methodology. More precisely:

$$X_{j,t} = g(\beta^{\top}_{j|R_j} R_{j,t}) + \varepsilon_{j,t}, \tag{9}$$

$$\widehat{\mathrm{CoVaR}}^{TENET}_{j|\widetilde{R}_j,t,\tau} \overset{\mathrm{def}}{=} \widehat{g}(\widehat{\beta}^{\top}_{j|\widetilde{R}_j} \widetilde{R}_{j,t}), \tag{10}$$

$$\widehat{D}_{j|\widetilde{R}_j} \overset{\mathrm{def}}{=} \frac{\partial \widehat{g}(\widehat{\beta}^{\top}_{j|R_j} R_{j,t})}{\partial R_{j,t}}|_{R_{j,t}=\widetilde{R}_{j,t}} = \widehat{g}'(\widehat{\beta}^{\top}_{j|\widetilde{R}_j} \widetilde{R}_{j,t}) \widehat{\beta}_{j|\widetilde{R}_j}. \tag{11}$$

Here $R_{j,t} \overset{\mathrm{def}}{=} \{X_{-j,t}, M_{t-1}, B_{j,t-1}\}$ is the information set which includes p variables, $X_{-j,t} \overset{\mathrm{def}}{=} \{X_{1,t}, X_{2,t}, \cdots, X_{k,t}\}$ are the explanatory variables including the log returns of all financial institutions except for a financial institution j, k represents the number of financial institutions. $B_{j,t-1}$ are the firm characteristics calculated from their balance sheet information. Define the parameters as $\beta_{j|R_j} \overset{\mathrm{def}}{=} \{\beta_{j|-j}, \beta_{j|M}, \beta_{j|B_j}\}^{\top}$. Note that there is no time symbol t in the parameters, since our model is set up based on one fixed window estimation, we can then apply moving window estimation to estimate all parameters in different windows. We define $\widetilde{R}_{j,t} \overset{\mathrm{def}}{=} \{\widehat{VaR}_{-j,t,\tau}, M_{t-1}, B_{j,t-1}\}$, $\widehat{\mathrm{VaR}}_{-j,t,\tau}$ as the estimated VaRs from (8) for financial institutions except for j in step 1, and $\widehat{\beta}_{j|\widetilde{R}_j} \overset{\mathrm{def}}{=} \{\widehat{\beta}_{j|-j}, \widehat{\beta}_{j|M}, \widehat{\beta}_{j|B_j}\}^{\top}$. As in equation (10) CoVaR comprises of not only the influences of financial institutions except for j, but also incorporates non-linearity reflected in the shape of a link function $g(\cdot)$. Therefore, we name it $\widehat{\mathrm{CoVaR}}^{TENET}$ which stands for Tail-Event driven NETwork risk with SIM model.[1] $\widehat{D}_{j|\widetilde{R}_j}$ is the gradient measuring the marginal effect of covariates evaluated at $R_{j,t} = \widetilde{R}_{j,t}$, and the componentwise expression is $\widehat{D}_{j|\widetilde{R}_j} \overset{\mathrm{def}}{=} \{\widehat{D}_{j|-j}, \widehat{D}_{j|M}, \widehat{D}_{j|B_j}\}^{\top}$. In particular, $\widehat{D}_{j|-j}$ allows to measure spillover effects across the financial

[1] For simplicity we omit the subscript $_{j|\widetilde{R}_j,t,\tau}$ in $\widehat{\mathrm{CoVaR}}^{TENET}_{j|\widetilde{R}_j,t,\tau}$, and write $\widehat{\mathrm{CoVaR}}^{TENET}$.

Table 1
A $k \times k$ adjacency matrix for financial institutions at the sth window.

$$A_s = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ \vdots \\ I_k \end{array} \begin{array}{c} \begin{array}{ccccc} I_1 & I_2 & I_3 & \cdots & I_k \end{array} \\ \begin{pmatrix} 0 & |\widehat{D}^s_{1|2}| & |\widehat{D}^s_{1|3}| & \cdots & |\widehat{D}^s_{1|k}| \\ |\widehat{D}^s_{2|1}| & 0 & |\widehat{D}^s_{2|3}| & \cdots & |\widehat{D}^s_{2|k}| \\ |\widehat{D}^s_{3|1}| & |\widehat{D}^s_{3|2}| & 0 & \cdots & |\widehat{D}^s_{3|k}| \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ |\widehat{D}^s_{k|1}| & |\widehat{D}^s_{k|2}| & |\widehat{D}^s_{k|3}| & \cdots & 0 \end{pmatrix} \end{array}$$

institutions and to characterize their evolution as a system represented by a network. Note that in our network analysis we only include the partial derivatives of institution j with respect to the other financial institutions (i.e. $\widehat{D}_{j|-j}$). The partial derivatives with respect to institution's characteristic variables $\widehat{D}_{j|B_j}$ and macro state variables $\widehat{D}_{j|M}$ are not included. The reason is that we concentrate on spillover effects among firms in the network analysis.

The term *network* refers to a (directed) *graph* with a set of vertices and a set of links, or edges. We summarize the estimation results in a form of a weighted adjacency matrix. Let $\widehat{D}^s_{j|i}$ be one element in $\widehat{D}^s_{j|-j}$ at estimation window s, where j represents one financial institution as before, i stands for another institution which is one element in the other financial institutions set $-j$. Then a weighted adjacency matrix contains absolute values of $\widehat{D}^s_{j|i}$ (in upper triangular matrix) and $\widehat{D}^s_{i|j}$ (in lower triangular matrix), where $\widehat{D}^s_{j|i}$ is the impact from firm i to firm j and $\widehat{D}^s_{i|j}$ means the impact from firm j to firm i. Table 1 shows the adjacency matrix; note that in each window of estimation one has only one adjacency matrix estimated.

The above $k \times k$ matrix A_s in Table 1 represents total connectedness across variables at window s, and I_i represents the name of financial institution i. The adjacency matrix, or a total connectedness matrix, is sparse and off-diagonal since our model (by construction) does not allow for self-loop effects (namely one variable cannot be regressed on itself). The rows of this matrix correspond to incoming edges for a variable in a respective row and the columns correspond to outgoing edges for a variable in a respective column.

2.3.2. Spectral clustering

In this section, we apply spectral clustering technique, see Shi and Malik (2000), to detect the time varying risk clusters. The weighted adjacency matrix at window s is A_s, the corresponding unweighted matrix is defined by A^u_s, which means that non-zero values in A_s are set to be 1s, and zeros are still 0s. We take the symmetrized adjacency matrix $A^{u\top}_s A^u_s$, and the corresponding degree matrix Γ^2_s(a diagonal matrix with diagonal elements as row(column) sums). The spectral clustering algorithm is launched by looking at the eigenvalues and eigenvectors of the normalized Laplacian matrix $\Gamma^{-1}_s A^{u\top}_s A^u_s \Gamma^{-1}_s$. We would like to identify for each window risk clusters of financial institutions.

2.4. Step 3: Identification of systemic risk contributions

In the third step, TENET explains systemic risk measures. We define two indices to identify systemically important financial institutions. The idea is that we would like to measure the systemic risk relevance of a specific firm by its total in and out connections weighted by market capitalization.

The Systemic Risk Receiver Index for a firm j is therefore defined as:

$$SRR_{j,s} \stackrel{\text{def}}{=} MC_{j,s} \left\{ \sum_{i \in k^{IN}_s} (|\widehat{D}^s_{j|i}| \cdot MC_{i,s}) \right\}, \tag{12}$$

the Systemic Risk Emitter Index for a firm j is defined as:

$$SRE_{j,s} \stackrel{\text{def}}{=} MC_{j,s} \left\{ \sum_{i \in k^{OUT}_s} (|\widehat{D}^s_{i|j}| \cdot MC_{i,s}) \right\}. \tag{13}$$

where k^{IN}_s and k^{OUT}_s are the sets of firms connected with firm j by incoming and outgoing links at window s respectively, and $MC_{i,s}$ represents the market capitalization of firm i at the starting point of window s. $|\widehat{D}^s_{j|i}|$ and $|\widehat{D}^s_{i|j}|$ are absolute partial derivatives derived from (11) which represent row (incoming) and column (outgoing) direction connectedness of firm j as in Table 1. Thus both $SRR_{j,s}$ and $SRE_{j,s}$ would take into account the firm j's and its connected firms' market capitalization as well as its connectedness within our network.

3. Results

3.1. Data

Since the SIC code can be applied to classify the industries, according to the company list 2012 of US financial institutions from the NASDAQ webpage and the corresponding four-digit SIC codes from 6000 to 6799 for these financial institutions in COMPUSTAT database, we divide the US financial institutions into four groups: (1) depositories (6000–6099), (2) insurance companies (6300–6499), (3) broker–dealers (6200–6231), (4) others (the rest codes). For instance, the Goldman Sachs Group is classified as broker–dealers based on its SIC code 6211. We select top 25 institutions in each group according to the ranking of their market capitalization (like Billio et al. (2012) they apply a similar selection method), so that we can compare the difference among industry groups. Our analysis focuses on the panel of these 100 publicly traded US financial institutions between 5 January, 2007 and 4 January, 2013, see Table 2 for a complete list. The weekly price data are available in Yahoo Finance.[2]

To capture the company specific characteristics we adopt the following variables calculated from balance sheet information as proposed in AB: 1. leverage, defined as total assets/total equity (in book values); 2. maturity mismatch, calculated by (short term debt–cash)/total liabilities; 3. market-to-book, defined as the ratio of the market to the book value of total equity; 4. size, calculated by the log of total book equity. The quarterly balance sheet information is available on the COMPUSTAT database, and cubic interpolation is implemented in order to obtain the weekly data.

Apart from the data on the financial companies we use weekly observations of macro state variables which characterize the general state of the economy. These variables are defined as follows: (i) the implied volatility index, VIX, reported by the Chicago Board Options Exchange; (ii) short term liquidity spread denoted as the difference between the three-month repo rate (available on the Bloomberg database) and the three-month bill rate (from Federal Reserve Board) to measure short-term liquidity risk; (iii) the changes in the three-month Treasury bill rate from the Federal Reserve Board; (iv) the changes in the slope of the yield curve corresponding to the yield spread between the ten-year Treasury rate and the three-month bill rate from the Federal Reserve Board; (v) the changes in the credit spread between BAA-rated bonds and the Treasury rate from the Federal Reserve Board; (vi) the weekly S&P500 index returns from Yahoo finance, and (vii) the weekly Dow Jones US Real Estate index returns from Yahoo finance.

[2] We appreciate Mr. Lukas Borke, who is a doctoral student in LvB Chair of Statistics, with the help for optimizing our code and downloading data.

Table 2
Financial companies with tickers classified by industry: depositories (25), insurance (25), broker–dealers (25) and others (25).

	Depositories (25)		Insurances (25)
WFC	Wells Fargo & Company	AIG	American International Group, Inc.
JPM	J P Morgan Chase & Co	MET	MetLife, Inc.
BAC	Bank of America Corporation	TRV	The Travelers Companies, Inc.
C	Citigroup Inc.	AFL	Aflac Incorporated
USB	US Bancorp	PRU	Prudential Financial, Inc.
COF	Capital One Financial Corporation	CB	Chubb Corporation (The)
PNC	PNC Financial Services Group, Inc. (The)	MMC	Marsh & McLennan Companies, Inc.
BK	Bank Of New York Mellon Corporation (The)	ALL	Allstate Corporation (The)
STT	State Street Corporation	AON	Aon plc
BBT	BB&T Corporation	L	Loews Corporation
STI	SunTrust Banks, Inc.	PGR	Progressive Corporation (The)
FITB	Fifth Third Bancorp	HIG	Hartford Financial Services Group, Inc. (The)
MTB	M&T Bank Corporation	PFG	Principal Financial Group Inc
NTRS	Northern Trust Corporation	CNA	CNA Financial Corporation
RF	Regions Financial Corporation	LNC	Lincoln National Corporation
KEY	KeyCorp	CINF	Cincinnati Financial Corporation
CMA	Comerica Incorporated	Y	Alleghany Corporation
HBAN	Huntington Bancshares Incorporated	UNM	Unum Group
HCBK	Hudson City Bancorp, Inc.	WRB	W.R. Berkley Corporation
PBCT	People's United Financial, Inc.	FNF	Fidelity National Financial, Inc.
BOKF	BOK Financial Corporation	TMK	Torchmark Corporation
ZION	Zions Bancorporation	MKL	Markel Corporation
CFR	Cullen/Frost Bankers, Inc.	AJG	Arthur J. Gallagher & Co.
CBSH	Commerce Bancshares, Inc.	BRO	Brown & Brown, Inc.
SBNY	Signature Bank	HCC	HCC Insurance Holdings, Inc.
	Broker–Dealers (25)		Others (25)
GS	Goldman Sachs Group, Inc. (The)	AXP	American Express Company
BLK	BlackRock, Inc.	BEN	Franklin Resources, Inc.
MS	Morgan Stanley	CBG	CBRE Group, Inc.
CME	CME Group Inc.	IVZ	Invesco Plc
SCHW	The Charles Schwab Corporation	JLL	Jones Lang LaSalle Incorporated
TROW	T. Rowe Price Group, Inc.	AMG	Affiliated Managers Group, Inc.
AMTD	TD Ameritrade Holding Corporation	OCN	Ocwen Financial Corporation
RJF	Raymond James Financial, Inc.	EV	Eaton Vance Corporation
SEIC	SEI Investments Company	LM	Legg Mason, Inc.
NDAQ	The NASDAQ OMX Group, Inc.	CACC	Credit Acceptance Corporation
WDR	Waddell & Reed Financial, Inc.	FII	Federated Investors, Inc.
SF	Stifel Financial Corporation	AB	Alliance Capital Management Holding L.P.
GBL	Gamco Investors, Inc.	PRAA	Portfolio Recovery Associates, Inc.
MKTX	MarketAxess Holdings, Inc.	JNS	Janus Capital Group, Inc
EEFT	Euronet Worldwide, Inc.	NNI	Nelnet, Inc.
WETF	WisdomTree Investments, Inc.	WRLD	World Acceptance Corporation
DLLR	DFC Global Corp	ECPG	Encore Capital Group Inc
BGCP	BGC Partners, Inc.	NEWS	NewStar Financial, Inc.
PJC	Piper Jaffray Companies	AGM	Federal Agricultural Mortgage Corporation
ITG	Investment Technology Group, Inc.	WHG	Westwood Holdings Group Inc
INTL	INTL FCStone Inc.	AVHI	AV Homes, Inc.
GFIG	GFI Group Inc.	SFE	Safeguard Scientifics, Inc.
LTS	Ladenburg Thalmann Financial Services Inc	ATAX	America First Tax Exempt Investors, L.P.
OPY	Oppenheimer Holdings, Inc.	TAXI	Medallion Financial Corp.
CLMS	Calamos Asset Management, Inc.	NICK	Nicholas Financial, Inc.

3.2. Estimation results

We perform the TENET analysis in three steps: Firstly, the Tail Event VaR of all firms are estimated. Secondly, the NETwork analysis based on the SIM with variable selection technique is performed. Finally, the systemically important financial institutions are identified based on the *SRR*, *SRE* indices defined in Section 2.4.

To estimate VaR as in (7) and (8), we regress weekly log returns of each institution on macro state variables at the quantile level $\tau = 0.05$, with the whole period being $T = 266$, the number of independent variables is $p = 110$ (e.g. when JP Morgan is dependent variable, then the independent variables include 4 firm characteristics of JP Morgan, 99 other firms' returns and 7 macro state variables), and the rolling window size is set to be $n = 48$ corresponding to one year's weekly data. (We choose a small window size for the stationarity of the data process, and our methodology allows to work with the setting $p > n$. We acknowledge that by choosing a larger window size, and different data frequencies, the results may vary. We leave it as a further research topic to study what is an optimal window size and data frequency in this context.)

Fig. 1 is an example of estimated VaR (the thinner red line) for J P Morgan (with the SIC code 6020). In the second step a CoVaR based risk network is estimated by applying the SIM with variable selection, see (20) in Appendix A. Fig. 1 shows the $\widehat{\text{CoVaR}}^{TENET}$ (the thicker blue line) of J P Morgan. Then the network analysis induced by the $\widehat{\text{CoVaR}}^{TENET}$ is shown from Fig. 2 to Fig. 6. Recall that the adjacency matrix of Table 1 is constructed from $|\widehat{D}^s_{j|i}|$ and $|\widehat{D}^s_{i|j}|$. To aggregate the results over windows, we take the component-wise sum of the adjacency matrices. With the aggregation we will be able to understand the risk channels and the relative role of each firm or each sector in the whole financial network.

For this propose, we define three levels of connectedness: the overall level, the group level and the firm level connectedness. The overall level of risk is characterized by the total connectedness of the system and the averaged value of the tuning parameter λ. The

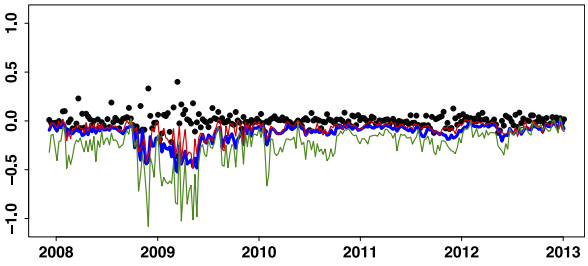


Fig. 1. Log return of J P Morgan (black points), $\widehat{\text{VaR}}$ (thinner red line), $\widehat{\text{CoVaR}}^{TENET}$ (thicker blue line), and $\widehat{\text{CoVaR}}^{L}$ (thinner green line) for J P Morgan, $\tau = 0.05$, window size $n = 48$, $T = 266$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

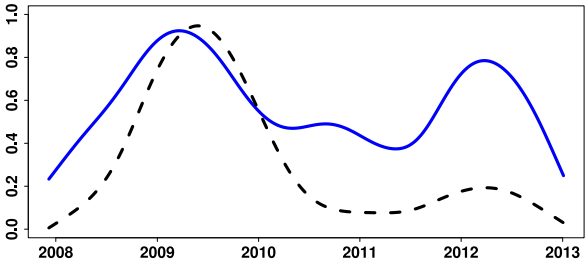


Fig. 2. Total connectedness (solid blue line) and average lambda (dashed black line) of 100 financial institutions from 20071207 to 20130105, $\tau = 0.05$, window size $n = 48, T = 266$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

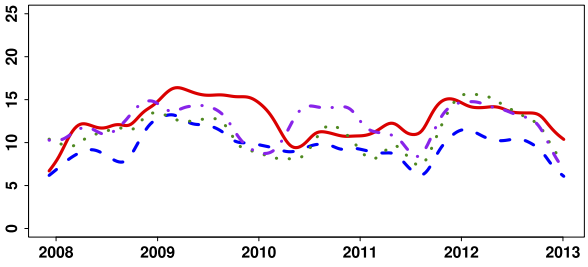


Fig. 3. Incoming links for four industry groups. Depositories: solid red line, Insurances: dashed blue line, Broker–Dealers: dotted green line, Others: dash–dot violet line. $\tau = 0.05$, window size $n = 48$, $T = 266$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

total connectedness of links is defined as $TC_s = TC_s^{IN} = TC_s^{OUT} \stackrel{\text{def}}{=} \sum_{i=1}^{k}\sum_{j=1}^{k} |\widehat{D}^s_{j|i}|$, where TC_s^{IN} and TC_s^{OUT} are the total incoming and outgoing links in this matrix respectively. The solid line of Fig. 2 shows the evolution of the total connectedness, and the dashed line of Fig. 2 shows the averaged λ values of the CoVaR estimations by using SIM with variable selection, where λ is the estimated penalization parameter, see Appendix A.

While at the beginning of 2008 there was lower connectedness and smaller averaged λ, from the second quarter of 2008 both connectedness and averaged λ began to increase sharply which corresponds to the bankruptcy of Bear Stearns and Lehman Brothers. As the crisis was unfolding, the system became more heavily interconnected and reached its peak at the beginning of 2009, the averaged λ stayed at peak level in the middle of 2009, which can be seen as the influence of the European sovereign debt crisis. Then the downward trend dominated the whole market, and lasted until end of 2010, the financial institutions were most least connected to each others in second quarter of 2011. From the third quarter of 2011 the averaged λ began to increase and lasted until beginning of 2012 which can be attributed to the impact of the US debt-ceiling crisis in July 2011. Total connectedness series

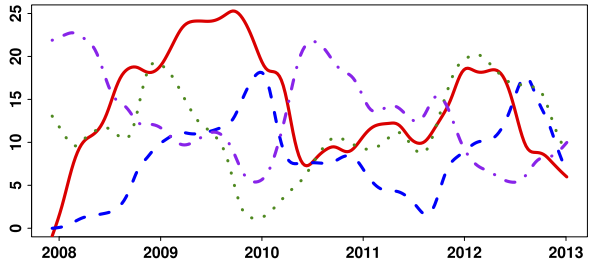


Fig. 4. Outgoing links for four industry groups. Depositories: solid red line, Insurances: dashed blue line, Broker–Dealers: dotted green line, Others: dash–dot violet line. $\tau = 0.05$, window size $n = 48$, $T = 266$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

increased again in second quarter of 2011. After the middle of 2012, both the averaged λ and the total connectedness series decreased. Since the evolution of averaged λ represents the variation of the systemic risk, Borke et al. (forthcoming) propose a Financial Risk Meter (FRM): http://sfb649.wiwi.hu-berlin.de/frm/index.html.

The group connectedness with respect to incoming links is defined as follows: $GC^{IN}_{g,s} \stackrel{\text{def}}{=} \sum_{i=1}^{k}\sum_{j\in g} |\widehat{D}^s_{j|i}|$, where $g = 1, 2, 3, 4$ correspond to the four aforementioned industry groups. The group connectedness with respect to outgoing links is defined as $GC^{OUT}_{g,s} \stackrel{\text{def}}{=} \sum_{i\in g}\sum_{j=1}^{k} |\widehat{D}^s_{j|i}|$. Fig. 3 shows the incoming links for these four groups. The patterns of these four groups are almost identical, i.e. there are more links during the end of 2008 and beginning of 2010, during the middle of 2011 and the end of 2012. Only for group "others", there are even more links between 2010 and 2012, this maybe because the heterogeneity of this group: AXP (American Express Company) is a credit card company, JLL (Jones Lang LaSalle Incorporated), CBG (CBRE Group, Inc) and AVHI (AV Homes, Inc.) are real estate firms, BEN (Franklin Resources, Inc.), IVZ (Invesco Plc) and AMG (Affiliated Managers Group) are investment management companies, whereas OCN (Ocwen Financial Corporation) and AGM (Federal Agricultural Mortgage Corporation) are mortgage loan companies. While the depositories group (solid line) received on average more risk than the other three groups, the insurance companies (dashed line) are less influenced by the financial crisis. This can be seen as evidence supporting the report of Systemic Risk in Insurance – An analysis of insurance and financial stability published by Geneva Association in 2010 stating that losses in the insurance industry have been only a sixth of those at banks. In contrast to the incoming links the outgoing links in Fig. 4 are more volatile. It is not surprising that the depositories sector dominates the others in the outgoing links, i.e. the bank group emits more risk to the system than the other groups. Broker–dealers and others fluctuate very much in the whole period, but they send out less risk compared with banks. And the insurers emit averagely less risk over all periods than the other groups.

Next we turn to analysing firm level interconnectedness. Firstly we focus on the directional connectedness from firm i to the firm j which is defined as follows: $DC^s_{j|i} \stackrel{\text{def}}{=} |\widehat{D}^s_{j|i}|$. The network in Fig. 5 shows one example of the firm level directional connectedness on 12 June 2009 which was in the financial crisis. There are several links emitted from C (Citigroup) in upper red ellipse and MS (Morgan Stanley) in lower green ellipse. To make the major connections more clearly, we apply a hard thresholding to omit the small values. That is, the values of absolute derivatives smaller than the average of the 100 largest absolute partial derivatives are set to be zeros. Fig. 6 is the network after the thresholding. We see that there are several strong connections, for example, in left violet ellipse the link from JLL to CBG (as we stated before they are both real estate companies, the connection induced by

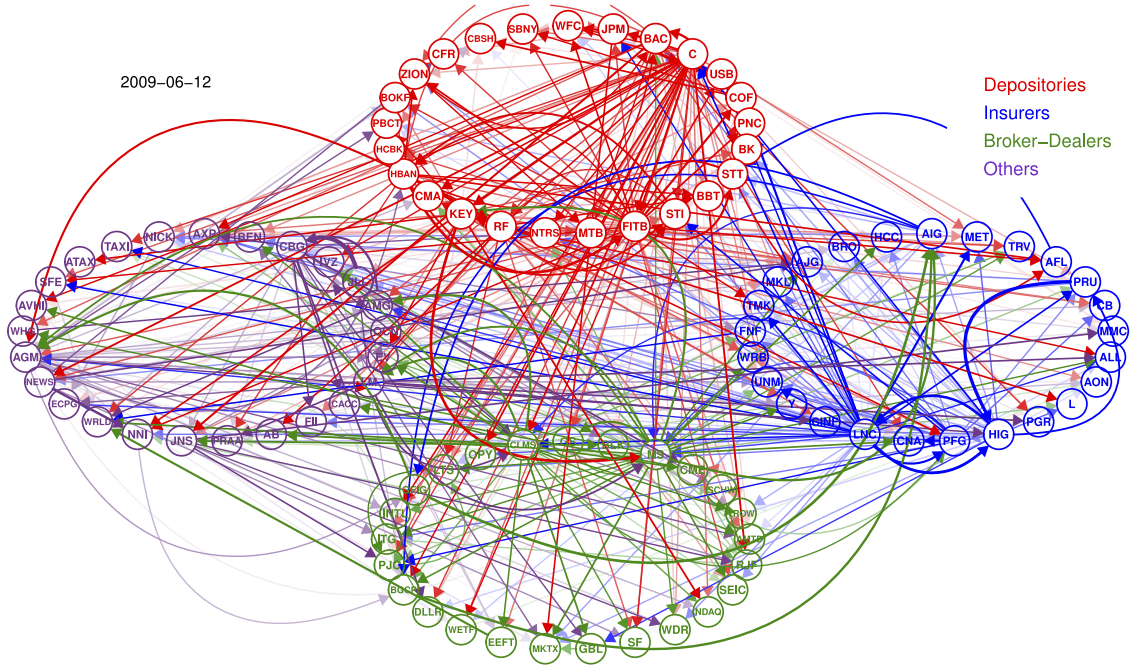


Fig. 5. A elliptical network representation of a weighted adjacency matrix without the thresholding. Depositories: clockwise 25 firms from WFC to SBNY (upper red), Insurance: clockwise 25 firms from AIG to HCC (right blue), Broker–Dealers: clockwise 25 firms from GS to CLMS (lower green), Others: clockwise 25 firms from AXP to NICK (left violet), date: 20090612, $\tau = 0.05$, window size $n = 48$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

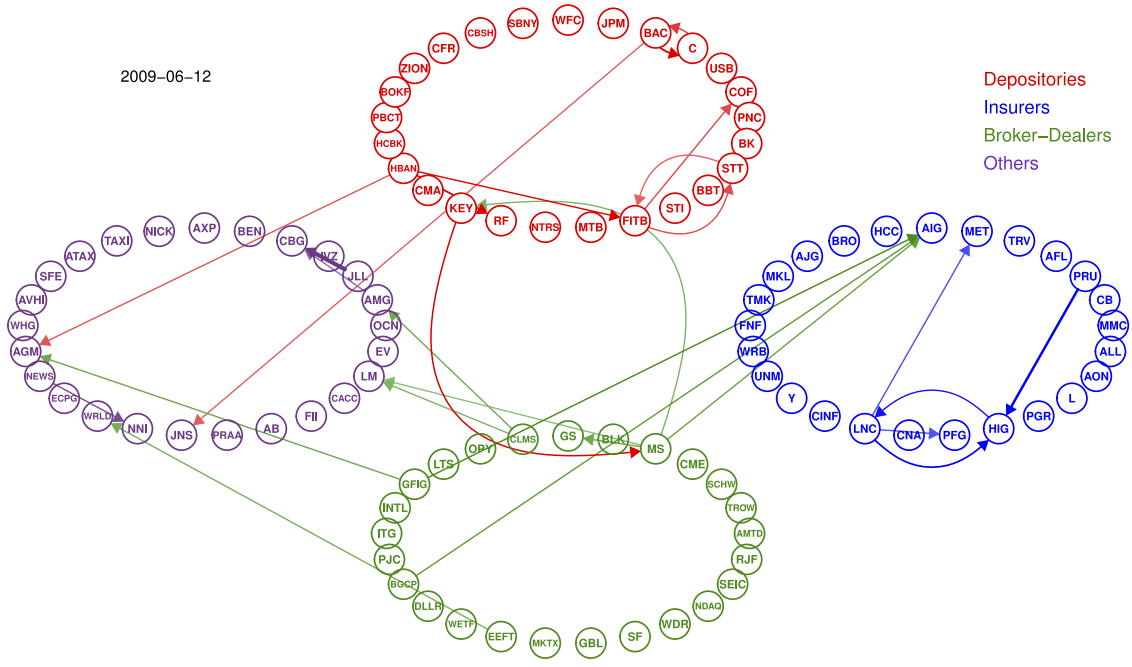


Fig. 6. A elliptical network representation of a weighted adjacency matrix after the thresholding (the values smaller than average of first 100 largest partial derivatives are set to be 0s). Depositories: clockwise 25 firms from WFC to SBNY (upper red), Insurance: clockwise 25 firms from AIG to HCC (right blue), Broker–Dealers: clockwise 25 firms from GS to CLMS (lower green), Others: clockwise 25 firms from AXP to NICK (left violet), date: 20090612, $\tau = 0.05$, window size $n = 48$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

spillover effects seems reasonable), and in right blue ellipse from PRU (Prudential Financial, Inc) to HIG (Hartford Financial Services Group), note that they are both insurances. Moreover there are also a couple of weak connections from MS to others. Furthermore, there are a lot of mutual connections, big banks like BAC (Bank of America) and C in upper red ellipse, STT (State Street Corporation) and FITB (Fifth Third Bancorp) in upper red ellipse, insurances: LNC (Lincoln National Corporation) and HIG in right blue ellipse, different groups, e.g. MS (a broker dealer) and KEY (KeyCorp, a big bank). We aggregate the directional connectedness by the sum of absolute value of $\widehat{D}^s_{j|i}$ and $\widehat{D}^s_{i|j}$ over $T = 266$ windows. The results for individual firm can be found in Table 3. For WFC (Wells Fargo) the strong incoming links come from STI (SunTrust Banks), C and BAC, the outgoing links go to USB (US Bancorp), STI and CBSH (Commerce Bancshares). We also see some pairs of mutual interacting firms, like BAC and C, AIG (American International

Table 3
Top 15 firms ranked according to market capitalization (MC) in the 100 company list, the received links from other firms and transmitted links to other firms are shown correspondingly. Note that only the first three most influential firms are listed for each ticker, $n = 48$, $T = 266$.

Rank	Ticker	$\tau = .05$		$\tau = 0.95$	
		Received link from	Transmitted link to	Received link from	Transmitted link to
1	WFC	STI, C, BAC	USB, STI, CBSH	C, BAC, HIG	SBNY, PNC, USB
2	JPM	C, MS, COF	C, CFR, MKTX	C, MS, BAC	MS, BAC, GS
3	BAC	C, MS, RF	C, JNS, WFC	C, ZION, FITB	ZION, WFC, JPM
4	C	BAC, LNC, MS	BAC, AIG, JPM	LNC, MS, NEWS	AIG, BAC, LNC
5	AXP	COF, C, WRLD	OCN, WRLD, MMC	C, COF, JNS	WRLD, MKTX, FNF
6	USB	LNC, RF, STI	WRLD, JPM, PNC	FITB, COF, ZION	PNC, JPM, HCBK
7	GS	MS, C, JNS	WETF, MS, WDR	MS, C, CBG	MS, WHG, WDR
8	AIG	C, LNC, MS	C, GFIG, MS	C, AGM, LNC	AGM, C, NEWS
9	MET	LNC, HIG, C	LNC, CNA, MKL	LNC, HIG, C	HIG, PRU, AFL
10	COF	C, FITB, ZION	AXP, JPM, CBSH	FITB, C, LNC	AXP, PNC, BAC
11	BLK	FNF, MKTX, EEFT	CLMS, C, JNS	CBG, C, JNS	STT, NDAQ, MKTX
12	MS	C, KEY, AIG	GS, AIG, BAC	C, HIG, KEY	GS, PJC, C
13	PNC	C, HBAN, LNC	USB, HCBK, BK	C, COF, ZION	TROW, BK, USB
14	BK	MS, ZION, C	MMC, NTRS, WETF	C, JNS, LNC	STT, WRB, NEWS
15	BEN	LNC, JNS, CLMS	WRB, WRLD, MMC	JNS, CLMS, LNC	CLMS, FNF, BRO

Table 4
Top 10 directional connectedness from one financial institution to another. The ranking is calculated by the sum of absolute value of the partial derivatives, $\tau = 0.05$, window size $n = 48$, $T = 266$.

Rank	From Ticker	To Ticker	Sum
1	JLL (Jones Lang LaSalle)	CBG (CBRE Group)	140.39
2	CBG (CBRE Group)	JLL (Jones Lang LaSalle)	116.86
3	LNC (Lincoln National Corp.)	PFG (Principal Financial Group)	96.78
4	PFG (Principal Financial Group)	LNC (Lincoln National Corp.)	90.43
5	C (Citigroup)	AIG (American International Group)	82.03
6	JNS (Janus Capital Group)	WDR (Waddell & Reed Financial)	65.75
7	RF (Regions Financial)	HBAN (Huntington Bancshares)	60.86
8	STI (SunTrust Banks)	FITB (Fifth Third Bancorp.)	57.95
9	LNC (Lincoln National Corp.)	MET (MetLife)	57.35
10	MS (Morgan Stanley)	GS (Goldman Sachs Group)	55.98

Table 5
Top 10 financial institutions ranked according to Incoming links calculated by the sum of absolute value of the partial derivatives, and the rank of market capitalization (MC) in this 100 financial institutions list in 2012 is also shown in this table, $\tau = 0.05$, window size $n = 48$, $T = 266$.

Rank	Ticker of IN	IN-Sum	Rank of MC (Value)
1	AGM (Federal Agricultural Mortgage)	235.55	89 (3.52E+08)
2	AIG (American International Group)	230.46	8 (4.82E+10)
3	HIG (Hartford Financial Services Group)	225.46	37 (9.24E+09)
4	CBG (CBRE Group)	221.86	32 (1.28E+10)
5	FITB (Fifth Third Bancorp)	202.00	30 (1.31E+10)
6	STI (SunTrust Banks)	199.85	29 (1.44E+10)
7	HBAN (Huntington Bancshares)	196.29	51 (5.23E+09)
8	BAC (Bank of America Corp.)	192.11	3 (1.05E+11)
9	C (Citigroup)	191.50	3 (1.05E+11)
10	LNC (Lincoln National Corp.)	189.59	43 (6.67E+09)

Table 6
Top 10 financial institutions ranked according to Outgoing links calculated by the sum of absolute value of the partial derivatives, and the rank of market capitalization (MC) in this 100 financial institutions list in 2012 is also shown in this table, $\tau = 0.05$, window size $n = 48$, $T = 266$.

Rank	Ticker of OUT	OUT-Sum	Rank of MC (Value)
1	LNC (Lincoln National Corp.)	1129.38	43 (6.67E+09)
2	C (Citigroup)	1097.93	3 (1.05E+11)
3	MS (Morgan Stanley)	626.12	37 (9.24E+09)
4	CBG (CBRE Group)	597.83	32 (1.28E+10)
5	RF (Regions Financial)	568.71	36 (9.30E+09)
6	JNS (Janus Capital Group)	558.06	76 (1.57E+09)
7	CLMS (Calamos Asset Management)	514.80	99 (1.94E+08)
8	HIG (Hartford Financial Services Group)	499.04	37 (9.24E+09)
9	ZION (Zions Bancorp.)	472.18	63 (3.72E+09)
10	AGM (Federal Agricultural Mortgage)	349.11	90 (3.52E+08)

Group) and MS. We show the directional connection in $\tau = 0.95$ case as well, the selected firms are mostly different from $\tau = 0.05$ case, which shows that our method can explain the asymmetric effects on the dependency structure at different price levels. See Table 3 for more details. The ranking of the directional connectedness is calculated by the sum of absolute value of $\widehat{D}^s_{j|i}$ over windows. The first two strongest mutual connections are between JLL and CBG, between LNC and PFG (Principal Financial Group), see Table 4. Secondly, the firm connectedness with respect to incoming links is defined as $FC^{IN}_{j,s} = \sum_{i=1}^{k} |\widehat{D}^s_{j|i}|$. Finally, the

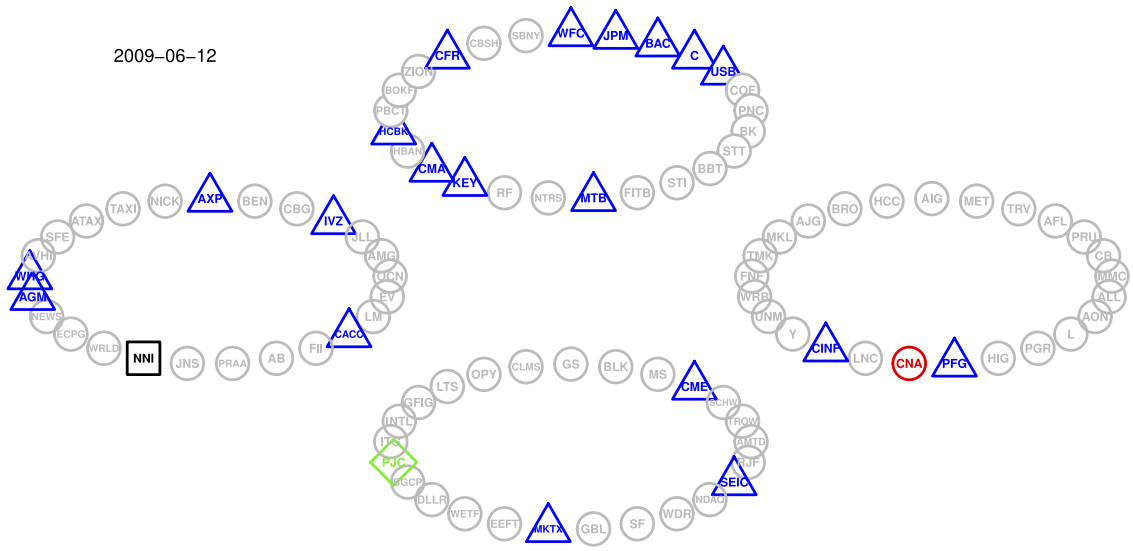


Fig. 7. A elliptical network representation of an unweighted adjacency matrix (1 and 0 representation of this matrix) without thresholding. Green, blue, red, black represent four different risk clusters, and grey represents unconnected firm. Date: 20090612, $\tau = 0.05$, window size $n = 48$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

firm connectedness with respect to outgoing links is: $FC^{OUT}_{j,s} = \sum_{j=1}^{k} |\widehat{D}^{s}_{j|i}|$. From Tables 5 and 6 we have the top 10 firms in terms of incoming links and outgoing links respectively. The most connected firm with incoming links is AGM (Federal Agricultural Mortgage Corporation) and the most connected firm with outgoing links is LNC (Lincoln National Corporation) which is a multiple insurance and investment management company. We have found out that among the top 10 IN-link and OUT-link companies, there are several big firms, such as AIG (American International Group) and BAC (Bank of America Corporation) with IN-link, and C (Citigroup) and MS (Morgan Stanley) with OUT-link. However, there are also firms with moderate or small sizes e.g. AGM and HBAN (Huntington Bancshares Incorporated) with IN-link, and CLMS (Calamos Asset Management) and JNS (Janus Capital Group) with OUT-link. This is connected with the Global Financial Stability Report (GFSR) of April 2009 which states that the crisis has shown that not only the banks but also other non-bank financial intermediaries can be systemically important and their failure can cause destabilizing effects. It also emphasizes that not only the largest financial institutions but also the smaller but interconnected financial institutions are systemically important and need to be regulated. "Too connected to fail" is an important issue. However, we see that small firms tend to have more connections with small firms, such as AGM (market cap $0.35 billion), which is with the largest sum of incoming links coming from GFIG (market cap. $0.29 billion), LTS (market cap. $0.22 billion), NEWS (market cap.$0.62 billion), OPY (market cap.$0.21 billion) and HBAN (market cap. $5.2 billion). Despite the heavy connections in the system, one would still not consider it as highly systemic risk relevant. So we try to account the three factors in the forthcoming systemic risk analysis: (1) a firm is big enough, (2) a firm is highly connected with other firms, (3) the connected firms are relative large in size. Therefore to identify the systemically important financial institutions, we add a weight of market capitalization in the network.

In addition, based on our network analysis we have the following findings: (1) the connections between institutions tend to increase before the financial crisis, (2) the network is characterized by numerous heavy links at the peak of a crisis, (3) the connections between institutions reflected by the absolute value of partial derivatives get weaker as the financial system stabilizes, (4) the incoming links are far less volatile than the outgoing links. Whereas banks dominate both incoming and outgoing links, the insurers are less affected by the financial crisis and exhibit less contribution in terms of risk transmission. The broker–dealer and others are highly volatile with respect to the risk contribution. (5) Several institutions with moderate or small sizes and also some non-bank institutions have higher connectedness, as they are too connected firms. (6) "Too connected" is not a sufficient condition to detect the importance of the firm. To identify the systemically important financial institutions we need a measure which combines the concepts "too connected to fail" and "too big to fail".

While in the first part of step 2 we detect connectedness by applying sum of the absolute derivatives, in the second part of step 2 we classify the risk clusters by using spectral clustering. Figs. 7 and 8 show the risk clusters in window starting on 6 June 2009 (during subprime crisis) and 10 Aug 2012. For Fig. 7, the biggest cluster with blue triangle includes some big banks, like WFC (Wells Fargo), JPM (J P Morgan), BAC (Bank of America), C (Citigroup), USB (US Bancorp), some insurances: PFG (Principal Financial Group) and CINF (Cincinnati Financial Corporation), broker–dealers: CME (CME Group Inc.), SEIC (SEI Investments Company) and MKTX (MarketAxess Holdings), and others like AXP (American Express Company) and IVZ (Invesco Plc). We see that during crisis, WFC (Wells Fargo), JPM (J P Morgan) and C (Citigroup) are very frequently classified into the same cluster. For Fig. 8, we see that the clusters are more widely spreading cross sectors.

In the third step we provide an exact systemic risk measure for each firm based on their connectedness structure. We consider the market capitalization of each firm as well as its connected firms with incoming or outgoing links, see Eqs. (12) and (13). Table 7 shows the ranking of the top 10 calculated Systemic Risk Receivers: JPM (J P Morgan), C (Citigroup), WFC (Wells Fargo), BAC (Bank of America), AIG (American International Group), GS (Goldman Sachs), USB (US Bancorp), MS (Morgan Stanley), AXP (American Express Company) and COF (Capital One Financial Corp.). As for the Systemic Risk Emitters, the corresponding ranking is presented in Table 8. Although the market capitalization of LNC and RF are moderate, they are still ranked in the top 10 largest systemic risk emitters list, as they have many strong outgoing links. Compared with the result of global systemically important banks (G-SIBs)

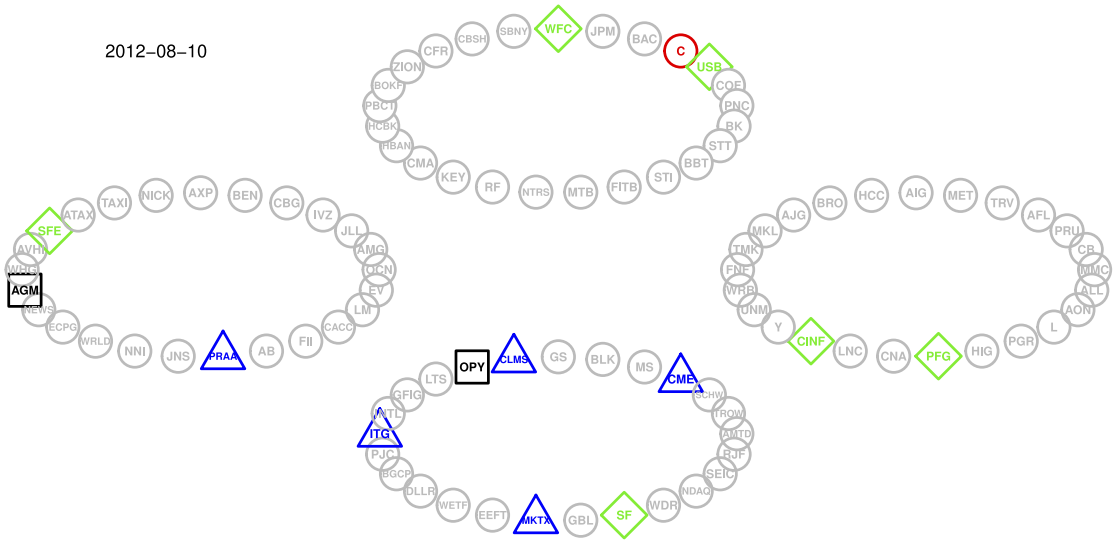


Fig. 8. A elliptical network representation of an unweighted adjacency matrix (1 and 0 representation of this matrix) without thresholding. Green, blue, red, black represent four different risk clusters, and grey represents unconnected firm. Date: 20120810, $\tau = 0.05$, window size $n = 48$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 7
Top 10 financial institutions ranked according to the index of Systemic Risk Receiver (SRR), the rank of market capitalization (MC) and their values (in brackets) of this 100 financial institutions in 2012 are also shown in this table.

Rank	Ticker	SRR	Rank of MC (Value)
1	JPM (J P Morgan Chase & Co)	4.63E+21	2 (1.55E+11)
2	C (Citigroup)	3.13E+21	3 (1.05E+11)
3	WFC (Wells Fargo & Company)	3.03E+21	1 (1.75E+11)
4	BAC (Bank of America)	2.90E+21	3 (1.05E+11)
5	AIG (American International Group)	1.15E+21	8 (4.82E+10)
6	GS (Goldman Sachs Group)	1.00E+21	8 (5.53E+10)
7	USB (US Bancorp)	8.57E+20	6 (6.03E+10)
8	MS (Morgan Stanley)	8.29E+20	12 (3.21E+10)
9	AXP (American Express Company)	7.71E+20	5 (6.26E+10)
10	COF (Capital One Financial Corp.)	6.64E+20	10 (3.39E+10)

Table 8
Top 10 financial institutions ranked according to the index of Systemic Risk Emitter (SRE), the rank of market capitalization (MC) and their values (in brackets) of this 100 financial institutions in 2012 are also shown in this table.

Rank	Ticker	SRE	Rank of MC (Value)
1	C (Citigroup)	1.18E+22	3 (1.05E+11)
2	BAC (Bank of America)	3.89E+21	3 (1.05E+11)
3	MS (Morgan Stanley)	2.11E+21	12 (3.21E+10)
4	WFC (Wells Fargo & Company)	1.37E+21	1 (1.75E+11)
5	AIG (American International Group)	7.01E+20	8 (4.82E+10)
6	COF (Capital One Financial Corp.)	6.18E+20	10 (3.39E+10)
7	LNC (Lincoln National Corp.)	5.10E+20	43 (6.67E+09)
8	RF (Regions Financial Corp.)	4.10E+20	36 (9.30E+09)
9	STI (SunTrust Banks, Inc.)	4.03E+20	29 (1.44E+10)
10	CBG (CBRE Group, Inc.)	3.73E+20	32 (1.28E+10)

published by Financial Stability Board 2012, six of our top ten systemic risk receivers appear in this report: JPM (J P Morgan), C (Citigroup), WFC (Wells Fargo), BAC (Bank of America), GS (Goldman Sachs), MS (Morgan Stanley), whereas four of our top ten systemic risk emitters appear in this report: C (Citigroup), BAC (Bank of America), WFC (Wells Fargo & Company) and MS (Morgan Stanley). Also we compare our result with the global systemically important insurers (G-SIIs) published by the Financial Stability Board 2013, AIG (American International Group) is present in their list. We compare with the list of all domestic systemically important banks (D-SIBs) in US published by Board of Governors of the Federal Reserve System 2013 as well, USB (US Bancorp), AXP (American Express), COF (Capital One Financial Corp.), RF (Regions Financial Corp.) and STI (SunTrust Banks, Inc.) are on that list. In total all our top 10 Systemic Risk Receivers and 8 of our top 10 Systemic Risk Emitters are identified as systemically important financial institutions. In this step, we could identify "too big as well as too connected" firms which need to be well supervised and regulated.

3.3. Model validation

3.3.1. Comparison with linear models

To evaluate the accuracy of the estimated VaR in the first step, we count the firms' VaRs violations, which is meant to be the situation when the stock losses exceed the estimated VaRs. In Fig. 1 there is no violation in the series of estimated VaR (thinner red line) for J P Morgan. The average violation rate for 100 financial institutions is $\widehat{\tau} = 0.0006$, which is much smaller than the nominal

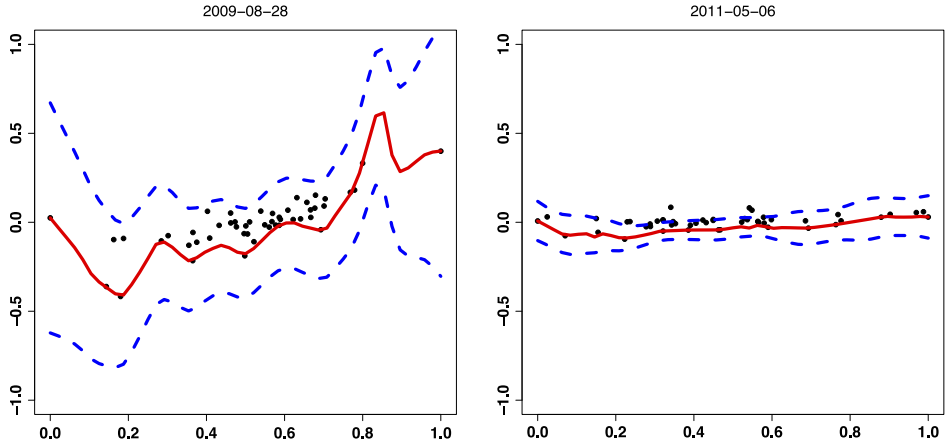


Fig. 9. Left: the estimated link function ($\widehat{\text{CoVaR}}^{TENET}$ of J P Morgan) (solid line) with $h = 0.05$, and the estimated index (points), time period: 20081003–20090828. Right: the estimated link function ($\widehat{\text{CoVaR}}^{TENET}$ of J P Morgan) (solid line) with $h = 0.03$, and estimated the index (points), time period: 20100604–20110506. $\tau = 0.05$, window size $n = 48$, 95% confidence bands (dashed lines).

Table 9

The average p-values of CaViaR test in overall and crisis periods for $\widehat{\text{CoVaR}}^{TENET}$, and the $\widehat{\text{CoVaR}}^{L}$, the standard deviations are given in the brackets.

Average p-value of CaViaR test	$\widehat{\text{CoVaR}}^{TENET}$	$\widehat{\text{CoVaR}}^{L}$
The overall period	0.63(0.33)	0.37(0.41)
The crisis Period	0.72(0.24)	0.51(0.42)

rate $\tau = 0.05$. Since our observations are $T = 266$, most of estimated VaRs do not have violation, the CaViaR test for VaR cannot be performed.

In step 2 we apply the SIM with variable selection to calculate CoVaR. We also compare our results with linear quantile LASSO models in this step to justify the necessity of having a nonlinear model. The benchmark linear LASSO model is written as follows:

$$X_{j,t} = \alpha_{j|R_j} + \beta_{j|R_j}^{L\top} R_{j,t} + \varepsilon_{j,t}, \tag{14}$$

$$\widehat{\text{CoVaR}}^{L}_{j|\widetilde{R}_j,t,\tau} \overset{\text{def}}{=} \widehat{\alpha}_{j|\widetilde{R}_j} + \widehat{\beta}^{L\top}_{j|\widetilde{R}_j} \widetilde{R}_{j,t}, \tag{15}$$

where $\alpha_{j|R_j}$ is a constant term, $R_{j,t}, X_{-j,t}, B_{j,t-1}$, $\widehat{\text{VaR}}_{-j,t,\tau}$ and $\widetilde{R}_{j,t}$ are defined in Section 2.3. The parameters $\beta^L_{j|R_j} \overset{\text{def}}{=} \{\beta^L_{j|-j}, \beta^L_{j|M}, \beta^L_{j|B_j}\}^\top$, and $\widehat{\beta}^L_{j|R_j} \overset{\text{def}}{=} \{\widehat{\beta}^L_{j|-j}, \widehat{\beta}^L_{j|M}, \widehat{\beta}^L_{j|B_j}\}^\top$ which are estimated by using linear quantile regression with variable selection. Then $\widehat{\text{CoVaR}}^{L}$ can be simply calculated.[3]

Recall that we denote our estimated CoVaR in step 2 as $\widehat{\text{CoVaR}}^{TENET}$. Now we compare the performance of $\widehat{\text{CoVaR}}^{TENET}$ and $\widehat{\text{CoVaR}}^{L}$. In Fig. 1 the thinner green line represents the $\widehat{\text{CoVaR}}^{L}$ of J P Morgan, there is 1 violation during the whole time period of $T = 266$, whereas there are 4 violations in the estimated $\widehat{\text{CoVaR}}^{TENET}$ series in Fig. 1 (thicker blue line). We apply the CaViaR test proposed by Berkowitz et al. (2011). While the p-values of $\widehat{\text{CoVaR}}^{TENET}$ in overall period is 0.63, $\widehat{\text{CoVaR}}^{L}$ is only 0.37. Also in crisis period (from 15 September 2008 to 26 February 2010) $\widehat{\text{CoVaR}}^{TENET}$ performs better than $\widehat{\text{CoVaR}}^{L}$, see Table 9 for more details.

Further, we examine the shape of the link functions in the crisis period as well as in the period of relative financial stability. We find out that for almost all firms in a financial crisis period, the link functions are in most of the windows, non-linear, while in a stable period, the link functions tend to be more linear. Take the $\widehat{\text{CoVaR}}^{TENET}$ for J P Morgan as an example. The left panel of Fig. 9 displays the shape of the estimated link function in one window in crisis time and its 95% confidence bands, see Carroll and Härdle (1989). In a stable period one observes in some windows the shape of the link function as on the right panel of Fig. 9. It confirms Chao et al. (2015)'s results stating that the nonlinear model performs better especially in a financial crisis period. We conclude the outperformance of our method over the linear model conditional on the network effects.

Table 10

Pre-Crisis analysis. Top 10 financial institutions ranked according to Incoming links calculated by the sum of absolute value of the partial derivatives, the values of market capitalization (MC) in 2008 are also shown in this table, $\tau = 0.05$, window size $n = 48$, $T = 41$.

Rank	Ticker of IN	IN-Sum	Value of MC
1	FRE (Freddie Mac)	43.95	2.20E+10
2	OCN (Ocwen Financial Corp.)	40.12	3.87E+08
3	NDAQ (The NASDAQ OMX Group)	39.54	6.69E+09
4	FNM (Fannie Mae)	34.07	3.80E+10
5	CACC (Credit Acceptance Corp.)	32.97	5.39E+08
6	KEY (KeyCorp)	32.49	7.98E+09
7	EV (Eaton Vance Corp.)	30.58	3.73E+09
8	PRAA (Portfolio Recovery Associates)	30.07	5.92E+08
9	HBAN (Huntington Bancshares)	29.92	3.36E+09
10	PJC (Piper Jaffray Companies)	29.66	5.75E+08

3.3.2. Pre-Crisis analysis

In this part we would like to test whether our model can detect in advance financial firms which had knock-on effects for the financial systems. We consider mainly five financial firms: FNM (Fannie Mae), FRE (Freddie Mac), LEH (Lehman Brothers), MER (Merrill Lynch) and WB (Wachovia Corp.). The weekly historical returns of these firms are available on the CRSP database. The above mentioned exercise has been carried out again with these five firms (a total of 105 firms in this case). The time period is from 7 December 2007 to 12 September 2008 includes 41 estimates in moving windows. Firstly we show the results from step 2, which checks the connectedness of these firms. Table 10 shows the ranking of total incoming links, where FRE receives most incoming links from other firms, and FNM is ranked as the 4th. From Table 11 it can be seen that the firm with the strongest outgoing links is FNM. Moreover FRE is ranked as the third one, and LEH is ranked in the 7th place. Table 12 presents the direct incoming links and

[3] For simplicity we omit the subscript $_{j|\widetilde{R}_j,t,\tau}$ in $\widehat{\text{CoVaR}}^{L}_{j|\widetilde{R}_j,t,\tau}$, and write $\widehat{\text{CoVaR}}^{L}$.

Table 11
Pre-Crisis analysis. Top 10 financial institutions ranked according to Outgoing links calculated by the sum of absolute value of the partial derivatives, the values of market capitalization (MC) in 2008 are also shown in this table, $\tau = 0.05$, window size $n = 48$, $T = 41$.

Rank	Ticker of OUT	OUT-Sum	Value of MC
1	FNM (Fannie Mae)	252.43	3.80E+10
2	CBG (CBRE Group, Inc.)	157.93	4.04E+09
3	FRE (Freddie Mac)	144.44	2.20E+10
4	WRLD (World Acceptance Corp.)	89.05	5.11E+08
5	CLMS (Calamos Asset Management)	81.13	3.79E+08
6	NEWS (NewStar Financial)	80.71	2.91E+08
7	LEH (Lehman Brothers)	75.73	3.50E+10
8	NNI (Nelnet, Inc.)	70.79	6.03E+08
9	PRAA (Portfolio Recovery Associates)	69.84	5.93E+08
10	C (Citigroup)	68.79	1.03E+11

Table 12
Pre-Crisis analysis. The five defaulted firms are ranked randomly, the received links from other firms and transmitted links to other firms are shown correspondingly. Note that only the first three most influential firms are listed for each ticker, $\tau = 0.05$, $n = 48$, $T = 41$.

Rank	Ticker	Received link from	Transmitted link to
1	FRE (Freddie Mac)	FNM, NS, OCN	FNM, FNF, SEIC
2	FNM(Fannie Mae)	FRE, CBG, HBAN	FRE, LEH, WB
3	LEH (Lehman Brothers)	FNM, WRLD, PJC	AGM, PJC, KEY
4	MER (Merrill Lynch)	FNM, LEH, NEWS	AVHI, JPM, MS
5	WB (Wachovia Corp.)	FNM, C, CMA	CMA, BEN, C

Table 13
Pre-Crisis analysis. Top 12 financial institutions ranked according to the index of Systemic Risk Receiver, the values of market capitalization (MC) in 2008 are also shown in this table.

Rank	Ticker	Value of *SRR*	Value of MC
1	WFC (Wells Fargo & Company)	8.47E+22	1.24E+11
2	C (Citigroup)	8.01E+22	1.03E+11
3	WB (Wachovia Corp.)	6.61E+22	7.30E+10
4	JPM (J P Morgan Chase & Co)	5.26E+22	1.48E+11
5	BAC (Bank of America)	4.20E+22	1.54E+11
6	FRE (Freddie Mac)	3.67E+22	2.20E+10
7	AIG (American International Group)	3.58E+22	3.71E+09
8	MER (Merrill Lynch)	2.81E+22	6.40E+10
9	FNM (Fannie Mae)	2.74E+22	3.80E+10
10	AXP (American Express Company)	2.60E+22	1.79E+10
11	GS (Goldman Sachs Group)	2.41E+22	6.73E+10
12	LEH (Lehman Brothers)	1.80E+22	3.50E+10

outgoing links in terms of other firms. Besides, FRE and FNM is the most connected pair, they send risk to each other. FNM dominates the incoming link tables, which can also be confirmed in Table 10. According to the selected variables in step 2, we perform the methodology in step 3. Table 13 shows the ranking of the systemic risk receivers according to our *SRR* values, where WB is third largest risk receiver, FRE is ranked as the sixth, AIG, MER, FNM follow subsequently, and LEH is ranked as the 12th. The systemic risk emitters according to our *SRE* values are presented in Table 14. We see that FNM is the biggest risk emitter, WB is the third one, FRE and LEH are 4th and 5th risk transmitters and the ranking of MER is 8th. In summary, all these five firms are identified as systemically important institutions which shows the validation of our methodology. Finally, we compare our ranking of systemic risk emitters in Table 14 with Hautsch et al. (2015) and Brownlees and Engle (2015). In the pre-crisis results of Hautsch et al. (2015), they involve five firms in the case study, i.e. AIG, FNM, FRE, LEH and MER. MER is not in their top ten list, whereas we did not identify AIG in our top ten list. Compared with the pre-crisis results with Brownlees and Engle (2015), where the firm Bear Stearns is also involved in their analysis. Their rankings of the aforementioned five firms between 2007 and 2008 are relative similar with ours.

Table 14
Pre-Crisis analysis. Top 10 financial institutions ranked according to the index of Systemic Risk Emitter, the values of market capitalization (MC) in 2008 are also shown in this table.

Rank	Ticker	Value of *SRE*	Value of MC
1	FNM (Fannie Mae)	2.61E+23	3.80E+10
2	C (Citigroup)	1.29E+23	1.03E+11
3	WB (Wachovia Corp.)	9.68E+22	7.30E+10
4	FRE (Freddie Mac)	8.97E+22	2.20E+10
5	LEH (Lehman Brothers)	5.71E+22	3.50E+10
6	CBG (CBRE Group, Inc.)	3.40E+22	4.04E+09
7	COF (Capital One Financial Corp.)	2.85E+20	1.69E+10
8	MER (Merrill Lynch)	2.32E+22	6.40E+10
9	RF (Regions Financial Corp.)	7.37E+21	1.04E+10
10	CMA (Comerica Inc.)	5.29E+21	4.79E+09

4. Conclusion

In this paper we propose TENET based on a semiparametric quantile regression framework to assess the systemic importance of financial institutions conditional to their market capitalization and interconnectedness in tails. The semiparametric model allows for more flexible modelling of the relationship between the variables. This is especially justified in a (ultra) high-dimensional setting when the assumption of linearity is not likely to hold. In order to face these challenges statistically we estimate a SIM in a generalized quantile regression framework while simultaneously performing variable selection. (Ultra) high dimensional setting allows us to include more variables into the analysis.

Our empirical results show that there is growing interconnectedness during the period of a financial crisis, and a network-based measure reflecting the connectivity. Moreover, by including more variables into the analysis we can investigate the overall performance of different financial sectors, depositories, insurance, broker–dealers, and others. Estimation results show a relatively high connectivity of depository industry in the financial crisis. We also observe strong non-linear relationships between the variables, especially, in the period of relative financial instability. The Systemic Risk Receivers and Systemic Risk Emitters can be simply identified based on their connectedness structure and market capitalization. We conclude that both the largest systemic risk receivers and the largest systemic risk emitters are systemically important.

Appendix A. Statistical methodology

Let us denote $X_t \in \mathsf{R}^p$ as p dimensional variables $R_{j,t}$ in (9), p can be very large, namely of an exponential rate. We also drop the subscripts of the coefficients $\beta_{j|R_j}$, as we focus on one regression. The SIM of (9) is then rewritten as:

$$Y_t = g(X_t^\top \beta^*) + \varepsilon_t, \tag{16}$$

where $\{X_t, \varepsilon_t\}$ are strong mixing processes, $g(\cdot)$: $\mathsf{R}^1 \to \mathsf{R}^1$ is an unknown smooth link function, β^* is the vector of index parameters. Regressors X_t can be the lagged variables of Y_t. For the identification, we assume that $\|\beta^*\|^2 = 1$, and the first component of β^* is positive. We assume that there are q non-zero components in β^*.

Note that (16) can be formulated in a location model and identified in a quasi maximum likelihood framework: the direction β^* (for known $g(\cdot)$) is the solution of

$$\min_{\beta} \mathsf{E}\rho_\tau\{Y_t - g(X_t^\top \beta)\}, \tag{17}$$

with loss function

$$\rho_\tau(u) = \tau u \mathbf{1}(u > 0) + (1-\tau)u\mathbf{1}(u < 0), \tag{18}$$
$$\mathsf{E}[\psi_\tau\{Y_t - g(X_t^\top \beta^*)\}|X_t] = 0 \quad a.s.$$

where $\psi_\tau(\cdot)$ is the derivative (a subgradient) of $\rho_\tau(\cdot)$. It can be reformulated as $F^{-1}_{\varepsilon_t|X_t}(\tau) = 0$.

The model is similar to the location scale model considered in Franke et al. (2014). Note that it may be extended to a quantile AR-ARCH type of single index model,

$$Y_t = g(X_t^\top \beta^*) + \sigma(X_t^\top \gamma^*)\varepsilon_t. \tag{19}$$

To estimate the shape of a link function $g(\cdot)$ and β^*, we adopt minimum average contrast estimation approach (MACE) with penalization outlined in Fan et al. (in press). The estimation of β^* and $g(\cdot)$ is as follows:

$$\begin{aligned}\hat\beta_\tau, \hat g(\cdot) &\overset{\text{def}}{=} \arg\min_{\beta, g(\cdot)} -L_n(\beta, g(\cdot)) \\ &= \arg\min_{\beta, g(\cdot)}\ n^{-1}\sum_{j=1}^n\sum_{t=1}^n \rho_\tau\big\{X_t - g(\beta^\top X_j) \\ &\quad - g'(\beta^\top X_j)X_{tj}^\top\beta\big\}\omega_{tj}(\beta) + \sum_{l=1}^p \gamma_\lambda(|\beta_l|),\end{aligned} \tag{20}$$

where $\omega_{tj}(\beta) \overset{\text{def}}{=} \frac{K_h(X_{tj}^\top\beta)}{\sum_{t=1}^n K_h(X_{tj}^\top\beta)}$, $K_h(\cdot) = h^{-1}K(\cdot/h)$, $K(\cdot)$ is a kernel e.g. Gaussian kernel, h is a bandwidth and $L_n(\beta, g(\cdot))$ is defined as $-n^{-1}\sum_{j=1}^n\sum_{t=1}^n \rho_\tau\big\{X_t - g(\beta^\top X_j) - g'(\beta^\top X_j)X_{tj}^\top\beta\big\}\omega_{tj}(\beta) + \sum_{l=1}^p \gamma_\lambda(|\beta_l|)$. Since the data is not equally spaced we choose a bandwidth h based on k-nearest neighbour procedure (See Härdle et al. (2004) and Carroll and Härdle (1989)). The optimal k, number of neighbours, are selected based on a cross-validation criterion. The implementation involves an iteration between estimating β^* and $g(\cdot)$, with a consistent initial estimate for β^*, (Wu et al., 2010). $X_{tj} = X_t - X_j$, $\theta \geq 0$, and $\gamma_\lambda(t)$ is some non-decreasing function concave for $t \in [0, +\infty)$ with a continuous derivative on $(0, +\infty)$. Please note that this MACE functional (with respect to $g(\cdot)$) (20) is in fact only a finite dimensional optimization problem since the minimum over $g(\cdot)$ is to be determined at $a_j = g(\beta^\top X_j)$, $b_j = g'(\beta^\top X_j)$. There are several approaches for the choice of the penalty function. These approaches can be classified based on the properties desired for an optimal penalty function, namely, unbiasedness, sparsity and continuity. The L_1 penalty approach known as least absolute shrinkage and selection operator (LASSO) is proposed for mean regression by Tibshirani (1996). Numerous studies further adapt LASSO to a quantile regression framework, Yu et al. (2003), Li and Zhu (2008), Belloni and Chernozhukov (2011), among others. While achieving sparsity the L_1-norm penalty tends to over-penalize the large coefficients as the LASSO penalty increases linearly in the magnitude of its argument, and, thus, may introduce bias to estimation. As a remedy to this problem the adaptive LASSO estimation procedure was proposed (Zou (2006); Zheng et al. (2013)). Another approach to alleviate the LASSO bias was proposed by Fan and Li (2001) known as Smoothly Clipped Absolute Deviation (SCAD):

$$\gamma_\lambda(t) = \begin{cases}\lambda|t| & \text{for } |t| \leq \lambda, \\ -(t^2 - 2a\lambda|t| + \lambda^2)/2(a-1) & \text{for } \lambda < |t| \leq a\lambda, \\ (a+1)\lambda^2/2 & \text{for } |t| > a\lambda,\end{cases}$$

where $\lambda > 0$ and $a > 2$. Note that for $\lambda = \infty$, this is exactly LASSO.

As for selecting λ, there are two common ways: data-driven generalized cross-validation criterion (GCV) and likelihood-based Schwartz, or Bayesian information criterion-type criteria (SIC, or BIC), Schwarz (1978), Koenker et al. (1994), and their further modifications. The most commonly used criterion is GCV, however, it has been shown that it leads to an overfitted model. Therefore, we employ a modified BIC-type model selection criteria proposed by Wang et al. (2007) and use GCV criterion only to verify whether GCV and BIC diverge significantly. We need to introduce some more notation to present our theoretical results.

Define $\hat\beta_\tau \overset{\text{def}}{=} (\hat\beta_{\tau(1)}^\top, \hat\beta_{\tau(2)}^\top)^\top$ as the estimator for $\beta^* \overset{\text{def}}{=} (\beta_{(1)}^{*\top}, \beta_{(2)}^{*\top})^\top$ attained by the loss in (20). Here $\hat\beta_{\tau(1)}$ and $\hat\beta_{\tau(2)}$ refer to the first q components and the remaining $p - q$ components of $\hat\beta_\tau$ respectively. The same notional logic applies to β^*. For X_t, $X_{(1)t}$ corresponds to $\beta_{(1)}^{*\top}$ and $X_{(0)t}$ corresponds to $\beta_{(2)}^{*\top}$. If in the iterations, we have the initial estimator $\hat\beta_{(1)}^{(0)}$ as a $\sqrt{n/q}$ consistent one for $\beta_{(1)}^*$, we will obtain, with a very high probability, an oracle estimator of the following type, say $\tilde\beta_\tau = (\tilde\beta_{\tau(1)}^\top, \mathbf{0}^\top)^\top$, since the oracle knows the true model $\mathcal{M}_* \overset{\text{def}}{=} \{l : \beta_l^* \neq 0\}$. The following theorem shows that the penalized estimator enjoys the oracle property. Define $\hat\beta^0 \in \mathsf{R}^p$ as the minimizer with the same loss in (20) but within subspace $\{\beta \in \mathsf{R}^p : \beta_{\mathcal{M}_*^c} = \mathbf{0}\}$.

With all the above definitions and conditions, see Appendix B, we present the following theorems.

Theorem A.1. *Under Conditions 1–7, the estimators $\hat\beta^0$ and $\hat\beta_\tau$ exist and coincide on a set with probability tending to 1. Moreover,*

$$\mathrm{P}(\hat\beta^0 = \hat\beta_\tau) \geq 1 - (p-q)\exp(-C'n^\alpha) \tag{21}$$

for a positive constant C', where $\hat\beta^0$ is the "ideal" estimator with non-zero elements correctly specified.

This theorem implies the sign consistency.

Theorem A.2. *Under Conditions 1–7, we have*

$$\|\hat\beta_{\tau(1)} - \beta_{(1)}^*\| = \mathcal{O}_p\{(D_n + n^{-1/2})\sqrt{q}\}. \tag{22}$$

For any unit vector $\mathbf{b}$ in R^q, we have

$$\mathbf{b}^\top C_{0(1)}^{1/2}C_{1(1)}^{-1/2}C_{0(1)}^{1/2}\sqrt{n}(\hat\beta_{\tau(1)} - \beta_{(1)}^*) \xrightarrow{\mathcal{L}} N(0, 1) \tag{23}$$

where $C_{1(1)} \overset{\text{def}}{=} \mathsf{E}\{\mathsf{E}\{\psi_\tau^2(\varepsilon_t)|Z_t\}[g'(Z_t)]^2[\mathsf{E}(X_{(1)t}|Z_t) - X_{(1)t}][\mathsf{E}(X_{(1)t}|Z_t) - X_{(1)t}]^\top\}$, and $C_{0(1)} \overset{\text{def}}{=} \mathsf{E}\{\partial\mathsf{E}\psi_\tau(\varepsilon_t)|Z_t\}\{[g'(Z_t)]^2(\mathsf{E}(X_{(1)t}|Z_t) - X_{(1)t})(\mathsf{E}(X_{(1)t}|Z_t) - X_{(1)t})\}^\top$. Note that $\mathsf{E}(X_{(1)t}|Z_t)$ denotes a $p \times 1$ vector, and $Z_t \overset{\text{def}}{=} X_t^\top\beta^*$, $\psi_\tau(\varepsilon_t)$ is a choice of the subgradient of $\rho_\tau(\varepsilon_t)$ and

$\sigma_\tau^2 \overset{\text{def}}{=} \mathsf{E}[\psi_\tau(\varepsilon_t)]^2/[\partial\mathsf{E}\psi_\tau(\varepsilon_t)]^2$, where

$$\partial\mathsf{E}\psi_\tau(\cdot)|Z_t = \frac{\partial\mathsf{E}\psi_\tau(\varepsilon_t - v)^2|Z_t}{\partial v^2}\Big|_{v=0}. \tag{24}$$

Let us now look at the distribution of $\hat g(\cdot)$ and $\hat g'(\cdot)$, estimators of $g(\cdot), g'(\cdot)$.

Theorem A.3. *Under Conditions 1–7, for any interior point $z = x^\top\beta^*$, $f_Z(z)$ is the density of Z_t, $t = 1, \ldots, n$, if $nh^3 \to \infty$ and $h \to 0$, we have*

$$\begin{aligned}&\sqrt{nh}\sqrt{f_Z(z)/(\nu_0\sigma_\tau^2)}\Big\{\widehat{g}(x^\top\widehat{\beta}) - g(x^\top\beta^*) \\ &\quad - \frac{1}{2}h^2g''(x^\top\beta^*)\mu_2\partial\mathsf{E}\psi_\tau(\varepsilon_t)\Big\} \xrightarrow{\mathcal{L}} N(0,\ 1).\end{aligned}$$

Also, we have

$$\sqrt{nh^3}\sqrt{\{f_Z(z)\mu_2^2\}/(\nu_2\sigma_\tau^2)}\left\{\widehat{g}'(x^\top\widehat{\beta}) - g'(x^\top\beta^*)\right\} \xrightarrow{\mathcal{L}} N(0,\ 1).$$

The dependence does not have any impact on the rate of the convergence of our nonparametric link function. As the degree of the dependence is measured by the mixing coefficient, it is weak enough such that Condition 7 is satisfied. In fact we assume an exponential decaying rate here, which implies the (A.4) in Kong et al. (2010).

Appendix B. Proof

Condition 1. The kernel $K(\cdot)$ is a continuous symmetric function. The link function $g(\cdot) \in C^2$, let $\mu_j \overset{\text{def}}{=} \int u^j K(u)du$ and $\nu_j \overset{\text{def}}{=} \int u^j K^2(u)du, j = 0, 1, 2$.

Condition 2. The derivative (or a subgradient) of $\rho_\tau(x)$, satisfies $\mathsf{E}\psi_\tau(\varepsilon_t) = 0$ and $\inf_{|v|\le c} \partial \mathsf{E}\psi_\tau(\varepsilon_t - v) = C_1$ where $\partial \mathsf{E}\psi_\tau(\varepsilon_t - v)$ is the partial derivative with respect to v, and C_1 is a constant.

Condition 3. The density $f_Z(z)$ of $Z_t = \beta^{*\top} X_t$ is bounded with bounded absolute continuous first-order derivatives on its support. Assume $\mathsf{E}\{\psi_\tau(\varepsilon_t|X_t)\} = 0$ a.s., which means for a quantile loss we have $F^{-1}_{\varepsilon_t|X_t}(\tau) = 0$. Let $X_{(1)t}$ denote the sub-vector of X_t consisting of its first q elements. Let $Z_t \overset{\text{def}}{=} X_t^\top \beta^*$ and $Z_{tj} \overset{\text{def}}{=} Z_t - Z_j$. Define $C_{1(1)} \overset{\text{def}}{=} \mathsf{E}\{\mathsf{E}\{\psi_\tau^2(\varepsilon_t)|Z_t\}[g'(Z_t)]^2[\mathsf{E}(X_{(1)t}|Z_t) - X_{(1)t}][\mathsf{E}(X_{(1)t}|Z_t) - X_{(1)t}]^\top\}$, and $C_{0(1)} \overset{\text{def}}{=} \mathsf{E}\{\partial \mathsf{E}\psi_\tau(\varepsilon_t)|Z_t\}\{[g'(Z_t)]^2(\mathsf{E}(X_{(1)t}|Z_t) - X_{(1)t})(\mathsf{E}(X_{(1)t}|Z_t) - X_{(1)t})\}^\top$, and the matrix $C_{1(1)}$ satisfies $0 < L_1 \le \lambda_{\min}(C_{0(1)}) \le \lambda_{\max}(C_{0(1)}) \le L_2 < \infty$ for positive constants L_1 and L_2. A constant $c_0 > 0$ exists such that $\sum_{t=1}^n \{\|X_{(1)t}\|/\sqrt{n}\}^{2+c_0} \to 0$, with $0 < c_0 < 1$. $v_{tj} \overset{\text{def}}{=} Y_t - a_j - b_j X_{tj}^\top \beta$. Also, a constant C_3 exists such that for all β close to β^* ($\|\beta - \beta^*\| \le C_3$), let $X_{(1)tj}$ denote the subvector of X_{tj} consisting of its first q components, $X_{(0)tj}$ denote the subvector of X_{tj} consisting of its first $p - q$ components:

$$\left\| \sum_t \sum_j X_{(0)tj} \omega_{tj} X_{(1)tj}^\top \partial \mathsf{E}\psi_\tau(v_{tj}) \right\|_{2,\infty} = \mathcal{O}_p(n^{1-\alpha_1}).$$

Condition 4. The penalty parameter λ is chosen such that $\lambda = \mathcal{O}(n^{-1/2})$, with $D_n \overset{\text{def}}{=} \max\{d_l : l \in \mathcal{M}_*\} = \mathcal{O}(n^{\alpha_1 - \alpha_2/2}\lambda) = \mathcal{O}(n^{-1/2})$, $d_l \overset{\text{def}}{=} \gamma_\lambda(|\beta_l^*|)$, $\mathcal{M}_* = \{l : \beta_l^* \ne 0\}$ be the true model. Furthermore assume $qh \to 0$ and $h^{-1}\sqrt{q/n} = \mathcal{O}(1)$ as n goes to infinity, $q = \mathcal{O}(n^{\alpha_2})$, $p = \mathcal{O}\{\exp(n^\delta)\}$, $nh^3 \to \infty$ and $h \to 0$. Also, $0 < \delta < \alpha < \alpha_2/2 < 1/2, \alpha_2/2 < \alpha_1 < 1$.

Condition 5. The error term ε_t satisfies $\mathsf{Var}(\varepsilon_t) < \infty$. Assume that for any integer $m = 1, \ldots, \infty$

$$\sup_t \mathsf{E}\left|\psi_\tau^m(\varepsilon_t)/m!\right| \le s_0 M^m$$
$$\sup_t \mathsf{E}\left|\psi_\tau^m(x_{tj})/m!\right| \le s_0 M^m$$

where s_0 and M are constants, and $\psi_\tau(\cdot)$ is the derivative (a subgradient) of $\rho_\tau(\cdot)$.

Condition 6. The conditional density function $f(\varepsilon_t|Z_t = z)$ is bounded and absolutely continuously differentiable.

Condition 7. $\{X_{tj}, \varepsilon_t\}_{t=-\infty}^{t=\infty}$ is a strong mixing process for any j. Moreover, let $m1$ and $m2$ be constants, positive constants c_{m1} and c_{m2} exists such that the $\alpha-$ mixing coefficient for every $j \in \{1, \ldots, p\}$,

$$\alpha(l) \le \exp(-c_{m1} l^{c_{m2}}), \tag{25}$$

where $c_{m2} > 2\alpha$.

Recall (20) and $\hat\beta^0$ as the minimizer with the loss

$$\tilde L_n(\beta) \overset{\text{def}}{=} \sum_{j=1}^n \sum_{t=1}^n \rho_\tau\big(Y_t - a_j^* - b_j^* X_{tj}^\top \beta\big)\omega_{tj}(\beta^*) + n\sum_{l=1}^p d_l|\beta_l|,$$

but within the subspace $\{\beta \in \mathsf{R}^p : \beta_{\mathcal{M}_*^c} = \mathbf{0}\}$, and $a_j^* = g(\beta^{*\top} X_j)$, $b_j^* = g'(\beta^{*\top} X_j)$. The following lemma assures the consistency of $\hat\beta^0$.

Lemma B.1. *Under Conditions 1–7, recall $d_j = \gamma_\lambda\left(|\beta_j^*|\right)$, we have that*

$$\|\hat\beta^0 - \beta^*\| = \mathcal{O}_p\big(\sqrt{q/n} + \|\mathbf{d}_{(1)}\|\big) \tag{26}$$

where $\mathbf{d}_{(1)}$ is the subvector of $\mathbf{d} = (d_1, \ldots, d_p)^\top$ which contains q elements corresponding to the non-zero $\beta^*_{(1)}$.

Proof. Note that the last $p-q$ elements of both $\hat\beta^0$ and β^* are zero, so it is sufficient to prove $\|\hat\beta^0_{(1)} - \beta^*_{(1)}\| = \mathcal{O}_p\big(\sqrt{q/n} + \|\mathbf{d}_{(1)}\|\big)$.

Following Fan et al. (in press), it is not hard to prove that for $\gamma_n = \mathcal{O}(1)$:

$$\mathrm{P}\left[\inf_{\|\mathbf{u}\|=1}\left\{\tilde L_n(\beta^*_{(1)} + \gamma_n \mathbf{u},\ \mathbf{0}) > \tilde L_n(\beta^*)\right\}\right] \to 1.$$

Then a minimizer inside the ball exists $\{\beta_{(1)} : \|\beta_{(1)} - \beta^*_{(1)}\| \le \gamma_n\}$. Construct $\gamma_n \to 0$ so that for a sufficiently large constant B_0: $\gamma_n > B_0 \cdot \big(\sqrt{q/n} + \|\mathbf{d}_{(1)}\|\big)$. Then by the local convexity of $\tilde L_n(\beta_{(1)}, \mathbf{0})$ near $\beta^*_{(1)}$, a unique minimizer exists inside the ball $\{\beta_{(1)} : \|\beta_{(1)} - \beta^*_{(1)}\| \le \gamma_n\}$ with probability tending to 1. □

Recall that $X_t = (X_{(1)t}, X_{(0)t})$ and $\mathcal{M}_* = \{1, \ldots, q\}$ is the set of indices at which β are non-zero.

Lemma B.1 shows the consistency of $\hat\beta^0$, and we need to show further that $\hat\beta^0$ is the unique minimizer in R^p on a set with probability tending to 1.

Lemma B.2. *Under Conditions 1–7, minimizing the loss function $\tilde L_n(\beta)$ has a unique global minimizer $\hat\beta_\tau = (\hat\beta_{\tau(1)}^\top, \hat\beta_{\tau(2)}^\top)^\top = (\hat\beta_{\tau(1)}^\top, \mathbf{0}^\top)^\top$, if and only if on a set with probability tending to 1,*

$$\sum_{j=1}^n \sum_{t=1}^n \psi_\tau\big(Y_t - \hat a_j - \hat b_j X_{tj}^\top \hat\beta_\tau\big)\hat b_j X_{(1)tj}\omega_{tj}(\beta^*) + n\mathbf{d}_{(1)} \circ \mathrm{sign}(\hat\beta_\tau) = 0 \tag{27}$$

$$\|z(\hat\beta_\tau)\|_\infty \le n, \tag{28}$$

where

$$z(\hat\beta_\tau) \overset{\text{def}}{=} \mathbf{d}_{(0)}^{-1} \circ \left\{\sum_{j=1}^n \sum_{t=1}^n b_j^* \psi_\tau\big(Y_t - a_j^* - b_j^* X_{tj}^\top \hat\beta_\tau\big) X_{(0)tj}\omega_{tj}(\hat\beta_\tau)\right\} \tag{29}$$

where $\circ$ stands for multiplication element-wise.

Proof. According to the definition of $\hat\beta_\tau$, it is clear that $\hat\beta_{(1)}$ already satisfies condition (27). Therefore we only need to verify condition (28). To prove (28), a bound for

$$\sum_{i=1}^n \sum_{i=1}^n b_j^* \psi_\tau\big(Y_i - a_j^* - b_j^* X_{ij}^\top \beta^*\big)\omega_{ij} X_{(0)ij} \tag{30}$$

is needed, note that to be consistent with notations for $U-$ statistics we use j instead of t within this proof. Define the following kernel function

$$h_d(X_i, a_j^*, b_j^*, Y_i, X_j, a_i^*, b_i^*, Y_j) = \frac{n}{2}\left\{b_j^* \psi_\tau\big(Y_i - a_j^* - b_j^* X_{ij}^\top \beta^*\big)\omega_{ij} X_{(0)ij} + b_i^* \psi_\tau\big(Y_j - a_i^* - b_i^* X_{ij}^\top \beta^*\big)\omega_{ji} X_{(0)ji}\right\}_d,$$

where $\{.\}_d$ denotes the dth element of a vector, $d = 1, \ldots, p - q$.

According to Borisov and Volodko (2009), based on Condition 5:

Define $U_{n,d} \overset{\text{def}}{=} \frac{1}{n(n-1)} \sum_{1\le i<j\le n} h_d(X_i, a_j^*, b_j^*, Y_i, X_j, a_i^*, b_i^*, Y_j)$ as the $U-$ statistics for (30). We have, with sufficient large c_{m2} in Condition 7.

$$\mathrm{P}\{|U_{n,d} - \mathrm{E}U_{n,d}| > \varepsilon\} \le c_{m3} \exp(c_{m5}\varepsilon/(c_{m3} + c_{m4}\varepsilon^{1/2} n^{-1/2}))$$

where c_{m3}, c_{m4}, c_{m5} are constants. Moreover, let $\varepsilon = \mathcal{O}(n^{1/2+\alpha})$ and m_6 be a constant, as $\alpha < 1/2$, we can further have,

$$\mathrm{P}(\{|U_{n,d} - \mathrm{E}U_{n,d}| > \varepsilon\}) \le c_{m3} \exp(-c_{m6}\varepsilon/2),$$

Define

$$F_{n,d} \overset{\text{def}}{=} (n)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} b_j \psi_\tau\big(Y_i - a_j^* - b_j^* X_{ij}^\top \beta^*\big) \omega_{ij} X_{(0)ij},$$

also it is not hard to derive that $U_{n,d} = F_{n,d} n/(n-1)$.

It then follows that

$$\begin{aligned}\mathrm{P}(|F_{n,d} - \mathrm{E}F_{n,d}| > \varepsilon) &= \mathrm{P}(|U_{n,d} - \mathrm{E}U_{n,d}|(n-1)/n > \varepsilon) \\ &\le 2\exp\big(-Cn^{\alpha+1/2}\big).\end{aligned}$$

Define $\mathcal{A}_n = \{\|F_n - \mathrm{E}F_n\|_\infty \le \varepsilon\}$, thus

$$\begin{aligned}\mathrm{P}(\mathcal{A}_n) &\ge 1 - \sum_{d=1}^{p-q} \mathrm{P}(|F_{n,d} - \mathrm{E}F_{n,d}| > \varepsilon) \\ &\ge 1 - 2(p-q)\exp\big(-Cn^{\alpha+1/2}\big).\end{aligned}$$

Finally we get that on the set $\mathcal{A}_n$,

$$\begin{aligned}\|z(\hat\beta^0)\|_\infty &\le \|\mathbf{d}_{\mathcal{M}_*^c}^{-1} \circ F_n\|_\infty \\ &\quad + \|\mathbf{d}_{\mathcal{M}_*^c}^{-1} \circ \sum_{i=1}^{n}\sum_{j=1}^{n} b_j \big[\psi_\tau\big(Y_t - a_j^* - b_j^* X_{ij}^\top \hat\beta^0\big) \\ &\quad - \psi_\tau\big(Y_t - a_j^* - b_j^* X_{ij}^\top \beta^*\big)\big] \omega_{ij} X_{(0)ij}\|_\infty \\ &\le \mathcal{O}(n^{1/2+\alpha}/\lambda \\ &\quad + \|\mathbf{d}_{\mathcal{M}_*^c}^{-1} \circ \sum_{i=1}^{n}\sum_{j=1}^{n} \partial \mathrm{E}\psi_\tau(v_{ij}) b_j X_{(1)ij}^\top (\hat\beta_{(1)} - \beta_{(1)}^*) \omega_{tj} X_{(0)ij}\|_\infty),\end{aligned}$$

where v_{ij} is between $Y_i - a_j^* - b_j^* X_{ij}^\top \beta^*$ and $Y_i - a_j^* - b_j^* X_{ij}^\top \hat\beta^0$. From Lemma B.1,

$$\|\hat\beta^0 - \beta_{(1)}^*\|_2 = \mathcal{O}_p\Big(\|\mathbf{d}_{(1)}\| + \sqrt{q}/\sqrt{n}\Big).$$

Choosing $\|\sum_i \sum_j X_{(0)ij} \omega_{ij} X_{(1)ij}^\top \partial \mathrm{E}\psi_\tau(v_{ij})\|_{2,\infty} = \mathcal{O}_p(n^{1-\alpha_1})$, $q = \mathcal{O}(n^{\alpha_2})$, $\lambda = \mathcal{O}(\sqrt{q/n}) = n^{-1/2+\alpha_2/2}$, $0 < \alpha_2 < 1$, $\|\mathbf{d}_{(1)}\| = \mathcal{O}(\sqrt{q} D_n) = \mathcal{O}(n^{\alpha_2/2} D_n)$

$$\begin{aligned}n^{-1}\|z(\hat\beta^0)\|_\infty &= \mathcal{O}\{n^{-1}\lambda^{-1}(n^{1/2+\alpha} \\ &\quad + n^{1-\alpha_1}\sqrt{q}/\sqrt{n} + \|\mathbf{d}_{(1)}\| n^{1-\alpha_1})\} \\ &= \mathcal{O}(n^{-\alpha_2/2+\alpha} + n^{-\alpha_1} + n^{-\alpha_1+\alpha_2/2} D_n/\lambda),\end{aligned}$$

Condition 4 ensures $D_n = \mathcal{O}(n^{\alpha_1-\alpha_2/2}\lambda)$, and let $0 < \delta < \alpha < \alpha_2/2 < 1/2$, $\alpha_2/2 < \alpha_1 < 1$, with rate $p = \mathcal{O}\{\exp(n^\delta)\}$, then $(n)^{-1}\|z(\hat\beta^0)\|_\infty = \mathcal{O}_p(1)$. □

Proof of Theorem A.1. The results follows from Lemmas B.1 and B.2. □

Proof of Theorem A.2. By Theorem A.1, $\hat\beta_{\tau(1)} = \beta_{(1)}$ almost surely. It then follows from Lemma B.2 that

$$\|\hat\beta_{\tau(1)} - \beta_{(1)}^*\| = \mathcal{O}_p\{(D_n + n^{-1/2})\sqrt{q}\}.$$

This completes the first part of the theorem. The other part of proof follows largely from Fan et al. (in press). □

References

Acharya, V., Engle, R., Richardson, M., 2012. Capital shortfall: A new approach to ranking and regulating systemic risks. Amer. Econ. Rev. 102 (3), 59–64.

Adrian, T., Brunnermeier, M.K., 2011. CoVaR. Staff reports 348. Federal Reserve Bank of New York.

Beale, N., Rand, D.G., Battey, H., Croxson, K., May, R.M., Nowak, M.A., 2011. Individual versus systemic risk and the regulator's dilemma. Proc. Natl. Acad. Sci. 108 (31), 12647–12652.

Belloni, A., Chernozhukov, V., 2011. L1-penalized quantile regression in high-dimensional sparse models. Ann. Statist. 39 (1), 82–130.

Berkowitz, J., Christoffersen, P., Pelletier, D., 2011. Evaluating value-at-risk models with desk-level data. Manage. Sci. 57 (12), 2213–2227.

Betz, F., Hautsch, N., Peltonen, T., Schienle, M., 2016. Systemic risk spillovers in the European banking and sovereign network. J. Financ. Stabil. http://dx.doi.org/10.1016/j.jfs.2015.10.006.

Billio, M., Getmansky, M., Lo, A.W., Pelizzon, L., 2012. Econometric measures of connectedness and systemic risk in the finance and insurance sectors. J. Financ. Econ. 104 (3), 535–559.

Bisias, D., Flood, M., Lo, A.W., Valavanis, S., 2012. A survey of systemic risk analytics. Annu. Rev. Financ. Econ. 4 (1), 255–296.

Borisov, I., Volodko, N., 2009. Exponential inequalities for the distributions of canonical u-and v-statistics of dependent observations. Siberian Adv. Math. 19 (1), 1–12.

Borke, L., Yu, L., Benschop, T., 2016. FRM: Financial Risk Meter. SFB Discussion Paper 2016 (forthcoming).

Boss, M., Krenn, G., Puhr, C., Summer, M., 2006. Systemic risk monitor: A model for systemic risk analysis and stress testing of banking systems. Financ. Stab. Rep. 11, 83–95.

Brownlees, C.T., Engle, R.F., 2015. Srisk: a conditional capital shortfall index for systemic risk measurement. http://dx.doi.org/10.2139/ssrn.1611229.

Carroll, R.J., Härdle, W., 1989. Symmetrized nearest neighbor regression estimates. Statist. Probab. Lett. 7 (4), 315–318.

Chan-Lau, J., Espinosa, M., Giesecke, K., Solé, J., 2009. Assessing the systemic implications of financial linkages. IMF Global Financial Stability Report. 2.

Chao, S.-K., Härdle, W.K., Wang, W., 2015. Quantile regression in risk calibration. Handb. Financ. Econom. Stat. 1467–1489.

Diebold, F.X., Yilmaz, K., 2014. On the network topology of variance decompositions: Measuring the connectedness of financial firms. J. Econometrics 182, 119–134.

Eisenberg, L., Noe, T.H., 2001. Systemic risk in financial systems. Manage. Sci. 47 (2), 236–249.

Fan, Y., Härdle, W.K., Wang, W., Zhu, L., 2016. Single-index based CoVaR with very high dimensional covariates. SFB 649 Discussion Paper 2013-010, Humbold-Universität zu Berlin. J. Bus. Econom. Statist. (in press).

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96 (456), 1348–1360.

Franke, J., Mwita, P., Wang, W., 2014. Nonparametric estimates for conditional quantiles of time series. AStA Adv. Stat. Anal. 1–24.

Gertler, M., Kiyotaki, N., 2010. Financial intermediation and credit policy in business cycle analysis. Handb. Monet. Econ. 3 (11), 547–599.

Giglio, S., Kelly, B., Pruitt, S., Qiao, X., 2012. Systemic risk and the macroeconomy: An empirical evaluation. Fama-Miller Working Paper.

Härdle, W.K., Müller, M., Sperlich, S., Werwatz, A., 2004. Nonparametric and Semiparametric Models. Springer.

Hautsch, N., Schaumburg, J., Schienle, M., 2015. Financial network systemic risk contributions. Rev. Finance 19 (2), 685–738.

Huang, X., Zhou, H., Zhu, H., 2009. A framework for assessing the systemic risk of major financial institutions. J. Bank. Finance 33 (11), 2036–2049.

Koenker, R., Ng, P., Portnoy, S., 1994. Quantile smoothing splines. Biometrika 81 (4), 673–680.

Kong, E., Linton, O., Xia, Y., 2010. Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model. Econometric Theory 26 (05), 1529–1564.

Lehar, A., 2005. Measuring systemic risk: A risk management approach. J. Bank. Finance 29 (10), 2577–2603.

Li, Y., Zhu, J., 2008. L1-norm quantile regression. J. Comput. Graph. Statist. 17 (1).

Minsky, H.P., 1977. A theory of systemic fragility. Financial crises: Institutions and Markets in a Fragile Environment 138–152.

Rodriguez-Moreno, M., Peña, J.I., 2013. Systemic risk measures: The simpler the better? J. Bank. Finance 37 (6), 1817–1831.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6 (2), 461–464.

Shi, J., Malik, J., 2000. Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 22 (8), 888–905.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 267–288.

Wang, H., Li, R., Tsai, C.-L., 2007. Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika 94 (3), 553–568.

Wu, T.Z., Yu, K., Yu, Y., 2010. Single-index quantile regression. J. Multivariate Anal. 101 (7), 1607–1621.

Yu, K., Lu, Z., Stander, J., 2003. Quantile regression: applications and current research areas. J. R. Stat. Soc. Ser. D 52 (3), 331–350.

Zheng, Q., Gallagher, C., Kulasekera, K., 2013. Adaptive penalized quantile regression for high dimensional data. J. Statist. Plann. Inference 143 (6), 1029–1038.

Zou, H., 2006. The adaptive LASSSO and its oracle properties. J. Amer. Statist. Assoc. 101 (476), 1418–1429.

Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: https://www.tandfonline.com/loi/cjas20

Do maternal health problems influence child's worrying status? Evidence from the British Cohort Study

Xianhua Dai, Wolfgang Karl Härdle & Keming Yu

To cite this article: Xianhua Dai, Wolfgang Karl Härdle & Keming Yu (2016) Do maternal health problems influence child's worrying status? Evidence from the British Cohort Study, Journal of Applied Statistics, 43:16, 2941-2955, DOI: 10.1080/02664763.2016.1155203

To link to this article: https://doi.org/10.1080/02664763.2016.1155203

Published online: 22 Mar 2016.

JOURNAL OF APPLIED STATISTICS, 2016
VOL. 43, NO. 16, 2941–2955
http://dx.doi.org/10.1080/02664763.2016.1155203

Taylor & Francis
Taylor & Francis Group

Do maternal health problems influence child's worrying status? Evidence from the British Cohort Study

Xianhua Dai[a], Wolfgang Karl Härdle[b,c] and Keming Yu[d]

[a]Wuhan Institute of Technology, Wuhan, People's Republic of China; [b]C.A.S.E., Humboldt-Universität zu Berlin, Berlin, Germany; [c]School of Business, Singapore Management University, Singapore; [d]Department of Mathematical Sciences, Brunel University, Uxbridge, UK

ABSTRACT

Conventional methods apply symmetric prior distributions such as a normal distribution or a Laplace distribution for regression coefficients, which may be suitable for median regression and exhibit no robustness to outliers. This work develops a quantile regression on linear panel data model without heterogeneity from a Bayesian point of view, i.e. upon a location-scale mixture representation of the asymmetric Laplace error distribution, and provides how the posterior distribution is summarized using Markov chain Monte Carlo methods. Applying this approach to the 1970 British Cohort Study (BCS) data, it finds that a different maternal health problem has different influence on child's worrying status at different quantiles. In addition, applying stochastic search variable selection for maternal health problems to the 1970 BCS data, it finds that maternal nervous breakdown, among the 25 maternal health problems, contributes most to influence the child's worrying status.

ARTICLE HISTORY
Received 31 January 2015
Accepted 14 February 2016

KEYWORDS
British Cohort Study data; Bayesian inference; quantile regression; asymmetric Laplace error distribution; Markov chain Monte Carlo; stochastic search variableselection

JEL CLASSIFICATIONS
C11; C38; C63

1. Introduction

In many applications, conventional regression analysis focuses on the mean effect or optimal forecasting in a mean squared error sense. Since a set of quantiles often provides more complete description of the response distribution than the mean, or classical mean regression, quantile regression not only quantifies the relationship between quantiles of the response distribution and covariates, but also exhibits robustness to outliers and has a wide application [4,28,51], for example, to calculate value at risk and expected shortfall for financial risk management [45], to study the relationship between GDP and population [40,41], to study the correlation of the wage and the level of education [23], and to estimate the volatility of temperatures [20].

For classical quantile regression, the error distribution is often assumed to have the qth quantile equal to zero, see, for example, Yu and Stander (2007) [53], and classical quantile regression parameters depend on asymptotic normality which is assumed unbiased and normal. In addition, confidence intervals depend on the density function of model error

CONTACT Xianhua Dai xhrdai@gmail.com

which is difficult to estimate reliability. On the contrary, credible intervals from Bayesian inference can avoid these problems, whichever sample sizes. Aside from these, Bayesian inference can take historical information or expert opinion easily via prior information. Therefore, Bayesian quantile regression is naturally motivated.

Quantile regression is attempted in Bayesian framework in both theoretical and applied econometric analyses, for example, Walker and Mallick [47], Kottas and Gelfand [29], and Hanson and Johnson [22] on median regression (one special quantile regression), and Yu and Moyeed [52], Tsionas [46], and Kozumi and Kobayashi [31] on general quantile regression with the asymmetric Laplace density for the errors. In addition, on infinite mixture model, Kottas and Krnjajic [30] on Bayesian semi-parametric approach, Yu [49], Taddy and Kottas [44], and Yue and Rue [56] on Bayesian nonparametric approach. However, few studies have been on Bayesian quantile regression for panel data [37,55].

This paper explores a Bayesian quantile regression for linear panel data without heterogeneity. For posterior inference, upon a location-scale mixture representation of the asymmetric Laplace error distribution, we propose a Gibbs sampling algorithm and develop Markov chain Monte Carlo (MCMC) methods (see, e.g. [8,18,34]). All posterior densities are fully tractable and easy to sample, making the Gibbs sampler appealing when several quantile regressions are required at one time. In addition, the proposed Gibbs sampler can be applied for the calculation of the marginal likelihood and the variable selection.

For variable selection, several criteria have been proposed (see, e.g. [58]), though no agreement has emerged in the literature on optimal criterion. Aside from the classical literature, Bayesian approach focuses on an unknown number of variables [9,17]. Variable selection in modeling with Bayesian quantile regression is difficult due to the computational efficiency. This work applies stochastic search variable selection based on MCMC method.

We apply Bayesian approach to the 1970 British Cohort Study (BCS) to analyze the influence of maternal health problems on child's worrying status. This is the first instance, as we know, in which the influences of the maternal health problems are estimated to account for child's worrying status. We find that the different maternal health problems have different influence on child's worrying status at different quantiles, moreover, maternal nervous breakdown, among the 25 maternal health problems, contributes most to influence the child's worrying status. Indeed Bayesian approach may be applied to empirical study of optimal taxation under prospect theory, or predictive asset return, see, for example, Kanbur *et al.* [27] and Dai [13] for optimal taxation under prospect theory, and Campbell and Yogo [5] and Dai *et al.* [15] for predictive asset return.

This paper joins the literature in health economics and personality psychology. While it is established in psychology on their importance (see, e.g. [21,38,39]), and in economics for the influence of personality traits on health [9,26] and health-related behaviors [10,11,24], it is less recognized in economics on the influence of maternal health problems on child's worrying status.

Using principal component analysis, a few economic result from the BCS data, for example, psychological and behavioral development influences education and labor market outcomes [16], intergenerational income persistence rises across the 1958 and the 1970 cohorts [3], and the standardized raw scores from the locus of control and self-esteem scales significantly predict self-reported poor health at age 30 [35]. Other data may be explored, see, for example [14]. This work goes beyond those studies, since Bayesian

inference is explored to examine the influence of maternal health problems on child's worrying status.

The remainder of the paper is structured as follows. In the next section, we describe the BCS data. Section 3 outlines the basic model, while Section 4 develops MCMC method for quantile regression model and explains how the MCMC output may be used to compute the marginal likelihoods and for variable selection. Empirical implementation and results for our Bayesian approach are shown in Section 5. Section 6 concludes.

2. Data: the BCS

The data, we use in this work, are from the BCS, a survey of all babies born (alive or dead) after the 24th week of gestation from 0.01 hours on Sunday, 5 April 1970 to 24.00 hours on Saturday, 11 April 1970 in places including England, Scotland, Wales, and Northern Ireland. Seven surveys, in detail, respectively in 1975, 1980, 1986, 1996, 2000, 2004, and 2008, are followed up so far to trace all members of the birth cohort. In this work, information on background characteristics is drawn from the survey in 1975 and 1980 on maternal health problems, and on child's worrying status from the survey in 1980 and 1986. Samples from the family of multiple children are excluded, and samples for the respondents with any missing information on those background characteristics are also excluded. A sample of size 3426 is left for our analysis in this paper.

2.1. Rutter score-derived variable for child

Applying the Rutter Behavior Scale question 'Often worried?' for child, the Rutter score-derived variable, Y , was derived, where the question was completed by the cohort member's parent (usually the mother) in the BCS 1980 and 1986 follow-up data sets. In the BCS, the Rutter score derived, and thus the response variable, is discrete choice. For our case, the response results are 1 (does not worry), 2 (somewhat worried), and 3 (certainly worried).

2.2. Mother Malaise score-derived variables

Applying the Malaise Inventory ('How you feel') completed by the cohort member's parent (usually the mother), the Mother Malaise score-derived variables were derived on behalf of the cohort member and included in the BCS 1975 and 1980 follow-up data sets. These 25 variables were named in the Mother Malaise data sets as follows:

(1) Do you often have backache? (X_1)
(2) Do you feel tired most of the time? (X_2)
(3) Do you often feel depressed? (X_3)
(4) Do you often have bad headaches? (X_4)
(5) Do you often get worried about things? (X_5)
(6) Do you usually have great difficulty in falling or staying asleep? (X_6)
(7) Do you usually wake unnecessarily early in the morning? (X_7)
(8) Do you wear yourself out worrying about your health? (X_8)
(9) Do you often get into a violent rage? (X_9)
(10) Do people annoy and irritate you? (X_{10})

(11) Have you at times had a twitching of the face, head or shoulders? (X_{11})
(12) Do you suddenly become scared for no good reason? (X_{12})
(13) Are you scared to be alone when there are not friends near you? (X_{13})
(14) Are you easily upset or irritated? (X_{14})
(15) Are you frightened of going out alone or of meeting people? (X_{15})
(16) Are you constantly keyed up and jittery? (X_{16})
(17) Do you suffer from indigestion? (X_{17})
(18) Do you suffer from an upset stomach? (X_{18})
(19) Is your appetite poor? (X_{19})
(20) Does every little thing get on your nerves and wear you out? (X_{20})
(21) Does your heart often race like mad? (X_{21})
(22) Do you often have bad pain in eyes? (X_{22})
(23) Are you troubled with rheumatism or fibrosis? (X_{23})
(24) Have you ever had a nervous breakdown? (X_{24})
(25) Do you have other health problems? (X_{25})

3. Potential outcome model

Let Y_{it+1} be the Rutter score-derived variable for the ith cohort member surveyed at the $(t+1)$th sweep, and $X_{1,it}, X_{2,it}, \ldots, X_{25,it}$ the Mother Malaise score-derived variables for the ith cohort member's parent (usually the mother) surveyed at the tth sweep. We introduce the linear panel data model without heterogeneity as follows:

$$Y_{it+1} = \beta_0 + \sum_{j=1}^{25} \beta_j X_{j,it} + \varepsilon_{it}, \tag{1}$$

for $i = 1, 2, \ldots, 3426$, and $t = 1,2$, where $\beta_.$ is unknown parameter, and ε_{it} is an idiosyncratic error term assumed to be independent of the Rutter score-derived variable and Mother Malaise score-derived variables.

4. Bayesian inference and variable selection

In this study, we consider quantile regression to estimate β from

$$\min \sum_{i=1}^{3426} \sum_{t=1}^{2} \rho_q \left(Y_{it} - \sum_{j=1}^{25} \beta_j X_{j,it} - \beta_0 \right), \tag{2}$$

where $\rho_q(.)$ in Equation (2) is the check function defined by

$$\rho_q(u) \equiv \{q - I(u < 0)\} \cdot u, \tag{3}$$

for $0 < q < 1$, where $I(.)$ is the indicator function. Instead of classical approach, a Bayesian approach and MCMC algorithm will be developed for posterior inference.

4.1. Asymmetric Laplace distribution

For Bayesian inference of Equation (2), an assumption on the data distribution is required to construct a likelihood function. The error term ε_{it} is assumed, following Yu and Moyeed

[52], to follow the asymmetric Laplace distribution (ALD) with density

$$f_{\mathrm{AL}}(\varepsilon_{it}) = \frac{q(1-q)}{\sigma} \exp\left\{-\rho_q\left(\frac{\varepsilon_{it}}{\sigma}\right)\right\}, \tag{4}$$

where σ is the scale parameter. For the properties of this distribution, see, for example [52,54]. Note that the qth quantile of ε_{it} is zero, $\mathrm{E}(\varepsilon_{it}) = (1-2q)/q(1-q)$, and $\mathrm{Var}(\varepsilon_{it}) = (1-2q+2q^2)/q^2(1-q)^2$.

To develop MCMC algorithm for the quantile regression, a location-scale mixture representation is applied, i.e.

$$\varepsilon_{it} = \theta v_{it} + \tau\sqrt{\sigma v_{it}}u_{it}, \tag{5}$$

where $\theta = (1-2q)/q(1-q)$, $\tau^2 = 2/q(1-q)$, $v_{it} \sim \varepsilon(\sigma)$ and $u_{it} \sim \mathrm{N}(0,1)$ are mutually independent random variables, and $\varepsilon(\sigma)$ is the exponential distribution with mean σ [31]. Thus the panel data model without heterogeneity can be represented as follows:

$$Y_{it} = \beta_0 + \sum_{j=1}^{25} \beta_j X_{j,it} + \theta v_{it} + \tau\sqrt{\sigma v_{it}}u_{it}, \tag{6}$$

where $v_{it} \sim \varepsilon(\sigma)$ and $u_{it} \sim \mathrm{N}(0,1)$ are mutually independent random variables.

To begin posterior inference, some prior distributions are supposed as follows: (1) $\beta \sim \mathrm{N}(\beta_0, B_0)$, where $\beta \equiv (\beta_0, \beta_1, \ldots, \beta_{25})$, and β_0 and B_0 are specified parameters; (2) $\sigma \sim \mathrm{IG}(n_0/2, s_0/2)$, where $\mathrm{IG}(a,b)$ is the inverse Gamma distribution with the parameters a and b, and n_0 and s_0 are specified parameters. These priors are chosen for computational reasons, but are flexible enough when analyzing BCS to represent various prior beliefs about the parameters. Next to construct a MCMC algorithm with those prior distributions.

4.2. MCMC algorithm

A MCMC algorithm (see, e.g. [8,18,34]) for the quantile regression is constructed by sampling $\{v_{it}\}$, β, and σ from their full conditional distributions applying the data augmentation techniques as [6]. A tractable and efficient Gibbs sampler is proposed for general $i = 1, 2, \ldots, N$ and $t = 1, T$ as follows. In the empirical part, $N = 3426$ and $T = 2$.

(1) Sample v_{it} $(i = 1, 2, \ldots, N; t = 1, T)$ from $\mathrm{GIG}(\frac{1}{2}, \hat{c}_{it}^2, \hat{d}_{it}^2)$, where

$$\hat{c}_{it}^2 = \frac{(Y_{it+1} - \beta^\top X_{it})^2}{\tau^2\sigma}, \tag{7}$$

$$\hat{d}_{it}^2 = \frac{\theta^2}{\tau^2\sigma} + \frac{2}{\sigma}, \tag{8}$$

and $\mathrm{GIG}(\nu, c, d)$ is the generalized inverse Gaussian distribution with the probability density function

$$f_{\mathrm{GIG}}(x \mid \nu, c, d) = \frac{\left(\frac{d}{c}\right)^\nu}{2K_\nu(cd)} x^{\nu-1} \exp\left\{-\frac{1}{2}(c^2x^{-1} + d^2x)\right\}, \tag{9}$$

for $x > 0$, $-\infty < \nu < \infty$, and c,$d > 0$, where $K_\nu(.)$ is a modified Bessel function of the third kind [2].

(2) Sample β from $\mathrm{N}(\hat{\beta}, \hat{B})$, where

$$\hat{\beta} = \hat{B}\left\{\sum_{i=1}^{N}\sum_{t=1}^{\mathrm{T}}\frac{(Y_{it+1} - \theta v_{it})X_{it}}{\tau^2\sigma v_{it}} + B_0^{-1}\beta_0\right\}, \tag{10}$$

$$\hat{B}^{-1} = \sum_{i=1}^{N}\sum_{t=1}^{\mathrm{T}}\frac{X_{it}X_{it}^{\top}}{\tau^2\sigma v_{it}} + B_0^{-1}. \tag{11}$$

(3) Sample σ from $\mathrm{IG}(\hat{n}/2, \hat{s}/2)$, where

$$\hat{n} = 3NT + n_0, \tag{12}$$

$$\hat{s} = \sum_{i=1}^{N}\sum_{t=1}^{\mathrm{T}}\frac{(Y_{it+1} - \beta^{\top}X_{it} - \theta v_{it})^2}{\tau^2 v_{it}} + 2\sum_{i=1}^{N}\sum_{t=1}^{\mathrm{T}} v_{it} + s_0. \tag{13}$$

The MCMC algorithm for the quantile regression model is constructed applying the data augmentation technique as [6]. From Equations (5) and (6) and the assumptions for some prior distributions, a tractable and efficient Gibbs sampler can be proposed as above. In addition, the proposed Gibbs sampler sample v_{it} from the generalized inverse Gaussian distribution. Efficient algorithms to simulate from the generalized inverse Gaussian distribution exist, see, for example, Dagpunar [12] and Hörman *et al.* [25], but our proposed Gibbs sampler is implemented easily without any further need for tuning. Similar to Alhamzawi and Yu [1], all those similar results and our assumptions can, applying Ghosh *et al.* [19] and Sriram *et al.* [43], guarantee the rationality of the MCMC algorithm mentioned above.

4.3. Marginal likelihood

The marginal likelihood, $m(Y)$, of the panel data model is defined as

$$m(Y) = \int f(Y \mid \eta)\pi(\eta)\,\mathrm{d}\eta, \tag{14}$$

where $f(Y \mid \eta)$ is the sampling density of the data $\{Y\}$ and $\pi(\eta)$ is the prior of the model specific parameter η.

The marginal likelihood, $m(Y)$, can be reformulated as

$$m(Y) = \frac{f(Y \mid \eta)\pi(\eta)}{\pi(\eta \mid Y)}, \tag{15}$$

from which Chib [7] suggests to estimate the marginal likelihood as follows:

$$\log m(Y) = \log f(Y \mid \eta^*) + \log \pi(\eta^*) - \log \pi(\eta^* \mid Y), \tag{16}$$

where η^* is a particular high density point, typically the posterior mean or mode.

For $\eta \equiv \{\beta, \sigma\}$ and $Y \equiv \{Y_{it}\}$ in the panel data model, the posterior ordinate $\pi(\eta^*|Y)$ is estimated by the following decomposition:

$$\pi(\eta^* \mid Y) = \pi(\sigma^* \mid Y)\pi(\beta^* \mid \sigma^*, Y), \tag{17}$$

marginalized over the latent variable $v \equiv \{v_{it}\}$, since the ordinates $\pi(\sigma^* \mid Y)$ and $\pi(\beta^* \mid \sigma^*, Y)$ can be estimated according to [7]. The likelihood ordinate, $f(Y \mid \eta^*)$, can be estimated by Chib method.

4.4. Variable selection

To perform the variable selection for the quantile regression, an indicator vector is defined as follows. $\gamma \equiv (\gamma_0, \gamma_1, \ldots, \gamma_{25})$, where $\gamma_0 = 1$, and $\gamma_i = 1$ for $i \geq 1$ if β_i is included in the model (i.e. $\beta_i \neq 0$), and $\gamma_i = 0$ for $i \geq 1$ if β_i is excluded in the model (i.e. $\beta_i = 0$).

Given the indicator γ, k_γ denote the size of the γth subset model, $k_\gamma = \gamma^\top 1$, and β_{k_γ} and $X_{k_\gamma, it}$ are $k_\gamma \times 1$ vectors corresponding to all the components of β and X_{it} such that the corresponding γ_i's are equal to 1. Given γ, the following prior assumptions are supposed.

(1) $\beta_{k_\gamma} | \sigma, \nu \sim \mathrm{N}(\beta_0, 2\sigma(X_{k_\gamma}^\top V X_{k_\gamma})^{-1})$, where $p(\sigma) \propto \sigma^{-1}$ and each $\nu_i \sim \varepsilon(\sigma/p(1-p))$.
(2) A prior distribution over model space γ is given by $p(\gamma \mid \pi) \propto \pi^{k_\gamma}(1-\pi)^{k-k_\gamma}$.
(3) $\pi \sim \mathrm{beta}(a_0, b_0)$.

Given γ and the prior assumptions above, there are several ways to develop, for examples, (a) a tractable and efficient Gibbs sampler can be proposed applying the data augmentation technique as [6], similarly to Section 4.2, then compare the posterior model probabilities for different γ; (b) following Smith and Kohn [42], Kuo and Mallick [33], Krishna *et al.* [32], Zou and Yuan [57], Wu and Liu [48], Alhamzawi and Yu [1], or Yu *et al.* [50], an efficient Gibbs sampler can be proposed for computing posterior model probabilities in quantile regression, which we will follow next.

Under the prior assumptions, a MCMC algorithm can be developed to compute posterior model probabilities in the quantile regression by running the Gibbs sampler, and the marginal likelihood of Y under model γ can be obtained by integrating out β_{k_γ} and σ

$$p(Y \mid \gamma, \nu, X) \propto \int p(\sigma)\,\mathrm{d}\sigma \int p(Y \mid \beta_{k_\gamma}, \gamma, \sigma, \nu, X) p(\beta_{k_\gamma} \mid \gamma, \sigma, \nu) p(\nu \mid \sigma)\,\mathrm{d}\beta_{k_\gamma}. \tag{18}$$

Integrating out β_{k_γ} and σ as a normal integral and an inverse gamma integral

$$Y \mid \gamma, \nu, X \sim t_{(2n)}\{X_{k_\gamma}\beta_0 + \xi\nu, \tfrac{1}{2}(V + V X_{k_\gamma}(X_{k_\gamma}^\top V X_{k_\gamma})^{-1} X_{k_\gamma}^\top V)\}. \tag{19}$$

Then, the Gibbs sampler can be implemented [32,42] to generate samples of

$$p(Y \mid \gamma, \nu, X) \propto p(Y, \gamma, \nu, X) p(\gamma \mid \pi). \tag{20}$$

5. Real data application

In this section, the Bayesian quantile regression is applied to analyze the BCS data. This data set was extensively investigated for many sorts of topics, but this paper examines the

influence of maternal health problems on child's worrying status. There are 3426 observations, 25 predictor variables, and 1 response variable. We assume the quantile regression model between the response variable and the 25 covariates, plus an intercept.

In Table 1, upon the Bayesian quantile regression applying the MCMC package in R [36], the model is evaluated at three different quantiles 0.05, 0.50, and 0.95. The maternal health problems have different influence on child's worrying status at different quantiles, through MCMC quantile regression iteration 50,001 of 51,000, in detail, β_i have different estimates at different quantiles for each $i = 0, \ldots, 25$. β_{24} and β_{25} have the biggest absolute value for the three quantiles, except for β_0.

Upon the Bayesian quantile regression applying the MCMC package in R [36], iterations $= 1001 : 50991$, thinninginterval $= 10$, numberofchains $= 1$, sample size per chain $= 5000$. Table 2 summarizes the empirical mean and standard deviation for each variable X_i $(i = 1, \ldots, 25)$, and standard error of the mean for the model at the quantile 0.05. In this case, X_{24} has the biggest standard deviation, and X_{25} has the next biggest standard deviation. Table 3 summarizes the quantiles for each variable X_i $(i = 1, \ldots, 25)$.

Tables 4 and 5 summarize the same contents for the quantile 0.50, and Tables 6 and 7 for the quantile 0.95.

Applying the stochastic search variable selection [36], quantreg iteration 50,001 of 51,000, the top models and the posterior model probabilities are summarized in Tables 8–10 for the different quantiles 0.05, 0.50, and 0.95. From the posterior model probabilities applying the stochastic search variable selection, SSVSquantreg, the top models picked have significantly different posterior model probabilities, and, in particular, the maternal nervous breakdown, X_{24}, and the other health problems, X_{25}, are the first two important

Table 1. β for the quantile q =0.05, 0.50, 0.95 (all figures e−3 units).

	$q = 0.05$	$q = 0.50$	$q = 0.95$
β_0	1126.80	2909.93	6219.29
β_1	5.13	0.56	0.95
β_2	−3.30	0.05	−8.85
β_3	0.23	−0.41	−0.30
β_4	−1.11	0.25	−3.58
β_5	−4.88	−0.09	0.93
β_6	−0.10	−0.20	−2.80
β_7	2.20	−0.55	−3.76
β_8	−1.81	2.09	1.86
β_9	−0.41	−1.19	−5.94
β_{10}	2.22	0.28	0.06
β_{11}	−14.86	−3.09	−7.68
β_{12}	−13.23	−0.79	−2.01
β_{13}	13.86	0.79	6.21
β_{14}	0.96	0.27	5.51
β_{15}	6.42	1.49	−6.35
β_{16}	2.87	0.41	−5.71
β_{17}	2.75	0.54	3.07
β_{18}	−0.85	−0.38	3.42
β_{19}	−3.20	0.32	2.77
β_{20}	6.24	−1.07	1.86
β_{21}	4.43	0.74	3.21
β_{22}	1.31	0.50	0.54
β_{23}	−3.54	0.10	−5.09
β_{24}	−194.69	63.94	317.94
β_{25}	79.96	−40.22	−289.67

Table 2. Empirical mean and standard deviation for each variable, and standard error of the mean for the quantile $q = 0.05$ (all figures e–3 units).

	Mean	SD	Naive SE	Time-series SE
(Intercept)	80,960.000	71,879.300	1017.000	1106.000
X_1	261.900	213.400	3.018	3.491
X_2	−149.800	237.600	3.360	3.822
X_3	185.400	340.400	4.814	5.271
X_4	−75.700	245.800	3.476	3.877
X_5	−254.900	257.300	3.638	4.211
X_6	−157.500	267.900	3.789	3.952
X_7	163.800	273.500	3.868	4.166
X_8	−186.800	461.000	6.519	7.460
X_9	−6554.000	335.300	4.742	5.084
X_{10}	1507.000	290.300	4.106	4.513
X_{11}	−313.300	557.300	7.881	8.479
X_{12}	−329.200	533.100	7.539	8.646
X_{13}	38.260	472.600	6.684	7.343
X_{14}	−4.005	288.400	4.079	4.303
X_{15}	237.800	352.400	4.984	5.331
X_{16}	49.760	423.700	5.992	6.617
X_{17}	−163.300	379.400	5.365	6.031
X_{18}	4.134	425.900	6.023	6.681
X_{19}	188.400	383.400	5.423	5.706
X_{20}	200.100	429.500	6.074	6.698
X_{21}	511.500	445.400	6.298	7.015
X_{22}	−145.200	456.700	6.459	6.873
X_{23}	50.030	266.600	3.771	3.990
X_{24}	8781.000	29,472.100	416.800	449.800
X_{25}	894.100	15,204.000	215.000	225.300

Table 3. Quantiles for each variable when the quantile $q = 0.05$ (all figures e–3 units).

	2.5%	25%	50%	75%	97.5%
X_1	−1.34700	1.13700	2.54300	3.98620	6.93500
X_2	−6.33800	−3.02200	−1.45200	0.12610	3.02100
X_3	−4.75700	−0.49200	1.84600	4.15660	8.71300
X_4	−5.81500	−2.35300	−0.73510	0.89110	3.99100
X_5	−7.70100	−4.26700	−2.55100	−0.78220	2.43500
X_6	−6.94900	−3.34100	−1.56900	0.25070	3.75800
X_7	−3.83500	−0.13280	1.64700	3.46200	7.04700
X_8	−10.26800	−5.08700	−2.13500	1.18830	7.48600
X_9	−7.25500	−2.92700	−0.70160	1.57420	6.00400
X_{10}	−5.59700	−1.81500	0.23440	2.10240	5.73600
X_{11}	−13.57600	−6.94100	−3.33700	0.59920	8.25000
X_{12}	−13.38400	−6.96400	−3.31500	0.28870	7.30000
X_{13}	−8.44000	−2.85800	0.24870	3.44650	10.18400
X_{14}	−5.73800	−1.93200	−0.06721	1.90910	5.67600
X_{15}	−3.93900	−0.09158	2.16200	4.58360	9.75500
X_{16}	−7.83700	−2.42900	0.44340	3.29010	8.77200
X_{17}	−9.35300	−4.09600	−1.53500	0.91150	5.69400
X_{18}	−8.17500	−2.86500	−0.01683	2.89110	8.58500
X_{19}	−5.68900	−0.60500	1.91100	4.39410	9.39800
X_{20}	−6.46200	−0.87610	1.97500	4.88770	10.38500
X_{21}	−3.12100	2.08200	4.94800	7.97480	14.39400
X_{22}	−10.27300	−4.54000	−1.46500	1.52430	7.74900
X_{23}	−4.87600	−1.25900	0.54290	2.24970	5.87200
X_{24}	−475.64400	−100.40000	74.99000	264.89090	698.47500
X_{25}	−292.12600	−91.20000	7.40400	108.54190	310.32500

Table 4. Empirical mean and standard deviation for each variable, and standard error of the mean for the quantile $q = 0.50$ (all figures e–4 units).

	Mean	SD	Naive SE	Time-series SE
(Intercept)	29,510.00000	1917.03100	27.11000	27.11000
X_1	0.66020	4.70700	0.06656	0.06889
X_2	0.42350	4.45500	0.06300	0.06300
X_3	2.91500	7.11300	0.10060	0.10060
X_4	−1.09500	4.83300	0.06835	0.06898
X_5	−1.02500	4.17200	0.05899	0.05899
X_6	−0.02617	6.51800	0.09217	0.09471
X_7	−1.56800	6.86200	0.09704	0.09704
X_8	2.16700	12.10100	0.17110	0.17110
X_9	−1.96000	7.36500	0.10420	0.10420
X_{10}	−45.60000	5.74800	0.08129	0.08129
X_{11}	−5.42100	13.93300	0.19700	0.19700
X_{12}	−6.85000	12.46000	0.17620	0.17250
X_{13}	2.50500	12.51200	0.17700	0.17700
X_{14}	−1.28200	6.02300	0.08517	0.08517
X_{15}	1.26500	9.32900	0.13190	0.13190
X_{16}	1.27600	9.62700	0.13610	0.13810
X_{17}	−0.27990	7.54500	0.10670	0.10670
X_{18}	2.56600	9.28200	0.13130	0.13130
X_{19}	1.81300	11.09900	0.15700	0.15350
X_{20}	−4.30400	10.52200	0.14880	0.14880
X_{21}	2.18700	11.71400	0.16570	0.16950
X_{22}	3.21700	9.07700	0.12840	0.11980
X_{23}	1.50600	6.13100	0.08671	0.08671
X_{24}	489.50000	832.31100	0.11770	11.77000
X_{25}	−172.10000	360.37700	5.09600	5.09600

Table 5. Quantiles for each variable when the quantile $q = 0.50$ (all figures e–4 units).

	2.5%	25%	50%	75%	97.5%
(Intercept)	25,445.4100	28,376.7290	29,630.0000	30,743.89190	33,003.4600
X_1	−8.6250	−2.4480	0.5838	3.6300	10.1710
X_2	−8.5120	−2.4630	0.4258	3.2990	9.3610
X_3	−10.8590	−1.7350	2.7430	7.4820	17.5130
X_4	−11.0650	−4.2140	−1.1040	2.0350	8.1710
X_5	−9.8440	−3.6330	−0.8936	1.6830	7.0560
X_6	−12.7880	−4.2380	−0.0366	4.2310	12.7960
X_7	−15.3660	−5.8820	−1.4430	3.0750	11.6360
X_8	−21.2970	−5.3980	2.0210	9.5740	26.5670
X_9	−16.9570	−6.6250	−1.9170	3.0030	12.1740
X_{10}	−12.1040	−4.0950	−0.2667	3.3190	10.6870
X_{11}	−33.8050	−14.5850	−4.8640	3.8510	21.6330
X_{12}	−34.0110	−14.3300	−6.1220	1.5110	16.1080
X_{13}	−22.5280	−5.1120	2.2030	9.8840	27.8070
X_{14}	−13.8990	−5.0690	−1.1530	2.6480	10.1430
X_{15}	−16.8440	−4.6930	1.0780	7.0880	20.4560
X_{16}	−17.3840	−5.0210	1.0840	7.3910	20.8090
X_{17}	−15.8120	−4.9400	−0.2400	4.6070	14.7560
X_{18}	−15.0820	−3.3770	2.3710	8.3070	21.7620
X_{19}	−19.4420	−5.2890	1.5910	8.5320	25.0950
X_{20}	−26.2960	−11.1120	−4.0050	2.6800	15.8220
X_{21}	−21.3700	−5.5090	2.1500	9.5880	25.7640
X_{22}	−14.1720	−2.7470	3.0950	9.0090	21.8540
X_{23}	−10.5890	−2.4880	1.3630	5.4100	13.9490
X_{24}	−968.4610	−48.3330	402.5000	955.8340	2391.5660
X_{25}	−927.5700	−399.4760	−156.6000	68.2750	502.9410

Table 6. Empirical mean and standard deviation for each variable, and standard error of the mean for the quantile $q = 0.95$ (all figures e−5 units).

	Mean	SD	Naive SE	Time-series SE
(Intercept)	543,695.460	89,426.900	1265.000	1526.000
X_1	−26.660	263.000	3.720	4.410
X_2	−315.410	285.900	4.044	5.278
X_3	56.890	380.900	5.387	6.252
X_4	−272.060	277.500	3.924	4.794
X_5	−188.940	267.000	3.776	4.756
X_6	209.250	330.700	4.677	5.667
X_7	−219.800	321.400	4.546	5.336
X_8	114.880	493.900	6.985	7.819
X_9	−323.280	383.700	5.426	6.256
X_{10}	−23.130	344.300	4.869	5.876
X_{11}	107.880	587.300	8.305	9.311
X_{12}	−288.510	506.800	7.167	7.714
X_{13}	−182.250	502.800	7.111	7.820
X_{14}	−119.030	348.300	4.925	5.872
X_{15}	−180.200	426.800	6.036	7.686
X_{16}	45.020	449.200	6.353	7.070
X_{17}	46.290	382.700	5.412	6.318
X_{18}	40.220	451.800	6.389	7.439
X_{19}	−283.000	463.500	6.555	7.313
X_{20}	−340.210	457.600	6.472	7.280
X_{21}	5380.900	451.000	6.378	7.051
X_{22}	596.060	476.700	6.742	7.828
X_{23}	−69.550	327.200	4.627	5.620
X_{24}	11,901.210	32,910.100	465.400	526.100
X_{25}	−17,966.530	18,277.500	258.500	324.300

Table 7. Quantiles for each variable when the quantile $q = 0.95$ (all figures e−3 units).

	2.5%	25%	50%	75%	97.5%
(Intercept)	3783.021000	4822.000000	5411.142900	6011.000000	7310.987000
X_1	−5.526000	−2.044000	−0.182300	1.547000	4.797000
X_2	−8.516000	−5.127000	−3.207600	−1.260000	2.565000
X_3	−7.097000	−2.014000	0.679300	3.184000	7.869000
X_4	−8.036000	−4.580000	−2.731600	−79.930000	2.706000
X_5	−7.234000	−3.644000	−1.910700	−705.400000	3.189000
X_6	−4.811000	−0.073560	2.210700	4.391000	8.246000
X_7	−8.675000	−4.335000	−2.161400	0.007187	3.938000
X_8	−8.859000	−2.114000	1.186400	4.435000	10.403000
X_9	−10.886000	−5.801000	−3.227300	−0.648600	4.243000
X_{10}	−7.099000	−2.538000	−0.192100	2.090000	6.376000
X_{11}	−11.012000	−2.672000	1.329400	5.071000	11.764000
X_{12}	−13.181000	−6.219000	−2.721400	0.581100	6.605000
X_{13}	−12.166000	−5.053000	−1.754700	1.648000	7.593000
X_{14}	−8.183000	−3.535000	−1.173300	1.203000	5.571000
X_{15}	−10.492000	−4.647000	−1.685700	1.147000	6.205000
X_{16}	−8.687000	−2.559000	0.555000	3.577000	8.928000
X_{17}	−7.463000	−2.024000	0.643100	3.161000	7.391000
X_{18}	−8.881000	−2.501000	0.658700	3.484000	8.691000
X_{19}	−12.285000	−5.818000	−2.713800	0.417100	5.652000
X_{20}	−12.563000	−6.433000	−3.268300	−0.329300	5.214000
X_{21}	−4.353000	2.590000	5.707900	8.514000	13.439000
X_{22}	−3.864000	2.872000	6.214400	9.278000	14.778000
X_{23}	−7.358000	−2.825000	−0.538400	1.571000	5.350000
X_{24}	−90.350000	−88.010000	147.516800	49.500000	696.594000
X_{25}	−555.750000	−300.300000	−172.533700	−49.340000	153.451000

Table 8. Variable selection for the quantile $q = 0.05$.

Models	Probability
(Intercept)	.9278
X_{24}	.0502
(Intercept), X_{24}	.0142
(Intercept), X_{25}	.0052
(Intercept), X_3	.0004

Table 9. Variable selection for the quantile $q = 0.50$.

Models	Probability
(Intercept)	.9954
(Intercept), X_{24}	.0040
(Intercept), X_{25}	.0004
(Intercept), X_2	.0002

Table 10. Variable selection for the quantile $q = 0.95$.

Models	Probability
(Intercept)	.9274
(Intercept), X_{24}	.0486
(Intercept), X_{25}	.0146
(Intercept), X_{20}	.0012
(Intercept), X_2	.0010

to influence child's worrying status. This indicates that the maternal nervous breakdown and the other health problems need be made enough attention to intervene early for the influence on child's worrying status.

6. Conclusions

In this paper, we developed a Bayesian quantile regression for linear panel data model without heterogeneity, in particular, upon a location-scale mixture representation of the asymmetric Laplace error distribution, this paper provides how the posterior distribution can be sampled and summarized by a MCMC method.

In addition, the influence of maternal health problems on child's worrying status was explored by this method to the 1970 BCS data, and we find that different maternal health problem has different influence on child's worrying status at different quantiles, also that maternal nervous breakdown and the other maternal health problem, by our method, are the first two important to influence the child's worrying status.

Our findings have high policy relevance in terms of the importance of the intervention of maternal nervous breakdown early for the influence on child's worrying status.

Disclosure Statement

No potential conflict of interest was reported by the authors.

Funding

This research was supported by China State Education Ministry through China Scholarship Council (201208420430), Deutsche Forschungsgemeinschaft through the SFB 649 'Economic Risk', and IRTG 1792 'High Dimensional Non Stationary Time Series'.

References

[1] R. Alhamzawi and K. Yu, *Conjugate priors and variable selection for Bayesian quantile regression*, Comput. Stat. Data Anal. 64 (2013), pp. 209–219.
[2] O.E. Barndorff-Nielsen and N. Shephard, *Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics*, J. R. Stat. Soc. Ser. B 63 (2001), pp. 167–241.
[3] J. Blanden, P. Gregg, and L. Macmillan, *Accounting for intergenerational income persistence: Non-cognitive skills, ability and education*, Econ. J. 117 (2007), pp. 43–60.
[4] M. Buchinsky, *Recent advances in quantile regression models: A practical guideline for empiricalresearch*, J. Hum. Resour. 33 (1998), pp. 88–126.
[5] J. Campbell and M. Yogo, *Efficient tests of stock return predictability*, J. Financ. Econ. 81 (2006), pp. 27–60.
[6] S. Chib, *Bayes inference in the Tobit censored regression model*, J. Econ. 51 (1992), pp. 79–99.
[7] S. Chib, *Marginal likelihood from the Gibbs output*, J. Am. Stat. Assoc. 90 (1995), pp. 1313–1321.
[8] S. Chib, *Markov chain Monte Carlo methods: Computation and inference*, in *Handbook of Econometrics*, Vol. 5, J.J. Heckman and E. Leamer, eds., North-Holland, Amsterdam, 2001, pp. 3569–3649.
[9] G. Conti, S. Frühwirth-Schnatter, J.J. Heckman, and R. Piatek, *Bayesian exploratory factor analysis*, J. Econ. 183 (2014), pp. 31–57.
[10] G. Conti and J.J. Heckman, *Understanding theearly origins of the education-health gradient: A framework that can also be applied to analyze gene-environment interactions*, Perspect. Psychol. Sci. 5 (2010), pp. 585–605.
[11] D.M. Cutler and A. Lleras-Muney, *Understanding differences in health behaviors by education*, J. Health Econ. 29 (2010), pp. 1–28.
[12] J.S. Dagpunar, *An easily implementedgeneralised inverse Gaussian generator*, Commun. Stat. Simul. Comput. 18 (1989), pp. 703–710.
[13] X. Dai, *Optimal taxation under income uncertainty*, Ann. Econ. Financ. 12 (2011), pp. 121–138.
[14] X. Dai and J.J. Heckman, *Older siblings' contributions to young child's cognitive skills*, Econ. Model. 35(C) (2013), pp. 235–248.
[15] X. Dai, H. Li, and Y. Wang, *A predictive functional regression model for asset return*, J. Math. Financ. 3 (2013), pp. 307–311.
[16] L. Feinstein, *The Relative Economic Importance of Academic, Psychological and Behavioural Attributes Developed on Chilhood*, CEP Discussion Paper, 2000.
[17] S. Frühwirth-Schnatter and H.F. Lopes, *Parsimonious Bayesian Factor Analysis When the Number of Factors Is Unknown*, Unpublished Tech. Report, 2009.
[18] D. Gamerman and H.F. Lopes, *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, 2nd ed., Chapman and Hall/CRC, Boca Raton, FL, 2006.
[19] J.K. Ghosh, D. Mohan, and S. Tapas, *An Introduction to Bayesian Analysis: Theory and Methods*, Springer, New York, 2006.
[20] M. Guo and W.K. Härdle, *Simultaneous confidence bands for expectile functions*, Adv. Stat. Anal. 96 (2012), pp. 517–541.
[21] S.E. Hampson and H.S. Friedman, *Personality and health: A lifespan perspective*, in *The Handbook of Personality: Theory and Research*, 3rd ed., O.P. John, R. Robins, and L. Pervin, eds., Guilford, New York, 2008, pp. 770–794.
[22] T. Hanson and W.O. Johnson, *Modeling regression error with a mixture of Polya trees*, J. Am. Stat. Assoc. 97 (2002), pp. 1020–1033.
[23] W.K. Härdle and S. Song, *Confidence bands in quantile regression*, Economet. Theory 26(04) (2010), pp. 1180–1200.

[24] J.J. Heckman, J. Stixrud, and S. Urzua, *The effects of cognitive andnoncognitive abilities on labor market outcomes and social behavior*, J. Lab. Econ. 24 (2006), pp. 411–482.
[25] W. Hörman, J. Leydold, and G. Derflinger, *Automatic Nonuniform Random Variable Generation*, Springer, Berlin, 2004.
[26] R. Kaestner and K. Callison, *Adolescent cognitive andnoncognitive correlates of adult health*, J. Hum. Capital. 5 (2011), pp. 29–69.
[27] R. Kanbur, J. Pirttila, and M. Tuomala, *Moral hazard, income taxation and prospect theory*, Scand. J. Econ. 110 (2008), pp. 321–337.
[28] R. Koenker, *Quantile Regression*, Cambridge University Press, New York, 2005.
[29] A. Kottas and A.E. Gelfand, *Bayesian semiparametric median regression modeling*, J. Am. Stat. Assoc. 96 (2001), pp. 1458–1468.
[30] A. Kottas and M. Krnjajic, *Bayesian semiparametric modelling in quantile regression*, Scand. J. Stat. 36 (2009), pp. 297–319.
[31] H. Kozumi and G. Kobayashi, *Gibbs sampling methods for Bayesian quantile regression*, J. Stat. Comput. Simul. 81 (2010), pp. 1565–1578.
[32] A. Krishna, H.D. Bondell, and S.K. Ghosh, *Bayesian variable selection using an adaptive powered correlation prior*, J. Stat. Plan. Inference 139 (2008), pp. 2665–2674.
[33] L. Kuo and B. Mallick, *Variable selection for regression models*, Sankhya B60 (1998), pp. 65–81.
[34] J.S. Liu, *Monte Carlo Strategies in Scientific Computing*, Springer, New York, 2001.
[35] J.E. Murasko, *A lifecourse study on education and health: The relationshipbetween childhood psychosocial resources and outcomes in adolescence and young adulthood*, Soc. Sci. Res. 36 (2007), pp. 1348–1370.
[36] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, 2011. Available at http://www.R-project.org.
[37] B.J. Reich, H.D. Bondell, and H.J. Wang, *Flexible Bayesian quantile regression for independent and clustered data*, Biostatistics 11 (2010), pp. 337–352.
[38] B.W. Roberts, P. Harms, J.L. Smith, D. Wood, and M. Webb, *Using multiple methods in personality psychology*, in *Handbook of Multimethod Measurement in Psychology*, M. Eid and E. Diener, eds., American Psychological Association, Washington, DC, 2006, pp. 321–335.
[39] B.W. Roberts, N.R. Kuncel, R.L. Shiner, A. Caspi, and L.R. Goldberg, *The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes*, Perspect. Psychol. Sci. 2 (2007), pp. 313–345.
[40] S.K. Schnabel and P.H.C. Eilers, *An analysis of life expectancy and economic production using expectile frontier zones*, Demogr. Res. 21 (2009), pp. 109–134.
[41] S.K. Schnabel and P.H.C. Eilers, *Optimal expectile smoothing*, Comput. Stat. Data Anal. 53 (2009), pp. 4168–4177.
[42] M. Smith and R. Kohn, *Nonparametric regression using Bayesian variable selection*, J. Econ. 75 (1996), pp. 317–343.
[43] K. Sriram, R.V. Ramamoorthi, and P. Ghosh, *Posterior Consistency of Bayesian Quantile Regression under a Mis-specified Likelihood Based on Asymmetric Laplace Density*, Indian Institute of Management Bangalore and Michigan State University, 2011.
[44] M.A. Taddy and A. Kottas, *A Bayesian nonparametric approach to inference for quantile regression*, J. Bus. Econ. Stat. 28 (2010), pp. 357–369.
[45] J. Taylor, *Estimating value at risk and expected shortfall using expectiles*, J. Financ. Econ. 6 (2008), pp. 231–252.
[46] E.G. Tsionas, *Bayesian quantile inference*, J. Stat. Comput. Simul. 73 (2003), pp. 659–674.
[47] S.G. Walker and B.K. Mallick, *A Bayesian semiparametric accelerated failure time model*, Biometrics 55 (1999), pp. 477–483.
[48] Y. Wu and Y. Liu, *Variable selection in quantile regression*, Stat. Sin. 19 (2009), p. 801.
[49] K. Yu, *Quantile regression using RJMCMC algorithm*, Comput. Stat. Data Anal. 40 (2002), pp. 303–315.
[50] K. Yu, C.W.S. Chen, C. Reed, and D.B. Dunson, *Bayesian variable selection in quantile regression*, Stat. Interface 6 (2013), pp. 261–274.

[51] K. Yu, Z. Lu, and J. Stander, *Quantile regression: Applications and current research area*, Statistics 52 (2003), pp. 331–350.
[52] K. Yu and R.A. Moyeed, *Bayesian quantile regression*, Stat. Probab. Lett. 54 (2001), pp. 437–447.
[53] K. Yu and J. Stander, *Bayesian analysis of a Tobit quantile regression model*, J. Econ. 137 (2007), pp. 260–276.
[54] K. Yu and J. Zhang, *Athree-parameter asymmetric Laplace distribution and its extension*, Commun. Stat. Theory Methods 34 (2005), pp. 1867–1879.
[55] Y. Yuan and G. Yin, *Bayesian quantile regression for longitudinal studies with nonignorable missing data*, Biometrics 66 (2010), pp. 105–114.
[56] Y.R. Yue and H. Rue, *Bayesian inference for additive mixed quantile regression models*, Comput. Stat. Data Anal. 55 (2011), pp. 84–96.
[57] H. Zou and M. Yuan, *Composite quantile regression and the oracle model selection theory*, Ann. Stat. 36 (2008), pp. 1108–1126.
[58] W.R. Zwick and W.F. Velicer, *Comparison of five rules for determining the number of components to retain*, Psychol. Bull. 99 (1986), pp. 432–442.

Scandinavian Journal of Statistics Theory and Applications

Scandinavian Journal of Statistics, Vol. 43: 1140–1152, 2016
doi: 10.1111/sjos.12233

An Extended Single-index Model with Missing Response at Random

QIHUA WANG
Institute of Statistical Science, Shenzhen University

Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences

TAO ZHANG
School of Science, Guangxi University of Science and Technology

WOLFGANG KARL HÄRDLE
School of Business, Singapore Management University

Center for Applied Statistics and Economics (CASE), Humboldt-Universität zu, Berlin

ABSTRACT. An extended single-index model is considered when responses are missing at random. A three-step estimation procedure is developed to define an estimator for the single-index parameter vector by a joint estimating equation. The proposed estimator is shown to be asymptotically normal. An algorithm for computing this estimator is proposed. This algorithm only involves one-dimensional nonparametric smoothers, thereby avoiding the data sparsity problem caused by high model dimensionality. Some simulation studies are conducted to investigate the finite sample performances of the proposed estimators.

Key words: asymptotic normality, estimating equations, missing data, single-index models

1. Introduction

The single-index model has been paid considerable attention recently because it is useful in several areas of science such as econometrics, biostatistics, finance and so on. The single-index model, which is investigated extensively, is of the following form:

$$Y = g\left(\beta^\top X\right) + \varepsilon, \tag{1}$$

where Y is the univariate response and X is a d-dimensional covariable vector, β is an unknown index parameter vector of interest, the function $g(\cdot)$ is an unknown link function and $\mathsf{E}(\varepsilon|X) = 0$. The single-index model provides dimension reduction in the sense that, if one can estimate the index β efficiently, the univariate index $\beta^\top X$ serves as a covariable to estimate the nonparametric link $g(\cdot)$. Much effort has been devoted to estimating the index β efficiently. Hall (1989), Zhu & Fang (1992) considered a projection pursuit framework. Härdle *et al.* (1993) employed the kernel smoothing method to study model (1) and gave an empirical rule for bandwidth selection. Ichimura (1993) studied the properties of a semi-parametric least-squares estimator in a general single-index model. Ichimura (1987) showed that the parameter vector β can be estimated root-n consistently. Härdle *et al.* (1993) and Hristache *et al.* (2001) obtained a $\sqrt{n}$ consistent estimator of the index vector β using the average derivative method. The technology of sliced inverse regression can also be used to achieve $\sqrt{n}$ consistent estimator (Li, 1991; Zhu & Fang, 1996). Xue & Zhu (2006) constructed the confidence region of the regression parametric vector for single-index regression models using the empirical likelihood method and proved that an estimated empirical log-likelihood ratio is asymptotically a weighted sum of independent χ_1^2 variables with unknown weights. Chang *et al.* (2010) proposed an asymptotically more

efficient estimation of the single-index model in terms of transferring restricted least squares to unrestricted least squares. Zhu *et al.* (2014) considered estimation and hypothesis testing in single-index panel data models with individual effects and obtain a double robustness property.

Let (Y_i, X_i) denote the observed values with Y_i being the response variable and X_i being the vector of d explanatory variables. In this paper, we consider an extended single-index model (ESIM) which specifies the relationship of the mean and variance of Y_i as follows

$$\mathsf{E}(Y_i|X_i) = \mu\left\{g\left(\beta^\top X_i\right)\right\}, \mathsf{Var}(Y_i|X_i) = \sigma^2 V\left\{g\left(\beta^\top X_i\right)\right\}, \tag{2}$$

where $\mu(\cdot)$ is a known monotonic function, $V(\cdot)$ is a known covariance function, $g(\cdot)$ is an unknown univariate link function and β is an unknown index vector, which belongs to the parameter space $\Theta = \left\{\beta = (\beta_1, \ldots, \beta_d)^\top : \|\beta\| = 1, \beta_1 > 0, \beta \in \mathbb{R}^d\right\}$. Cui *et al.* (2011) developed a method of estimating function (EFM) to study the ESIM. They investigated the efficiency and computation of the estimates for the ESIM and obtained the asymptotic properties of the EFM. However, the existing work is for the case where data are observed fully.

In practice, some responses may be missing, by design (as in two-stage studies) or by circumstance. For example, the response $Y's$ may be very expensive to measure, and only part of $Y's$ are available. Another example is that the $Y's$ represent the responses to a set of questions and some sampled individuals refuse to supply the desired information. Actually, missingness of responses is very common in opinion polls, market research surveys, mail enquiries, social-economic investigations, medical studies and other scientific experiments. Missing data issues have been investigated extensively (e.g. Rosenbaum & Rubin (1983), Robins *et al.* (1994), Robins *et al.* (1995), Wang & Rao (2002), Wang *et al.* (2004) and among others). To the best of our knowledge, the literature reduces to just a few recent papers for the single-index models (1) with $\mu\left\{g\left(\beta^\top X_i\right)\right\} = g\left(\beta^\top X_i\right)$ and $V\left\{g\left(\beta^\top X_i\right) = 1\right.$ for missing data. For this special case, Wang *et al.* (2010) derived semi-parametric nonlinear least squares estimators with complete case (CC) method by incorporating missing mechanism into the least-squares loss function suggested by Härdle *et al.* (1993) and minimizing the loss function with respect to the bandwidth and the parameters simultaneously. They obtained the central limit theorem, the law of the iterated logarithm for the estimator of β and the optimal convergence rate for the estimator of $g(\cdot)$. However, the computational burden of solving the minimization problem is very high when the dimension of explanatory variable vector is large.

In this paper, we extend the EFM due to Cui *et al.* (2011) to the missing response case for estimating both β and $g(\cdot)$ in model (2). That is, we consider the case where some Y-values may be missing and X is observed completely. The data we observe are

$$\{(Y_i, \delta_i, X_i)\}_{i=1}^n,$$

where $\delta_i = 0$ if Y_i is missing, otherwise $\delta_i = 1$. Throughout this paper, it is assumed that Y is missing at random (MAR). The MAR assumption implies that δ and Y are conditionally independent given X. That is, $\mathsf{P}(\delta = 1|Y, X) = \mathsf{P}(\delta = 1|X)$. MAR is a common assumption for statistical analysis with missing data and is reasonable in many practical situations (Little & Rubin, 1987).

In this paper, we develop a three-step estimating approach for estimating both β and $g(\cdot)$ by extending the EFM due to Cui *et al.* (2011) to the missing response problem. Unlike the two-step estimating approach of Cui *et al.* (2011), the three-step estimating approach can define an estimator of $g(\cdot)$ in addition to defining an estimator of β. For the estimating approach, the estimating function system only involves one-dimensional nonparametric smoothers, thereby avoiding the data sparsity problem caused by high dimensionality. Firstly, unlike the method proposed by Wang *et al.* (2010) for the special case of the ESIM where the minimization is

difficult to implement when d is large, our method is easy to implement. Secondly, unlike the method proposed by Wang *et al.* (2010) where the methodology can only be applied to the case of homogeneous errors, our method can be applied to the case of heterogeneous errors. Hence, the proposed methodology based on model (2) has more wide application and much more flexible framework. Cui *et al.* (2011) define the estimator of β only when data are observed fully. However, we define the estimators of both β and $g(\cdot)$ and investigate their asymptotic properties with data missing. It is more challenging to investigate the asymptotic properties because of the estimator of $g(\cdot)$ and the treatment of missing data.

This paper is organized as follows. In Section 2, we describe the estimating procedures. In Section 3, we establish the asymptotic theory for the proposed procedure. Some simulation studies are provided in Section 4. In Section 5, we analyse a real data set to illustrate the proposed procedures, and all proofs are included in the Supporting Information.

2. Three-step estimation

We develop the following three-step approach to define the estimators of β and $g(\cdot)$, respectively.

Step 1: We use the nonparametric fusion-refinement approach to get the initial estimate of β, denoted by $\tilde{\beta}$ with $\|\tilde{\beta}\| = 1$ (Ding & Wang, 2011).

Step 2: Define the estimator of $g(\cdot)$ and $g'(\cdot)$.

Note that under MAR, we have

$$\mu\{g(t)\} = \mathsf{E}\left[\delta Y | \beta^\top X = t\right] / \mathsf{E}\left[\delta | \beta^\top X = t\right].$$

We then may obtain an initial estimator of $\mu\{g(t)\}$

$$\mu\left\{\tilde{g}(t)\right\} = \left(\sum_{j=1}^n \delta_j Y_j H_{h_n}\left(t - \tilde{\beta}^\top X_j\right)\right) / \left(\sum_{j=1}^n \delta_j H_{h_n}\left(t - \tilde{\beta}^\top X_j\right)\right),$$

where $H(\cdot)$ is a kernel function with support on $(-1, 1)$, h_n is a bandwidth sequence and $H_{h_n}(\cdot) = H(\cdot/h_n)$.

Denote by α_0 and α_1 the values of $g(\cdot)$ and $g'(\cdot)$ evaluating at $\beta^\top x$, respectively. The local linear approximation for $g\left(\beta^\top X\right)$ in a neighbourhood of $\beta^\top x$ is $g_0\left(\beta^\top X\right) = \alpha_0 + \alpha_1\left(\beta^\top X - \beta^\top x\right)$. The estimators $G\left(\beta^\top x\right) \stackrel{\text{def}}{=} \left(g\left(\beta^\top x\right), g'\left(\beta^\top x\right)\right)$ are obtained by solving the kernel estimating equations:

$$\begin{aligned}
&\sum_{j=1}^n K_{b_n}\left(\tilde{\beta}^\top X_j - \beta^\top x\right) \mu'\left\{g_0\left(\tilde{\beta}^\top X_j\right)\right\} V^{-1}\left\{g_0\left(\tilde{\beta}^\top X_j\right)\right\} \\
&\times\left[\delta_j Y_j + \left(1 - \delta_j\right) \mu\left\{\tilde{g}\left(\tilde{\beta}^\top X_j\right)\right\} - \mu\left\{g_0\left(\tilde{\beta}^\top X_j\right)\right\}\right] = 0, \\
&\sum_{j=1}^n \left(\tilde{\beta}^\top X_j - \beta^\top x\right) K_{b_n}\left(\tilde{\beta}^\top X_j - \beta^\top x\right) \mu'\left\{g_0\left(\tilde{\beta}^\top X_j\right)\right\} V^{-1}\left\{g_0\left(\tilde{\beta}^\top X_j\right)\right\}, \\
&\times\left[\delta_j Y_j + \left(1 - \delta_j\right) \mu\left\{\tilde{g}\left(\tilde{\beta}^\top X_j\right)\right\} - \mu\left\{g_0\left(\tilde{\beta}^\top X_j\right)\right\}\right] = 0
\end{aligned} \tag{3}$$

where $K_{b_n}(\cdot)$ is the symmetric kernel density function satisfying $K_{b_n}(\cdot) = K(\cdot/b_n)$ and b_n is a bandwidth, with respect to α_0 and α_1, yielding $\widehat{G}\left(\beta^\top x\right) = \left(\widehat{g}\left(\beta^\top x\right), \widehat{g}'\left(\beta^\top x\right)\right) = \left(\widehat{\alpha}_0, \widehat{\alpha}_1\right)$.

Step 3: Obtain the estimator of β. Similar to Cui *et al.* (2011), by eliminating β_1, the parameter space Θ can be rearranged to the form $\Theta = \left\{\left(1 - \sum_{r=2}^d \beta_r^2\right)^{1/2}, \beta_2 \ldots, \beta_d\right)^\top : \sum_{r=2}^d \beta_r^2 < 1\Big\}$.

We turn to the estimation of $\beta \in \Theta$. First, we estimate $\beta^{(1)} = (\beta_2, \ldots, \beta_d)$, which can be obtained by solving the following equation:

$$\begin{aligned} &\textstyle\sum_{j=1}^{n} \left[\partial\mu\left\{\widehat{g}\left(\beta^\top X_j\right)\right\}/\partial\beta^{(1)}\right] V^{-1}\left\{\widehat{g}\left(\beta^\top X_j\right)\right\} \\ &\times\left[\delta_j Y_j + \left(1-\delta_j\right)\mu\left\{\widetilde{g}\left(\beta^\top X_j\right)\right\} - \mu\left\{\widehat{g}\left(\beta^\top X_j\right)\right\}\right] = 0. \end{aligned} \tag{4}$$

The solution is defined as $\hat{\beta}^{(1)}$, and hence, we obtain $\hat{\beta}$ by the transformation.

From Ding & Wang (2011), the fusion-refinement approach performs better than the CC for estimating β. This is why we select the fusion-refinement estimator as the initial estimator instead of the CC estimator in the step one. Indeed, the simulation results show that the proposed estimators perform better when the initial value takes the fusion-refinement estimator. However, the fusion-refinement procedure due to Ding & Wang (2011) is a nonparametric method, and hence, the fusion-refinement estimator does not use the model information. This means that it is not suitable to use it directly as the final estimator for the model considered in this paper. However, it is fine to use it as the initial estimator for the three-step estimating method, which uses model information in Step 2 and Step 3, and hence, it is expected to define a more efficient estimator of β. One reviewer finds that the fusion-refinement approach in the paper relies on the assumption that the missing data mechanism model carries information about the full data model, which may need attention.

3. Asymptotic theory

To establish asymptotic theory, we firstly give some notations. Let $\rho_l(z) = \left\{\mu'(z)\right\}^l V^{-1}(z)$, $q_1(z,y) = \mu'(z)V^{-1}(z)\{y-\mu(z)\}$, $q_2(z,y) = \{y-\mu(z)\}\rho_1'(z) - \rho_2(z)$, $\pi(X) = \mathsf{P}(\delta = 1|X)$. Let

$$\gamma_j = \textstyle\int t^j K(t)dt \; and \; \nu_j = \int t^j K^2(t)dt, \quad j = 1, 2, \ldots$$

and $S = \begin{pmatrix} \gamma_0 & 0 \\ 0 & \gamma_2 \end{pmatrix}$, $S^* = \begin{pmatrix} \nu_0 & \nu_1 \\ \nu_1 & \nu_2 \end{pmatrix}$. Denote by $\beta^0 = \left(\beta_1^0, \beta^{(1)0\top}\right)^\top$ the true values of $\beta = \left(\beta_1, \beta^{(1)\top}\right)^\top$. Denote by $J = \frac{\partial\beta}{\partial\beta^{(1)}}$ the Jacobian matrix of size $d \times (d-1)$ with

$$J = \begin{pmatrix} -\beta^{(1)\top}/\sqrt{1-\|\beta^{(1)}\|^2} \\ I_{d-1} \end{pmatrix}.$$

Denote $C = (1-\delta)E\left\{X|\beta^\top X\right\} + \left(X - \mathsf{E}\left\{X|\beta^\top X\right\}\right) g'\left\{\left(\beta^\top X\right)\right\}$. Let

$$\begin{aligned} A &= J^\top \mathsf{E}\left[\rho_2\left\{g\left(\beta^\top X\right)\right\} C^\top C\right] J, \\ B &= J^\top \mathsf{E}\left[\delta\rho_2\left\{g\left(\beta^\top X\right)\right\}\sigma^2 C^\top C\right] J. \end{aligned}$$

In order to prove the asymptotic normality of the estimators, we also introduce some regularity conditions.

(a) $\mu(\cdot)$, $V(\cdot)$ and $g(\cdot)$ have bounded and continuous derivatives order two. $V(\cdot)$ is uniformly bounded and bounded away from 0.
(b) Assume that $q_2(z,y) < 0$ for $z \in \mathbb{R}$ and y in the range of the response variable.
(c) Define the block partition of matrix Ω as follows:

$$\Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix},$$

where Ω_{11} is a positive constant, Ω_{12} is a $(d-1)$-dimensional row vector, Ω_{21} is a $(d-1)$-dimensional column vector and Ω_{22} is a $(d-1)\times(d-1)$ nonnegative definite matrix. The largest eigenvalues of Ω_{22} is bounded away from infinity.

(d) The density function of X has a continuous second derivative on its support A. The density function $f_{\beta^\top X}\left(\beta^\top X\right)$ of random variable $\beta^\top X$ is bounded away from 0 on T_β and satisfies the Lipschitz condition of order 1 on T_β, where $T_\beta = \left\{\beta^\top X : X \in T\right\}$ and T is the compact support set of X.

(e) The kernel $K(\cdot)$ is a bounded and symmetric density function with a bounded derivative and satisfies

$$\int_{-\infty}^{+\infty} |t|^2 K(t)dt < \infty,$$

$H(\cdot)$ is a bounded kernel function of order 2 with bounded support.

(f) $\pi(\cdot) > 0$ and $\mu_{(\cdot)} \neq 0$.

We are ready to present the asymptotic results of the proposed estimators. The proof of the theorem is provided in the Supporting Information.

Theorem 1. *Suppose that conditions* $(a)-(f)$ *hold, if* $nb_n^4 \to 0$, $nh_n^4 \to 0$, $nh_n^2/\log(1/h_n) \to \infty$ *and* $nb_n^2 h_n^2 \to 0$, *then*

$$\sqrt{n}\left(\widehat{\beta}^{(1)} - \beta^{(1)0}\right) \xrightarrow{\mathcal{L}} \mathsf{N}_{d-1}(0, \Omega),$$

where $\Omega = A^{-1}BA^{-1}|_{\beta^{(1)}=\beta^{(1)0}}$.

Remark 1. When $\delta = 1$, the asymptotic co-variance matrix reduces to that of Cui *et al.* (2011).

To define a consistent estimator of the asymptotic variance, a natural way is first to define estimators of $h(t) = \mathsf{E}\{X|\beta^\top X = t\}$ using the local linear estimate as

$$\widehat{h}(t) = \sum_{i=1}^{n} b_i(t) X_i \Big/ \sum_{i=1}^{n} b_i(t),$$

where $b_i(t) = K_{b_n}\left(\widehat{\beta}^\top X_i - t\right)\left\{S_{n,2}(t) - \left(\widehat{\beta}^\top X_i - t\right) S_{n,1}(t)\right\}$ and $S_{n,k}(t) = K_{b_n}\left(\widehat{\beta}^\top X_i - t\right)\left(\widehat{\beta}^\top X_i - t\right)^k, k = 1, 2$. Let $\widehat{C}_i = (1-\delta_i) J^\top \widehat{h}\left(\widehat{\beta}^\top X_i\right) + J^\top\left(X_i - \widehat{h}\left(\widehat{\beta}^\top X_i\right)\right)\widehat{g}'\left\{\left(\widehat{\beta}^\top X_i\right)\right\}$. Then, the asymptotic variance Ω can be estimated by

$$\begin{aligned}\widehat{\Omega} = &\left[n^{-1}\textstyle\sum_{i=1}^n \rho_2\left\{\widehat{g}\left(\widehat{\beta}^\top X_i\right)\right\}\widehat{C}_i\widehat{C}_i^\top\right]^{-1} \\ &\times\left\{n^{-1}\textstyle\sum_{i=1}^n \delta_i q_1^2\left[\widehat{g}\left(\widehat{\beta}^\top X_i\right), Y_i\right]\widehat{C}_i\widehat{C}_i^\top\right\}\left[n^{-1}\textstyle\sum_{i=1}^n \rho_2\left\{\widehat{g}\left(\widehat{\beta}^\top X_i\right)\right\}\widehat{C}_i\widehat{C}_i^\top\right]^{-1}.\end{aligned}$$

Remark 2. If $\mu\left\{g\left(\beta^\top X\right)\right\} = g\left(\beta^\top X\right)$, $\sigma^2 V\left\{g\left(\beta^\top X\right)\right\} = \sigma^2$, then the matrix Ω in Theorem 3.1 reduces to

$$\begin{aligned}A^{-1}BA^{-1} = \mathsf{E}\Big[&\left\{(1-\delta)\, J^\top \mathsf{E}\left(X|\beta^\top X\right) + J^\top\left(X - \mathsf{E}\left(X^\top|\beta^\top X\right)\right)\left[g'\left(\beta^\top X\right)\right]\right\} \\ &\times\left\{(1-\delta) J^\top \mathsf{E}\left(X|\beta^\top X\right) + J^\top\left(X - \mathsf{E}\left(X^\top|\beta^\top X\right)\right)\left[g'\left(\beta^\top X\right)\right]\right\}^\top \sigma^2\Big].\end{aligned}$$

The asymptotic normality of $\widehat{\beta} = \left(\widehat{\beta}_1, \widehat{\beta}^{(1)\top}\right)^\top$ follows from Theorem 1 with a simple application of the multivariate delta-method, because $\widehat{\beta}_1 = \sqrt{1 - \|\widehat{\beta}^{(1)\top}\|}$.

Corollary 1. *Under the conditions of Theorem 1, we have*

$$\sqrt{n}\left(\widehat{\beta} - \beta^0\right) \xrightarrow{\mathcal{L}} \mathsf{N}_{d-1}\left(0, \Lambda_{\beta^0}\right),$$

where $\Lambda_{\beta^0} = J\Omega J^\top|_{\beta=\beta^0}$.

Using the plug in method, the asymptotic variance Λ_{β^0} can be estimated by $\widehat{J}\widehat{\Omega}\widehat{J}^\top$, where $\widehat{J}$ is J with β replaced by $\widehat{\beta}$.

Let $\hat{\beta}^{cc}$ be the CC estimator defined by the EFM method due to Cui *et al.* (2011). Similar to Cui *et al.* (2011), it can be shown that $\hat{\beta}^{cc}$ is asymptotically normal with mean zero and variance

$$\Lambda^{cc}_{\beta^0} = J\Omega^{cc}J^\top,$$

where $\Omega^{cc} = A_{cc}^{-1}$ with

$$A_{cc} = J^\top E\left[\delta\left(X - E\left\{\delta X|\beta^\top X\right\}\right)\left(X - E\left\{\delta X|\beta^\top X\right\}\right)^\top\right] \times\rho_2\left\{g\left(\beta^\top X\right)\right\}\left\{g'\left(\beta^\top X\right)\right\}^2/\sigma^2\Big] J.$$

Both the asymptotic variances of the proposed estimator and the CC estimator are of complex structure, and hence, it is hard to compare them in terms of their asymptotic variances. We will compare their finite sample properties in the following simulation section.

Theorem 2. *Suppose that conditions of Theorem 1 hold, we have*

$$\sqrt{nb_n}\left(\widehat{g}\left(\widehat{\beta}^\top x\right) - g\left(\beta^{0\top}x\right) - \frac{\mu^{(2)}\{g(\beta^\top x)\}}{2} e_1 S^{-1} U b_n^2\right) \xrightarrow{\mathcal{L}} \mathsf{N}(0, \Lambda_1),$$

where $U = (\mu_2, \mu_3)$, $e_1 = (1, 0)$ *and* $\Lambda_1 = \frac{\sigma^2}{\pi(x)\rho_2\{g(\beta^\top x)\} f_{\beta^\top x}(\beta^\top x)} e_1 S^{-1} S^* S^{-1}$.

Let $Z_i^* = \left(1, \frac{\widehat{\beta}^\top X_i - x}{b_n}\right)^\top$. The asymptotic variance Λ_1 can be estimated by

$$\begin{aligned}\widehat{\Lambda}_1 &= e_1\left[n^{-1}\textstyle\sum_{i=1}^n \delta_i q_2\left[\widehat{g}\left(\widehat{\beta}^\top X_i\right), Y_i\right] Z_i^* Z_i^{*\top} K_{b_n}\left(\widehat{\beta}^\top X_i - x\right)\right]^{-1} \\ &\quad\times n^{-1}\textstyle\sum_{i=1}^n \delta_i q_1^2\left[\widehat{g}\left(\widehat{\beta}^\top X_i\right), Y_i\right] Z_i^* Z_i^{*\top} K_{b_n}^2\left(\widehat{\beta}^\top X_i - x\right)\Big\} \\ &\quad\times\left[n^{-1}\textstyle\sum_{i=1}^n \delta_i q_2\left[\widehat{g}\left(\widehat{\beta}^\top X_i\right), Y_i\right] Z_i^* Z_i^{*\top} K_{b_n}\left(\widehat{\beta}^\top X_i - x\right)\right]^{-1}.\end{aligned}$$

where $q_1(z, y)$ and $q_2(z, y)$ are defined at the beginning of this section.

Remark 3. The choice of bandwidth is a very important topic in nonparametric regression estimation. The popular method such as cross-validation, generalized cross-validation and the rule of thumb can be used to select the optimal bandwidth for the estimator of $g(\cdot)$.

4. Simulation studies

We conducted some Monte Carlo simulation studies to evaluate the performances of the proposed estimators for finite samples. In our simulation, kernel functions $H(\cdot)$ and $K(\cdot)$ were taken as Gaussian kernel. The optimal bandwidths for b_n and h_n were chosen by using the cross-validation method.

Example 1. To illustrate how the initial estimate of β affects the estimate of β, we consider the following *simple* single-index model

$$Y = 6\left(X^\top \beta\right)^2 + \varepsilon, \tag{5}$$

where X is generated from $\mathsf{N}_d(2, I)$ for $d = 50$, $\varepsilon \sim \mathsf{N}(0, 0.2)$, the true parameter vector $\beta = \left(\sqrt{2}/2, \sqrt{2}/2, 0, \ldots, 0\right)$. Take the missing mechanism:

$$\text{logit}\{\mathsf{P}(\delta = 1|Y, X)\} = \gamma^\top X + c_0, \tag{6}$$

where $\text{logit}(a) = \log\{a/(1-a)\}$, $\gamma = \left(\sqrt{2}/4, \ldots, \sqrt{2}/4, 0, c_1\right)^\top / \sqrt{1 + c_1^2}$, c_0 is a constant to control missing proportion and c_1 is a constant to control the distance between γ and β. The number of replications is 500. The size of the sample was taken to be $n = 60$, 90 and 120, respectively.

The proposed estimator $\widehat{\beta}$ uses the fusion-refinement estimate as initial estimate in Step 1. Let $\hat{\beta}^{cc}$ be the CC estimator defined by the EFM method due to Cui *et al.* (2011). When the initial estimate is taken to be $(1, \ldots, 1)/\sqrt{d}$ or $\hat{\beta}^{cc}$, we define the resulting estimators to be $\widehat{\beta}^{(I)}$ or $\widehat{\beta}^{C}$. We compare $\widehat{\beta}$ with $\widehat{\beta}^{(I)}$ and $\hat{\beta}^{C}$, respectively, in terms of the average absolute bias (AB) and the square root of the trace of the standard covariance matrix (SRTSC). The AB is defined by

$$AB = \frac{1}{500}\sum_{i=1}^{500}\left(\frac{1}{d}\sum_{s=1}^{d}|\widehat{\beta}_{n,s}^{i} - \beta_s|\right),$$

and the SRTSC is defined by

$$SRTSC = \sqrt{\frac{1}{499}\sum_{i=1}^{500}\left\{\frac{1}{d}\left(\widehat{\beta}_n^i - \bar{\widehat{\beta}}\right)\left(\widehat{\beta}_n^i - \bar{\widehat{\beta}}\right)^\top\right\}},$$

where $\widehat{\beta}_{n,s}^{i}$ is the sth component of $\widehat{\beta}_n^i$ and $\widehat{\beta}_n^i$ is one of $\widehat{\beta}$, $\widehat{\beta}^{(I)}$ and $\hat{\beta}^{C}$ at the ith run and $\bar{\widehat{\beta}} = \frac{1}{500}\sum_{i=1}^{500}\widehat{\beta}_n^i$. The simulation results of AB and SRTSC for $\widehat{\beta}$, $\widehat{\beta}^{(I)}$ and $\widehat{\beta}^{(C)}$ with about 25%, 50% and 75% missing proportions reported in Table 1. From Table 1, we can see that the initial estimate does not affect the resulting estimators seriously. But, it also can be seen that the fusion-refinement initial estimate used in this paper is a relatively better choice in terms of AB and SRTSC of these estimators.

Example 2. To compare the proposed method with Wang *et al.* (2010), we first consider the following *simple* single-index model

$$Y = \left(X^\top \beta\right)^2 + \varepsilon, \tag{7}$$

Table 1. *AB and SRTSC of* $\widehat{\beta}$, $\widehat{\beta}^I$ *and* $\widehat{\beta}^C$ *with different MP and different sample sizes*

		AB		
n	MP	$\widehat{\beta}$	$\widehat{\beta}^I$	$\widehat{\beta}^C$
60	0.25	0.0628	0.0735	0.0788
	0.50	0.0947	0.1086	0.1123
	0.75	0.1184	0.1346	0.1352
90	0.25	0.0569	0.0620	0.0699
	0.50	0.0870	0.0917	0.0910
	0.75	0.1157	0.1226	0.1230
120	0.25	0.0523	0.0607	0.0587
	0.50	0.0768	0.0863	0.0833
	0.75	0.1070	0.1147	0.1112
		SRTSC		
n	p	$\widehat{\beta}$	$\widehat{\beta}^I$	$\widehat{\beta}^C$
60	0.25	0.0203	0.0247	0.0230
	0.50	0.0347	0.0446	0.0406
	0.75	0.0520	0.0617	0.0603
90	0.25	0.0137	0.0231	0.0169
	0.50	0.0279	0.0384	0.0318
	0.75	0.0471	0.0538	0.0505
120	0.25	0.0111	0.0197	0.0167
	0.50	0.0250	0.0346	0.0283
	0.75	0.0429	0.0483	0.0472

AB, absolute bias; MP, missing proportions; SRTSC, square root of the trace of the standard covariance.

where X is generated from $\mathsf{N}_d(2, I)$ for $d = 50$, $\varepsilon \sim \mathsf{N}(0, 0.2)$, the true parameter is $\beta = \left(2/\sqrt{5}, 1/\sqrt{5}, 0, \ldots, 0\right)$ and the missing mechanism follows model (6). The number of replications is 500. The size of the sample was taken to be $n = 60$, 90,120, respectively.

The proposed estimator $\widehat{\beta}$ is compared with $\widehat{\beta}^{wang}$ due to Wang *et al.* (2010) and the CC estimator $\widehat{\beta}^{cc}$ as mentioned in Example 1 and the full data estimator (denoted by $\widehat{\beta}^{full}$) due to Cui *et al.* (2011). The estimator $\widehat{\beta}^{full}$ can be served as a gold standard, although it can not be achievable in practice. We computed AB and SRTSC in Table 2 for $\widehat{\beta}$, $\widehat{\beta}^{cc}$, $\widehat{\beta}^{wang}$ and $\widehat{\beta}^{full}$ with about 25%, 50% and 75% missing proportions.

Several observations can be made from Table 2. Firstly, we can see that AB and SRTSC of all the estimators decrease as the sample size increases or the missing rate decreases, as expected. Secondly, we also see that $\widehat{\beta}$ outperforms $\widehat{\beta}^{wang}$ and perform better than $\widehat{\beta}^{cc}$ in terms of AB and SRTSC. It should be pointed out that the proposed method performs slightly better only than $\widehat{\beta}^{cc}$ when the sample size is large and the missing proportion is small. The reason may be that the covariables of the subjects with missing responses provide relatively more covariable information for the proposed method. Also, the proposed method is a three-step estimating approach with the fusion-refinement estimator used as the initial estimator. And the simulation results show that it performs better slightly than the three-step estimating approach with the CC estimator used as the initial estimator. Hence, it is possible that the resulting estimator performs better than both the initial estimators because the three-step estimating method uses more data and model information than the initial ones. When the missing proportion is small, the proposed estimator is comparable with the full data estimator $\widehat{\beta}^{full}$, the gold standard, and hence the proposed method performs well in terms of AB and SRTSC.

Table 2. *AB and SRTSC of* $\widehat{\beta}^{full}$, $\widehat{\beta}$, $\widehat{\beta}^{cc}$ *and* $\widehat{\beta}^{wang}$ *with different MP and different sample sizes*

			AB		
n	$\widehat{\beta}^{full}$	MP	$\widehat{\beta}$	$\widehat{\beta}^{cc}$	$\widehat{\beta}^{wang}$
60	0.0742	0.25	0.0790	0.0953	0.1034
		0.50	0.1009	0.1210	0.1297
		0.75	0.1214	0.1396	0.1388
90	0.0540	0.25	0.0584	0.0744	0.0773
		0.50	0.0869	0.0951	0.1229
		0.75	0.1087	0.1143	0.1348
120	0.0477	0.25	0.0462	0.0533	0.0578
		0.50	0.0679	0.0755	0.1038
		0.75	0.0980	0.1017	0.1100
			SRTSC		
60	0.0230	0.25	0.0243	0.0276	0.0259
		0.50	0.0305	0.0436	0.0445
		0.75	0.0572	0.0684	0.0659
90	0.0172	0.25	0.0187	0.0199	0.0200
		0.50	0.0293	0.0388	0.0342
		0.75	0.0471	0.0579	0.0586
120	0.0148	0.25	0.0156	0.0188	0.0179
		0.50	0.0227	0.0254	0.0307
		0.75	0.0411	0.0506	0.0511

AB, absolute bias; MP, missing proportions; SRTSC, square root of the trace of the standard covariance.

Example 3. In this study, we consider the following the ESIM:

$$\begin{aligned} &\mathsf{E}(Y|X) = \exp\left\{g\left(\beta^\top X\right)\right\}, \; g\left(\beta^\top X\right) = \sin\left(X^\top \beta\right) \\ &\mathsf{Var}(Y|X) = \sigma^2, \qquad\qquad \sigma = 0.2. \end{aligned} \tag{8}$$

The true parameter is $\beta = \left(2/\sqrt{5}, 1/\sqrt{5}, 0, \ldots, 0\right)$, X is generated from $\mathsf{N}_d(2, I)$ for $d = 50$, $\varepsilon \sim \mathsf{N}(0, 0.04)$ and the missing mechanism follows model (6). We calculated AB and SRTSC for $\widehat{\beta}$, $\widehat{\beta}^{full}$ and $\widehat{\beta}^{cc}$, where $\mu(\cdot) = \exp(\cdot)$ in (2). At the same time, AB and SRTSC for $\widehat{\beta}^{wang}$ were also computed where we treated model (8) as a simple single-index model. For each sample size of $n = 60$, 90 and 120, 500 replications were taken. The simulation results are summarized in Tables 3.

From Table 3, the similar observations to Example 2 can be found. This shows that the proposed method is attractive for the ESIM (8).

Example 4. To illustrate the adaptivity of our algorithm to heterogeneous errors, we consider model (9),

$$\begin{aligned} &\mathsf{E}(Y|X) = \left\{g\left(\beta^\top X\right)\right\}^2, g\left(\beta^\top X\right) = X^\top \beta \\ &\mathsf{Var}(Y|X) = \sigma^2 \exp\left\{\tfrac{\sqrt{5}}{7} g\left(\beta^\top X\right)\right\}, \qquad \sigma^2 = 1, \end{aligned} \tag{9}$$

where the true parameter is $\beta = \left(2/\sqrt{5}, 1/\sqrt{5}, 0, \ldots, 0\right)$, X is generated from $\mathsf{N}_d(2, I)$ for $d = 50$ and the missing mechanism follows model (6). We calculated AB and SRTSC for $\widehat{\beta}$, $\widehat{\beta}^{full}$ and $\widehat{\beta}^{cc}$. For each sample size of $n = 60$, 100, 200 and 300, 500 replications were calculated. The simulation results are also summarized in Table 4.

Table 3. *AB and SRTSC of* $\widehat{\beta}^{full}$, $\widehat{\beta}$, $\widehat{\beta}^{cc}$ *and* $\widehat{\beta}^{wang}$ *with different MP and different sample sizes*

			AB		
n	$\widehat{\beta}^{full}$	MP	$\widehat{\beta}$	$\widehat{\beta}^{cc}$	$\widehat{\beta}^{wang}$
60	0.0922	0.25	0.0958	0.1124	0.1243
		0.50	0.1119	0.1233	0.1341
		0.75	0.1298	0.1321	0.1398
90	0.0813	0.25	0.0858	0.1099	0.1184
		0.50	0.1051	0.1201	0.1233
		0.75	0.1260	0.1304	0.1380
120	0.0679	0.25	0.0744	0.0837	0.0943
		0.50	0.0889	0.0993	0.1109
		0.75	0.1094	0.1211	0.1236
			SRTSC		
60	0.0239	0.25	0.0259	0.0346	0.0396
		0.50	0.0397	0.0488	0.0503
		0.75	0.0608	0.0706	0.0718
90	0.0172	0.25	0.0184	0.0305	0.0327
		0.50	0.0322	0.0437	0.0459
		0.75	0.0513	0.0649	0.0665
120	0.0108	0.25	0.0116	0.0186	0.0254
		0.50	0.0230	0.0268	0.0396
		0.75	0.0463	0.0582	0.0599

AB, absolute bias; MP, missing proportions; SRTSC, square root of the trace of the standard covariance.

Table 4. *AB and SRTSC of* $\widehat{\beta}^{full}$, $\widehat{\beta}$ *and* $\widehat{\beta}^{cc}$ *with different MP and different sample sizes*

			AB	
n	$\widehat{\beta}^{full}$	MP	$\widehat{\beta}$	$\widehat{\beta}^{cc}$
60	0.1005	0.25	0.1053	0.1127
		0.50	0.1128	0.1247
		0.75	0.1293	0.1306
90	0.0954	0.25	0.0937	0.1055
		0.50	0.1056	0.1187
		0.75	0.1246	0.1269
120	0.0807	0.25	0.0803	0.0922
		0.50	0.0944	0.1050
		0.75	0.1185	0.1226
			SRTSC	
60	0.0648	0.25	0.0670	0.0783
		0.50	0.0846	0.0920
		0.75	0.1019	0.1123
90	0.0522	0.25	0.0567	0.0689
		0.50	0.0779	0.0834
		0.75	0.0936	0.1001
120	0.0467	0.25	0.0444	0.0570
		0.50	0.0623	0.0695
		0.75	0.0878	0.0939

AB, absolute bias; MP, missing proportions; SRTSC, square root of the trace of the standard covariance.

For the heteroscedastic setting, $\widehat{\beta}^{wang}$ cannot be calculated because it is for the simple single-index model. Hence, we compare $\widehat{\beta}$ with $\widehat{\beta}^{full}$ and $\widehat{\beta}^{cc}$ only. From Table 4, the similar observations to Example 2 can be found. Therefore, our estimation method also implements well for the heteroscedastic case.

5. Real data analysis

ACTG 175 data have been studied by some authors (e.g. Hammer *et al.*, 1996; Davidian *et al.*, 2005; Ding and Wang 2011; Hu *et al.*, 2010). In an HIV clinical trial, 2139 HIV positive patients were involved. The patients were randomized into four arms to receive monotherapy (zidovudine) or combined therapy (adefovir + didanosine, zidovudine + zalcitabine and didanosine). We apply the proposed methods to this data set. The response Y = I ("the CD4 count at 96ś5 weeks" $\geq$ 300). The predictors X are six baseline characteristics: age, weight, CD4 counts at baseline and 20 ± 5 weeks, CD8 counts at baseline and 20 ± 5 weeks. Let T denote the received therapy, that is, $T = 1$ if receiving combined therapy, and $T = 0$ otherwise. Among the 746 patients, there were 473 patients with observations in Y, including 105 patients receiving monotherapy and 368 patients receiving other therapies, and due to death and dropout, there were 273 patients with missing observations in Y, including 74 patients with $T = 0$ and 199 patients with $T = 1$. All the patients had predictors X observed.

The single-index model is used to model the relationship between the CD4 count at 96 ± 5 weeks and the relevant 6 predictors $X = (X_1, \ldots, X_6)^\top$:

$$\mathsf{P}\,(the\ CD4\ count\ at\ 96 \pm 5\ weeks \geq 300|X) = \exp\left\{g\left(\beta^\top X\right)\right\} / \left[1 + \exp\left\{g\left(\beta^\top X\right)\right\}\right], \tag{10}$$

where $\beta = (\beta_1, \ldots, \beta_6)^\top$. We first focused on the subset of data labelled by $T = 0$. We obtained the proposed estimator $\widehat{\beta} = (0.1289,\ 0.9195,\ 0.0161, 0.3546 \text{ and } -0.0677)^\top$. For the subset of data labelled by $T = 1$, we obtained $\widehat{\beta} = (0.1927,\ -0.9792,\ -0.0058,\ -0.0079,\ 0.0582 \text{ and } 0.0244)^\top$.

As one can see from two estimates, 'weight' has the larger positive influence when patients receive combined therapy. On the contrary, there is a negative influence when patients receive monotherapy for proposing method. 'Age' has the positive influence in the two setting; this is true because resistance become more and more weak with increasing age.

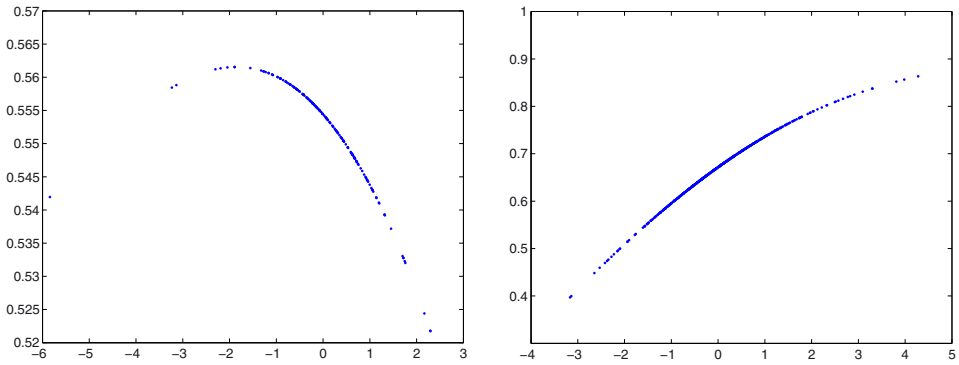


Fig. 1. Left: the estimated curve $\widehat{g}\left(\widehat{\beta}^\top X\right)$ against $\widehat{\beta}^\top X$ for the setting of $T = 0$. Right: the estimated curve $\widehat{g}\left(\widehat{\beta}^\top X\right)$ against $\widehat{\beta}^\top X$ for the setting of $T = 1$.

We also plot the scatter plot of the estimated single-index $\widehat{g}\left(\widehat{\beta}^{\top}X\right)$ against $\widehat{\beta}^{\top}X$ in the setting of $T = 0$ and $T = 1$, respectively. The scatter plot suggests a curvature relationship between the response and covariates. The pattern is displayed' in Fig. 1.

It is seen that there is a nonlinear trend. Therefore, using model (10) in the regression is perhaps more appropriate than using the internally linear model (11):

$$\mathsf{P}\left(the\ CD4\ count\ at\ 96 \pm 5\ weeks \geq 300|X\right) = \exp\left(\beta^{\top}X\right) / \left\{1 + \exp\left(\beta^{\top}X\right)\right\}. \quad (11)$$

Acknowledgement

Wang's research was supported by the National Science Fund for Distinguished Young Scholars in China (10725106), the National Natural Science Foundation of China (general programme 11171331 and key programme 11331011), a grant from the Key Lab of Random Complex Structure and Data Science, CAS and the Natural Science Foundation of SZU. Zhang's research was supported by the National Natural Science Foundation of China (11561006), research projects of colleges and universities in Guangxi (KY2015YB171) and innovation project of Guangxi Graduate Education (JGY2015122), a grant from the Key Base of Humanities and Social Sciences in Guangxi College. Härdle's research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 'Economic Risk'.

Supporting information

Additional information for this article is available online including detailed proofs of Theorems 1 and 2.

References

Chang, Z. Q., Xue, L. G. & Zhu, L. X. (2010). On an asymptotically more efficient estimation of the single-index model. *J. Multivariate Anal.* **101**, 1898–1901.

Cui, X., Härdle, W. & Zhu, L. x. (2011). The EFM approach for single-index models. *Ann. Statist.* **39**, 1658–1688.

Davidian. M., Tsiatis, A. A. & Leon, S. (2005). Semiparametric estimation of treatment effect in a pretest-posttest study with missing data. *Statistical Science* **20**, (3), 261-301.

Ding, X. & Wang, Q. H. (2011). Fusion-refinement procedure for dimension reduction with missing response at random. *J. Amer. Statist. Assoc.* **106**, 1193–1207.

Hall, P. (1989). On projection pursuit regression. *Ann. Statist* **17**, 573–588.

Hammer, S. M., *et al.* (1996). A Trial Comparing Nucleotide Monotherapy with Combined Therapy in HIV-Infected Adults With CD4 Cell Counts from 200 to 500 per Cubic Millimeter. *New England Journal of Medicine* **335**, 1081–1090.

Härdle, W., Hall, P. & Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist* **21**, 157–178.

Hristache, M., Juditsky, A. & Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *Ann. Statist.* **29**, 595–623.

Hu, Z. H., Follmann, D. A. & Qin, J. (2010). Semiparametric dimension reduction estimation for mean response with missing data. *Biometrika* **97**, 305–319.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics* **58**, 71–120.

Ichimura, H. (1987). *Estimation of single index models*, Ph.D. Dissertation, Dept. Economics, MIT.

Li, K. C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc* **86**, 316–342.

Little, R. J. A & Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, Wiley, NewYork.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc* **89**, 846–866.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Amer. Statist. Assoc.* **90**, 106–121.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.
Wang, Q. H. & Rao, J. N. K. (2002). Empirical likelihood-based inference under imputation for missing response data. *Ann. Statist.* **30**, 896–924.
Wang, Q. H., Lindon, O. & Härdle, W. (2004). Semiparametric regression analysis with missing response at random. *J. Amer. Statist. Assoc.* **99**, 334–345.
Wang, Y. H., Shen, J. S., He, S. Y. & Wang, Q. H. (2010). Estimation of single index model with missing response at random. *J. Statist. Plann. Inference* **140**, 1671–1690.
Xue, L. G. & Zhu, L. X. (2006). Empirical likelihood for single-index models. *J. Multivariate Anal.* **97**, 1295–1312.
Zhu, L. P., You, J. H. & Xu, Q. F. (2014). Statistical inference for single-index panel data models. *Scand. J. Stat.* **41**, 830–843.
Zhu, L. X. & Fang, K. T. (1992). On projection pursuit approximation for nonparametric regression. In *Proceedings of order statistics and nonparametrics: Theory and applications* (eds Sen, P. S. & Salama, I. A.), Hong Kong Baptist College, Hong Kong.
Zhu, L. X. & Fang, K. T. (1996). Asymptotics for kernel estimate of sliced inverse regression. *Ann. Statist.* **14**, 1053–1068.

Received December 2013, in final form March 2016

Qihua Wang, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.
E-mail: qhwang@amss.ac.cn

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Localizing Temperature Risk

Wolfgang Karl Härdle, Brenda López Cabrera, Ostap Okhrin & Weining Wang

To cite this article: Wolfgang Karl Härdle, Brenda López Cabrera, Ostap Okhrin & Weining Wang (2016) Localizing Temperature Risk, Journal of the American Statistical Association, 111:516, 1491-1508, DOI: 10.1080/01621459.2016.1180985

To link to this article: https://doi.org/10.1080/01621459.2016.1180985

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2016, VOL. 111, NO. 516, 1491–1508, Applications and Case Studies
http://dx.doi.org/10.1080/01621459.2016.1180985

Taylor & Francis
Taylor & Francis Group

Localizing Temperature Risk

Wolfgang Karl Härdle[a], Brenda López Cabrera[a], Ostap Okhrin[b], and Weining Wang[a]

[a]Ladislaus von Bortkiewicz Chair of Statistics of Humboldt-Universität zu Berlin, Berlin, Germany; [b]Faculty of Transportation, Dresden University of Technology, Dresden, Germany

ABSTRACT

On the temperature derivative market, modeling temperature volatility is an important issue for pricing and hedging. To apply the pricing tools of financial mathematics, one needs to isolate a Gaussian risk factor. A conventional model for temperature dynamics is a stochastic model with seasonality and intertemporal autocorrelation. Empirical work based on seasonality and autocorrelation correction reveals that the obtained residuals are heteroscedastic with a periodic pattern. The object of this research is to estimate this heteroscedastic function so that, after scale normalization, a pure standardized Gaussian variable appears. Earlier works investigated temperature risk in different locations and showed that neither parametric component functions nor a local linear smoother with constant smoothing parameter are flexible enough to generally describe the variance process well. Therefore, we consider a local adaptive modeling approach to find, at each time point, an optimal smoothing parameter to locally estimate the seasonality and volatility. Our approach provides a more flexible and accurate fitting procedure for localized temperature risk by achieving nearly normal risk factors. We also employ our model to forecast the temperaturein different cities and compare it to a model developed in 2005 by Campbell and Diebold. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received August 2014
Revised October 2015

KEYWORDS
Local model selection; Localizing temperature residuals; Seasonality; Weather derivatives

1. Introduction

The pricing of contingent claims based on stochastic dynamics, for example, stocks or FX rates, is well known in financial engineering. An elegant approach to such a pricing task is based on self-financing replication arguments. An essential element of this approach is the tradeability of the underlying. This, however, does not apply to weather derivatives, contingent on temperature or rain, since the underlying is not tradeable. In this context, the proposed pricing techniques are based on either equilibrium ideas (Horst and Mueller 2007) or econometric modeling of the underlying dynamics (Campbell and Diebold 2005; Benth, Benth, and Koekebakker 2007) followed by risk neutral pricing.

The equilibrium approach relies on assumptions about preferences (with explicitly known functional forms) though. In this study we prefer a phenomenological approach since the underlying (temperature) we consider is of a varying local nature and our analysis aims at understanding the pricing at different locations and different time points around the world. A time series approach has been taken by Benth, Benth, and Koekebakker (2007), who corrects for seasonality (in mean), then for intertemporal correlation and finally as in Campbell and Diebold (2005), for seasonal variations. After these manipulations, a Gaussian risk factor needs to be isolated to apply continuous time pricing techniques (Karatzas and Shreve 2001).

Empirical studies following this econometrical route show evidence that the resulting temperature risk factor deviates severely from Gaussianity, which in turn challenges the pricing tools (Benth, Härdle, and López Cabrera 2011). In particular, for Asian cities, like, for example, Kaohsiung (Taiwan), one observes very distinctive nonnormality in the form of clearly visible heavy tails caused by extended volatility in peak seasons. This is visible from Figure 1 where a log density plot reveals a nonnormal shoulder structure (kurtosis $= 3.22$, skewness $= -0.08$, JB $= 128.74$).

The econometric analysis we apply, follows Benth, Benth, and Koekebakker (2007) where temperature is decomposed into a seasonality term and a stochastic part with seasonal variance. The fitted seasonality trend and seasonal variance are approximated with truncated Fourier series (and an additional GARCH term).

The upper panel of Figure 2 displays the seasonality and deseasonalized residuals over two years in Kaohsiung. The lower panel RHS displays the empirical and smoothed seasonal variance function, while the lower panel LHS shows the smoothed seasonal variance function over years. The Fourier series expansion fails, though in the volatility peak seasons. Even incorporating an asymmetry term for the dip of temperature in winter does not improve the closeness to normality. One may of course pursue fine tuning the Fourier method with more and more periodic terms but this will increase the number of parameters; we, therefore, propose a local parametric approach. The mean and the seasonality function estimated with local linear regression using the quartic kernel are also shown in Figure 2. We observe

CONTACT Weining Wang *wangwein@cms.hu-berlin.de* Ladislaus von Bortkiewicz Chair of Statistics, Humboldt-Universität zu Berlin, Spandauer Straße 1, 10178 Berlin, Germany.
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/JASA*.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

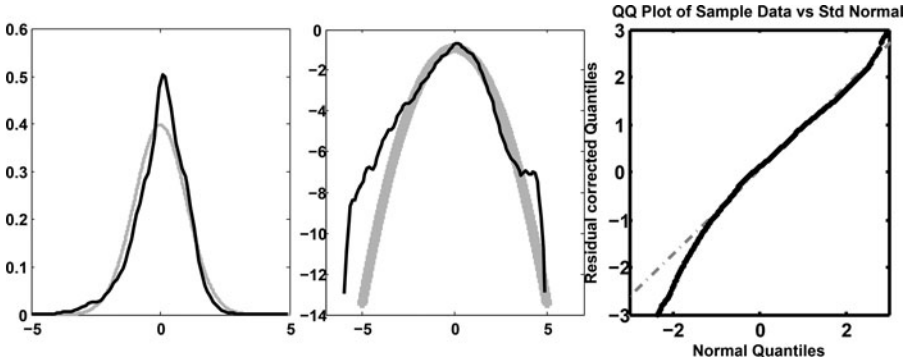


Figure 1. Kernel density estimates (left panel), log kernel density estimates (middle panel), and QQ-plots (right panel) of normal densities (gray lines) and Kaohsiung standardized residuals (black line).

high variance in winter and early summer and low variance in spring and late summer.

The scale correction of the obtained residuals (after seasonal and intertemporal fitting) is apparently not identical over a year. A very structured volatility pattern up to April is followed by a moderately constant period until an increasing peak starting in September. This motivates our research to localize temperature risk. The local smoothness of the seasonal variance function is of course not only a matter of one location (here Kaohsiung) but varies also over the different cities around the world that we are analyzing in this study. Our study is local in a double sense: local in time and space. We use adaptive methods to localise the underlying dynamics and with that being able to achieve Gaussian risk factors. This will justify the pricing via standard tools that are based on Gaussian risk drivers. The localization in time is based on adjusting the smoothing parameter. For a general framework on local parametric approximation we refer to Spokoiny (2009). As a result, we obtain better approximations to normality and, therefore, less biased prices.

This article is structured as follows. Section 2 describes the localizing approach. In Section 3, we present the data and conduct the analysis to different cities. Section 4 presents a forecasting exercise and Section 5 is devoted to an application where the pricing of weather derivative contract types is presented. Section 6 concludes the article. All quotations of currency in this article will be in USD unless otherwise stated and, therefore, we will omit the explicit notion of the currency. All the computations were carried out in Matlab version 7.6 and R. The

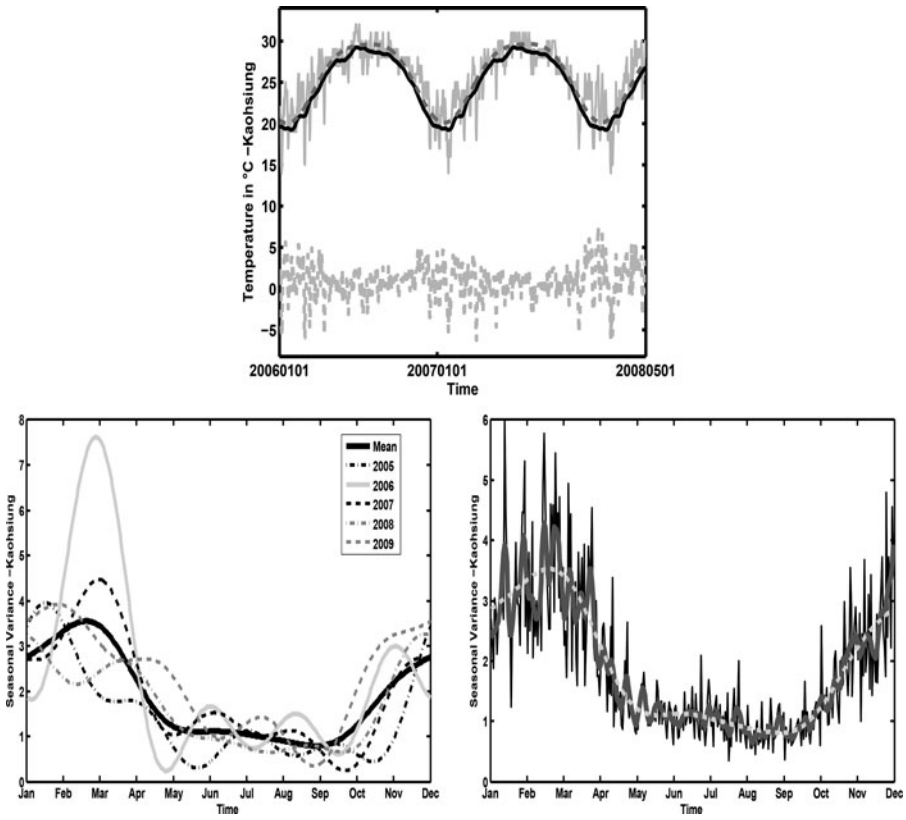


Figure 2. Upper panel: Kaohsiung daily average temperature (gray line), Fourier truncated (dotted gray line) and local linear seasonality function (black line), residuals in lower part. Lower left panel: Truncated Fourier seasonal variation, ($\hat{\sigma}_t^2$) over years. Lower right panel: Kaohsiung empirical (black line), truncated Fourier (dotted gray line), and local linear (gray line) seasonal variance ($\hat{\sigma}_t^2$) function.

Table 1. Seasonality estimates $\hat{\Lambda}_t$ of daily average temperature.

City	Period	$\hat{a}$	$\hat{b}$	$\hat{c}_1$	$\hat{d}_1$	$\hat{c}_2$	$\hat{d}_2$	$\hat{c}_3$	$\hat{d}_3$
Berlin	(19480101–20080527)	9.2173	0.0000	9.8932	−157.9123	0.2247	261.2850	0.1591	−127.7303
	(19730101–20080527)	9.3050	0.0001	10.0070	−161.2493	0.4601	−66.0530	−0.3723	−416.4776
	(19730101–20080527)	9.3050	0.0001	10.0070	−161.2493	0.4601	−66.0530	−0.3723	−416.4776
	(19830101–20080527)	9.4581	0.0001	10.0969	−161.7129	0.5205	−51.9929	0.3734	42.0874
	(19930101–20080527)	9.5923	0.0002	10.1995	−162.9774	0.6564	−37.1548	0.4241	41.9970
	(20030101–20080527)	9.6948	0.0007	10.1954	−162.3343	0.5554	−43.2293	0.3269	1.5998
Kaohsiung	(19730101–20081231)	24.2289	0.0001	0.9157	−145.6337	−4.0603	−78.1426	−1.0505	10.6041
	(19730101–19821231)	24.4413	0.0001	2.1112	−129.1218	−3.3887	−91.1782	−0.8733	20.0342
	(19830101–19921231)	25.0616	0.0003	2.0181	−135.0527	−2.8400	−89.3952	−1.0128	20.4010
	(19930101–20021231)	25.3227	0.0003	3.9154	−165.7407	−0.7405	−51.4230	−1.1056	19.7340
New York	(19490101–20081204)	53.1473	0.0001	18.6810	−143.4051	−3.3872	271.5072	−0.4203	−16.3125
	(19730101–20081204)	53.6992	0.0001	18.0092	−148.4124	−3.5236	279.6876	−0.4756	−21.8090
	(19730101–19821204)	53.6037	−0.0000	17.7446	−155.2453	−3.7769	289.7932	−0.8326	−4.2257
	(19830101–19921204)	54.8740	−0.0003	17.6924	−152.7461	−3.4245	284.6412	−0.4933	−218.9204
	(19930101–20021204)	53.8050	0.0003	17.6942	−153.3997	−3.4246	285.7958	0.5753	−315.2792
	(20030101–20081204)	52.9177	0.0012	17.8425	−151.2977	−3.8837	287.2022	−0.1290	−216.7298
Tokyo	(19730101–20081231)	15.7415	0.0001	8.9171	−162.3055	−2.5521	−7.8982	−0.7155	−15.0956
	(19730101–19821231)	15.8109	0.0001	9.2855	−162.6268	−1.9157	−16.4305	−0.5907	−13.4789
	(19830101–19921231)	15.4391	0.0004	9.4022	−162.5191	−2.0254	−4.8526	−0.8139	−19.4540
	(19930101–20021231)	16.4284	0.0001	8.8176	−162.2136	−2.1893	−17.7745	−0.7846	−22.2583
	(20030101–20081231)	16.4567	0.0001	8.5504	−162.0298	−2.3157	−18.3324	−0.6843	−16.5381

NOTE: All coefficients are nonzero at 1% significance level.

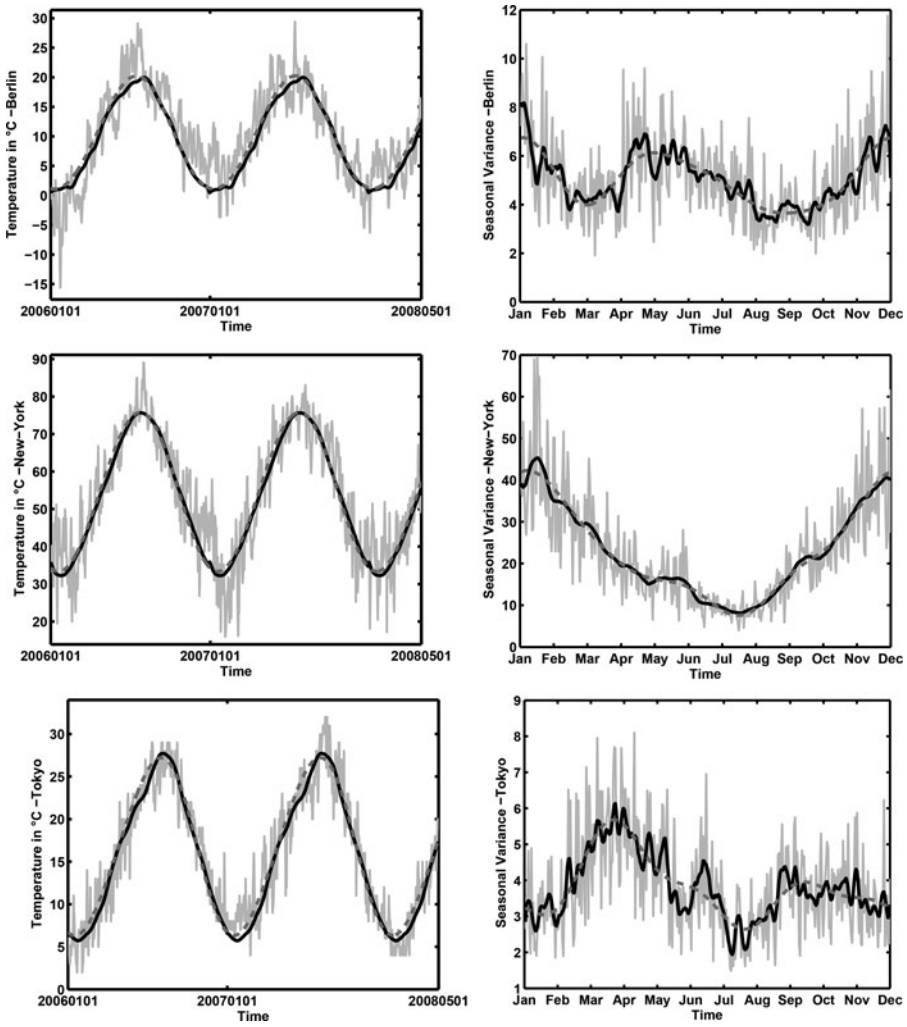


Figure 3. The empirical (gray line), the Fourier truncated (dotted gray line), and the local linear (black line) seasonal mean (left panel) and variance component (right panel) using quartic kernel and bandwidth $h = 4.49$.

Table 2. ADF and KPSS-Statistics for the detrended daily average temperature time series for different cities.

City	Period	ADF	KPSS
Atlanta	19480101–20081204	−55.55+	0.21***
Beijing	19730101–20090831	−30.75+	0.16***
Berlin	19480101–20080527	−40.94+	0.13**
Essen	19700101–20090731	−23.87+	0.11*
Houston	19700101–20081204	−38.17+	0.05*
Kaohsiung	19730101–20091210	−37.96+	0.05*
New York	19490101–20081204	−56.88+	0.08*
Osaka	19730101–20090604	−18.65+	0.09*
Portland	19480101–20081204	−45.13+	0.05*
Taipei	19920101–20090806	−32.82+	0.09*
Tokyo	19730101–20090831	−25.93+	0.06*

NOTE: '+' corresponds to a significance level of 0.01 for ADF test, and '*', '**' and '***' corresponds to significance levels of 0.1, 0.05 and 0.01, respectively, for KPSS test.

temperature data for different cities in the U.S., Europe, and Asia were obtained from the National Climatic Data Center (NCDC), the Deutscher Wetterdienst (DWD), Bloomberg Professional Service, and the Japanese Meteorological Agency (JMA). All data is converted to Celsius degrees. Weather derivative data from CME was extracted from Bloomberg. To simplify notation, dates are denoted using a yyyymmdd format.

2. Model

Although the temperature data are usually given in a discrete scale, temperature itself develops continuously over time. Thus, a continuous model for the futures price dynamics can be clearly formulated. We propose, as also suggested in Benth, Benth, and Koekebakker (2007) and Härdle and López Cabrera (2012), a mean reverted Ornstein-Uhlenbeck process for the modeling of detrended temperature variations in continuous time CAR(L):

$$d\mathbf{X}_t = \mathbf{A}\mathbf{X}_t dt + \mathbf{e}_L \sigma_t dB_t, \tag{1}$$

where $\sigma_t^2 > 0$ is a bounded deterministic seasonal variation, $\mathbf{X}_t \in \mathbb{R}^L$(detrended temperature) for $L \geq 1$ denotes a vectorial Ornstein-Uhlenbeck process, $\mathbf{e}_k$ a kth unit vector in $\mathbb{R}^L$ for $k = 1, \ldots, L$, B_t a Brownian motion, and an $L \times L$-matrix $\mathbf{A}$:

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & & \ddots & & 0 & \vdots \\ 0 & \cdots & \cdots & 0 & 0 & 1 \\ -\alpha_L & -\alpha_{L-1} & \cdots & & -\alpha_2 & -\alpha_1 \end{pmatrix}.$$

To bring the continuous time model in (1) to data, we consider a discretized version of it. The details of the discretization can be found in the Appendix. Let us first refine our notation from t to (t, j), with $t = 1, \ldots, \tau = 365$ days, $j = 0, \ldots, J$ years. The discrete time series model for calibration is given as:

$$\begin{aligned} X_{365j+t} &= T_{t,j} - \Lambda_t, \\ X_{365j+t} &= \sum_{l=1}^{L} \beta_{lj} X_{365j+t-l} + \varepsilon_{t,j}, \\ \varepsilon_{t,j} &= \sigma_t e_{t,j}, \\ e_{t,j} &\sim \mathrm{N}(0,1), \\ \hat{\varepsilon}_{t,j} &= X_{365j+t} - \sum_{l=1}^{L} \hat{\beta}_{lj} X_{365j+t-l}, \end{aligned} \tag{2}$$

where $T_{t,j}$ is the temperature at day t in year j, Λ_t denotes the seasonality effect and σ_t the seasonal variance. We adopt the model in (2) and estimate Λ_t, σ_t nonparametricly using adaptive methods proposed later in Section 2.1. Motivation for using this model can be found in Campbell and Diebold (2005) (CD), who proposes the model, see their Equations (1), (1a), (1b), (1c).

$$\begin{aligned} T_t &= \mathrm{Trend}_t + \mathrm{Seasonality}_t + \sum_{l=1}^{L} \rho_{t-l} T_{t-l} + \sigma_t \varepsilon_t, \\ \mathrm{Trend}_t &= \sum_{m=0}^{M} \beta_m t^m, \\ \mathrm{Seasonality}_t &= \sum_{p=1}^{P} \left[\delta_{c,p} \cos\left\{ 2\pi p \frac{d(t)}{365} \right\} + \delta_{s,p} \sin\left\{ 2\pi p \frac{d(t)}{365} \right\} \right], \end{aligned}$$

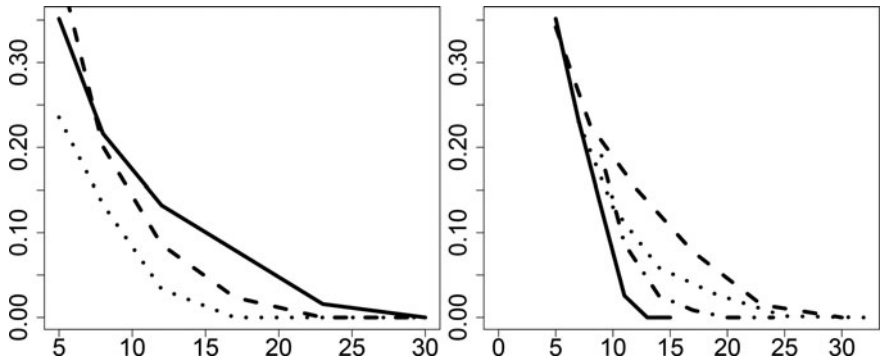
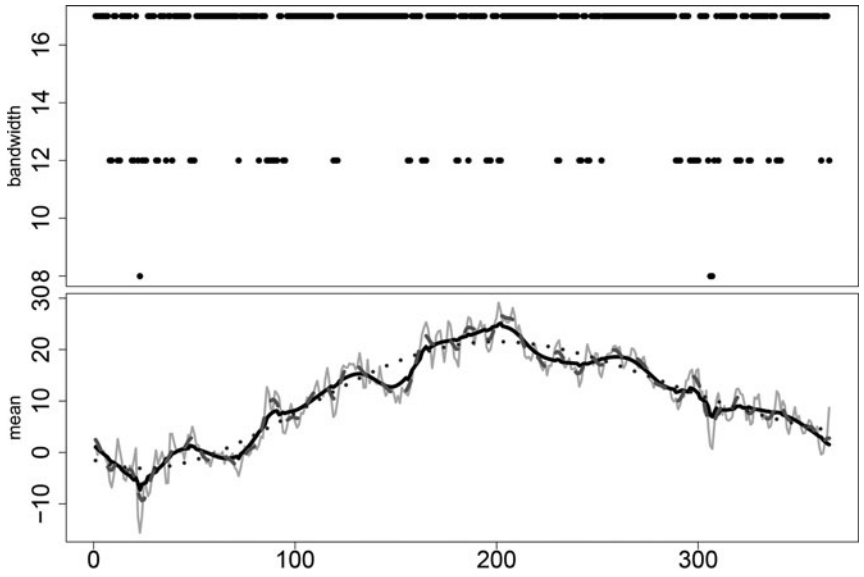
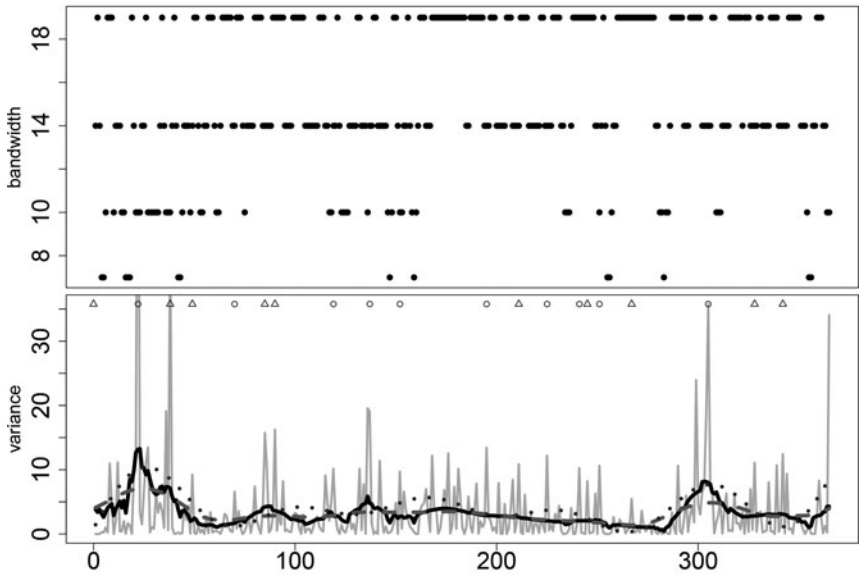


Figure 4. Simulated Critical Values for likelihood of seasonal variance (9) with $\theta^* = 1$, $r = 0.5$, number of simulation runs = 10,000 with $\alpha = 0.3$ (dotted), 0.5 (dashed), 0.7 (solid) for the bandwidth sequence (3, 5, 8, 12, 17, 23, 30) on the left plot and with $\alpha = 0.3$ and for sequences (3, 5, 7, 9, 11, 13, 15) (solid), (3, 5, 8, 12, 17, 23, 30) (dashed), (5, 7, 10, 14, 19, 25, 32) (dotted), and (7, 9, 11, 14, 17, 10, 24) (dot-dashed) on the right plot.



(a) Mean, 2007



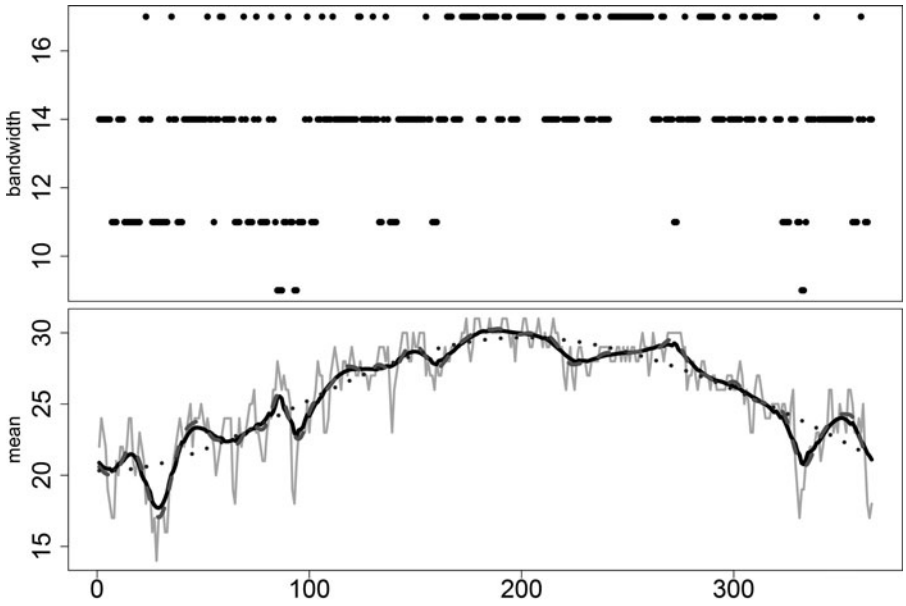
(b) Variance, 2007

Figure 5. Estimation of mean and variance for Berlin. In both figure sequence of bandwidths (upper panel), averaged observations (solid gray line), nonparametric function estimation with fixed bandwidth (dashed gray line), adaptive bandwidth (solid black line) and truncated Fourier (dotted line) (bottom panel of each figure). Circles and triangles in each bottom panel for variance represents the 10 smallest and the 10 largest outliers respectively.

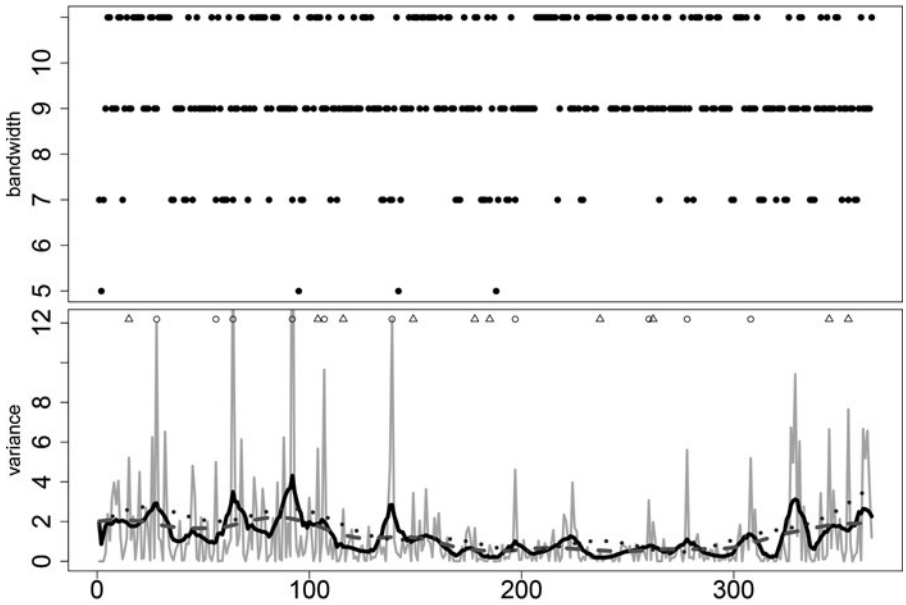
$$\sigma_t^2 = \sum_{q=1}^{Q}\left[\gamma_{c,q}\cos\left\{2\pi q\frac{d(t)}{365}\right\} + \gamma_{s,q}\sin\left\{2\pi q\frac{d(t)}{365}\right\}\right] + \sum_{r=1}^{R}\left\{\alpha_r(\sigma_{t-r}\varepsilon_{t-r})^2 + \sum_{s=1}^{S}\beta_s\sigma_{t-s}^2\right\}.$$

In all the comparisons below, we follow the setting proposed by Campbell and Diebold (2005) with $L = 25, M = 1, P = 3, Q = 3, R = 1$, and $S = 1$. The CD model is also based on a seasonal autoregressive process, but it is quite different from our model in (2). Instead of regressing the deseasonalized temperature on the lagged deseasonalized temperature as in (2), CD model regresses the present's deseasonalized temperature on the temperature in previous days. The trend function thus cannot be interpreted as "seasonal function" but a seasonal component. Also CD model suggests an additive structure instead of a multiplicative one for the seasonality and GARCH effect in the temperature volatility. Please refer to Benth and Benth (2012) for a detailed discussion of the differences between those two models.

We will use the CD model as a benchmark model for further analysis. Later studies, for example, Benth, Benth, and Koekebakker (2007) and Härdle and López Cabrera (2012), have provided evidence that the parameters β_{lj} are likely to be j independent and hence estimated consistently from a global autoregressive process model AR(L_j) with $L_j = L$. Also, Benth, Benth, and Koekebakker (2007) adopt the parameterization of



(a) Mean, 2008



(b) Variance, 2008

Figure 6. Estimation of mean and variance for Kaohsiung. In both figure sequence of bandwidths (upper panel), averaged observations (solid gray line), nonparametric function estimation with fixed bandwidth (dashed gray line), adaptive bandwidth (solid black line) and truncated Fourier (dotted line) (bottom panel of each figure). Circles and triangles in each bottom panel for variance represents the 10 smallest and the 10 largest outliers respectively.

Λ_t and σ_t as follows:

$$\Lambda_t = a + bt + \sum_{l=1}^{L_1} c_l \cos\left\{\frac{2\pi(t-d_l)}{l \cdot 365}\right\}, \tag{3}$$

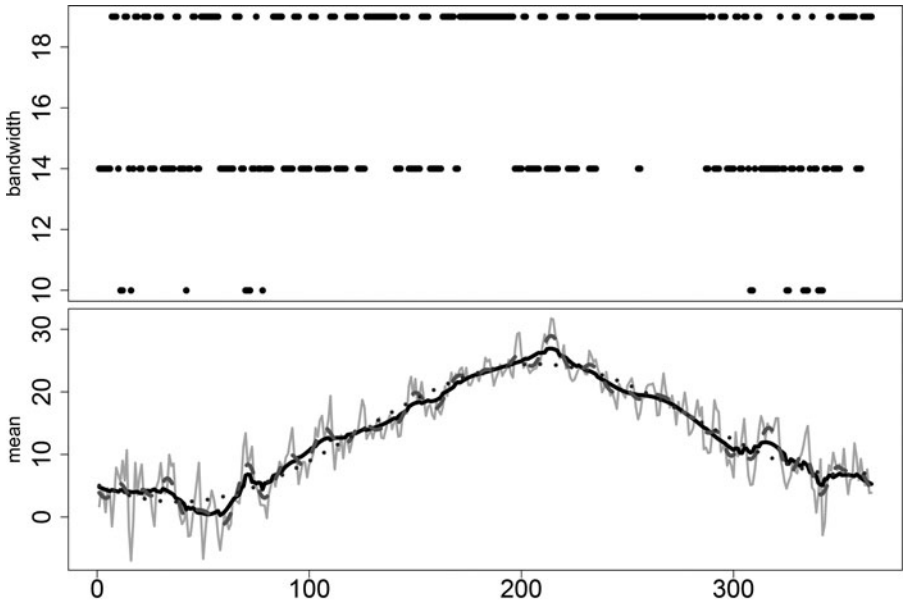
$$\sigma^2_{t,\mathrm{FTSG}} = c_{10} + \sum_{l=1}^{L_2}\left\{c_{2l}\cos\left(\frac{2l\pi t}{365}\right) + c_{2l+1}\sin\left(\frac{2l\pi t}{365}\right)\right\} + \alpha_1(\sigma_{t-1}\eta_{t-1})^2 + \beta_1\sigma^2_{t-1},$$
$$\eta_t \sim iid(0,1). \tag{4}$$

An alternative path to model Λ_t and σ_t is to use a nonparametric method: the local linear regression, where the seasonality Λ_s and σ_s are approximated with a local linear regression (LLR) estimator:

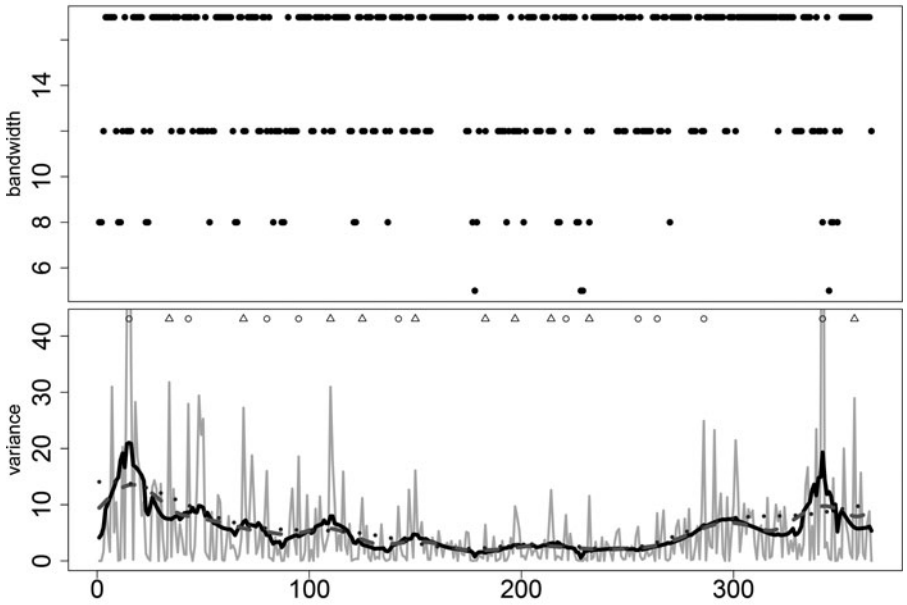
$$\underset{e,f}{\arg\min} \sum_{t=1}^{365}\left\{\bar{T}_t - e_s - f_s(t-s)\right\}^2 K\left(\frac{t-s}{h}\right), \tag{5}$$

$$\underset{g,v}{\arg\min} \sum_{t=1}^{365}\left\{\hat{\varepsilon}^2_t - g_s - v_s(t-s)\right\}^2 K\left(\frac{t-s}{h}\right), \tag{6}$$

where $\bar{T}_t$ is the mean (over years) of daily averages temperatures, $\hat{\varepsilon}^2_t$ the squared residual process (after seasonal and intertemporal fitting), h the bandwidth and $K(\cdot)$ is a kernel. Note, that due to the spherical character of the data, the kernel weights in (5)



(a) Mean, 2007



(b) Variance, 2007

Figure 7. Estimation of mean and variance for New York. In both figure sequence of bandwidths (upper panel), averaged observations (solid gray line), nonparametric function estimation with fixed bandwidth (dashed gray line), adaptive bandwidth (solid black line) and truncated Fourier (dotted line) (bottom panel of each figure). Circles and triangles in each bottom panel for variance represents the 10 smallest and the 10 largest outliers, respectively.

and (6) may be calculated from "wrapped around observations" thereby avoiding boundary bias. The estimates $\hat{\Lambda}_s$, $\hat{\sigma}_s^2$ are given by the minimisers $\hat{e}_s$, $\hat{g}_s$ of (5) and (6).

The seasonal trend function Λ_t and the seasonal variance function σ_t^2 affect, of course, the Gaussianity of the resulting normalized residuals. The commonly used approaches 1. truncated Fourier series, and 2. local polynomial regression (with fixed bandwidth) are rather restrictive and do not fit the data well since they do not necessarily yield normal risk factors. These observations motivated us to consider a more flexible approach. The main idea is to fit a local parametric model for the trend and variance with adaptively chosen window sizes. Specifically, we use kernel smoothing and employ an adaptive technique to choose the bandwidth over days. Other examples of this technique can be found in Cízek, Härdle, and Spokoiny (2009) and Chen, Härdle, and Pigorsch (2010).

It is worth noting that when we bring our model to the data, one can choose to estimate the mean function year by year as $\hat{\Lambda}_{t,j}$ or take the average over years as $\hat{\Lambda}_t$, this is later referred as the separately estimated mean and the jointly estimated mean methods, respectively. Regarding the estimate $\hat{\sigma}_t$, an aggregated approach is developed to tackle the problem of losing information when considering estimates at the individual level or averaging mean (variance) functions over time. This approach considers the minimum variance between the

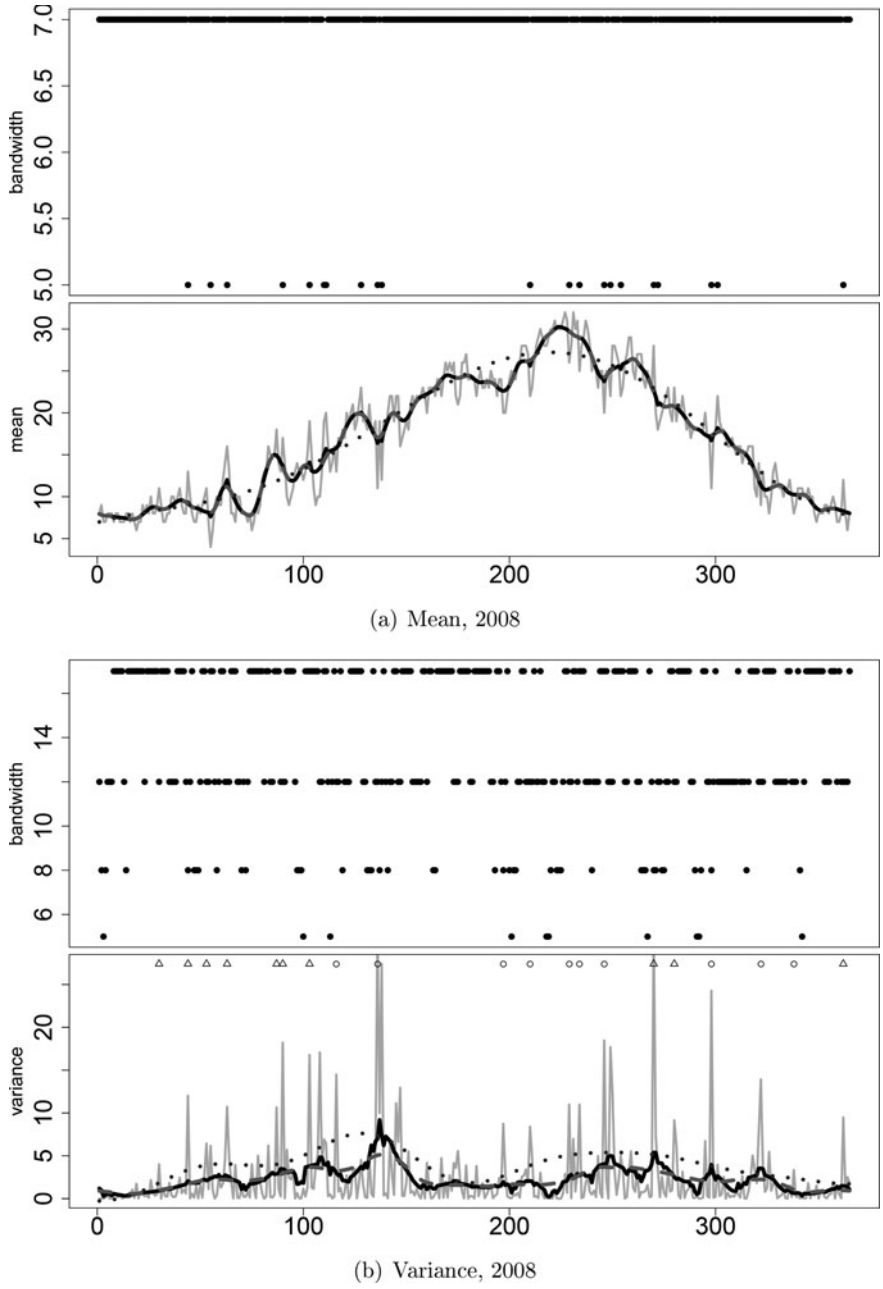


Figure 8. Estimation of mean and variance for Tokyo. In both figure sequence of bandwidths (upper panel), averaged observations (solid gray line), nonparametric function estimation with fixed bandwidth (dashed gray line), adaptive bandwidth (solid black line) and truncated Fourier (dotted line) (bottom panel of each figure). Circles and triangles in each bottom panel for variance represents the 10 smallest and the 10 largest outliers respectively.

aggregation of yearly local function estimates and an optimal local estimate θ^o. Once the sets of local functions have been identified, the aggregated local function can be defined as the weighted average of all the observations in a given time set. Formally, if $\hat{\theta}^j(t)$ is the localized estimation of the variance function σ^2 at time t of year j, the aggregated local function is given by:

$$\hat{\theta}_\omega(t) = \sum_{j=1}^{J} \omega_j \hat{\theta}^j(t). \tag{7}$$

With this aggregation step across J, we give the same weight to all observations, even to observations that were unimportant at the yearly level. Then a reasonable optimised estimate will be:

$$\arg\min_{\omega} \sum_{j=1}^{J} \sum_{t=1}^{365} \{\hat{\theta}_\omega(t) - \hat{\theta}_j^o(t)\}^2$$
$$\text{subject to } \Sigma_{j=1}^{J} \omega_j = 1;\ \omega_j > 0,\ j = 1, \ldots, J, \tag{8}$$

where the weights are assumed to be exogenous and nonstochastic, and $\hat{\theta}_j^o$ is defined as one of the following: 1 (Locave), $\hat{\theta}_j^o(t) = J^{-1} \sum_{j=1}^{J} \hat{\sigma}_j^2(t)$, the average of seasonal empirical variances over years, 2, (Locsep) $\hat{\theta}_j^o(t) = \hat{\sigma}_j^2(t)$, the yearly empirical variances, 3, one of above two approaches with maximised p-values over a year. One may interpret this normalization of weights as an optimization with respect to different frequencies (yearly, daily). In

Table 3. Summary of methods.

Method	Explanation
JoMe adMe adVa	Jointly estimated mean, adaptive bandwidth mean adaptive bandwidth variance
JoMe fiMe fiVa	Jointly estimated mean, fixed bandwidth mean fixed bandwidth variance
SeMe adMe adVa	Separately estimated mean, adaptive bandwidth mean adaptive bandwidth variance
SeMe fiMe fiVa	Separately estimated mean, fixed bandwidth mean fixed bandwidth variance
Locave	Aggregated approach with average of yearly empirical variance as the target
Locsep	Aggregated approach with each year's empirical variance as the target
Locmax	The optimal between Locave and Locsep (minimize the p-value)
Fourier	Method with Fourier series fitting for mean and variance
CD	Method adopted by Campbell and Diebold (2005)

the next subsection we describe the localization procedures for Λ_t and σ_t, which are going to be elements of estimation methods applied to the temperature data (our summary of the final estimation methods can be found in Table 3).

2.1. Adaptive Estimation

In this section we introduce adaptive procedures adopted for flexible estimation of Λ_t and σ_t. The time series $T_{t,j}$ are approximated at a fixed time point $s \in [1, 365]$. Our goal is to find a local window that possesses certain optimality properties, to be defined below. Specifically, for a specified weight sequence, we conduct a sequential likelihood ratio test (LRT) to choose an appropriate bandwidth. Different procedures of estimating seasonality and volatility are studied. Suppose that the object to be approximated is the seasonal variance $\theta(t) = \{\sigma_t^2\}$ (Λ_t can be estimated similarly). A weighted maximum likelihood approach is given by:

$$\begin{aligned}\tilde{\theta}_k(s) &\stackrel{\text{def}}{=} \arg\max_{\theta} L\{W^k(s), \theta\} \\ &= \arg\min_{\theta} \sum_{t=1}^{365}\sum_{j=0}^{J}\{\log(2\pi\theta)/2 \\ &\quad + \hat{\varepsilon}_{t,j}^2/2\theta\}w(s,t,h_k), \end{aligned} \tag{9}$$

with the "localizing scheme" $W^k(s) = \{w(s,1,h_k), w(s,2,h_k), \ldots, w(s,365,h_k)\}^\top$, where $w(s,t,h_k) = h_k^{-1}K\{(s-t)/h_k\}$, $k = 1, \ldots, K$, $h_1 < h_2 < h_3 < \cdots < h_K$ a prescribed sequence of bandwidths, and $K(u) = 15/16(1-u^2)^2\mathbf{I}(|u| \le 1)$ (quartic kernel).

Define confidence sets with critical values (Critical Values) $\mathfrak{z}_k$ to level α:

$$\mathfrak{E}_{\alpha,k} = \{\theta : L(W^k, \tilde{\theta}_k, \theta) \le \mathfrak{z}_k\}, \tag{10}$$

where the likelihood ratio is defined as

$$L(W^\ell, \tilde{\theta}_k, \theta) \stackrel{\text{def}}{=} L(W^\ell, \tilde{\theta}_k) - L(W^\ell, \theta). \tag{11}$$

Equipped with confidence sets (10), we launch the local model selection (LMS) algorithm:

Step 1. Fix a point $s \in \{1, 2, \ldots, 365\}$.

Step 2. Start with the smallest interval h_1: $\hat{\theta}_1 = \tilde{\theta}_1$

Step 3. For $k \ge 2$, $\tilde{\theta}_k$ is accepted and $\hat{\theta}_k = \tilde{\theta}_k$ if $\tilde{\theta}_{k-1}$ was accepted and $\tilde{\theta}_k \in \mathfrak{E}_{\alpha,l}$, $\forall \ell = 1, \ldots, k-1$, that is,

$$L(W^k, \tilde{\theta}_\ell, \tilde{\theta}_k) \le \mathfrak{z}_\ell, \forall \ell = 1, \ldots, k-1.$$

Otherwise, set $\hat{\theta}_k = \hat{\theta}_{k-1}$, where $\hat{\theta}_k$ is the latest accepted after first k steps.

Step 4. Define $\hat{k}$ as the kth step we stopped, and $\hat{\theta}_\ell = \tilde{\theta}_{\hat{k}}$, $\ell \ge k$.

The critical values $\mathfrak{z}_\ell$ used in the sequential test above are computed based on the following algorithm:.

Step 1. Consider first $\mathfrak{z}_1$ and let $\mathfrak{z}_2 = \mathfrak{z}_3 = \cdots = \mathfrak{z}_{K-1} = \infty$. This leads to the estimates $\hat{\theta}_k(\mathfrak{z}_1)$ and the value $\mathfrak{z}_1$ is selected as the minimal one for which

$$\sup_{\theta^*} \mathrm{E}_{\theta^*}|L\{W^k, \tilde{\theta}_k, \hat{\theta}_k(\mathfrak{z}_1)\}|^r \le \frac{\alpha \mathfrak{r}_r}{K-1}, k = 2, \ldots, K. \tag{12}$$

Step 2. Suppose $\mathfrak{z}_1, \ldots, \mathfrak{z}_{k-1}$ have been fixed, and set $\mathfrak{z}_k = \cdots = \mathfrak{z}_{K-1} = \infty$. With estimate $\hat{\theta}_m(\mathfrak{z}_1, \ldots, \mathfrak{z}_k)$ for $m = k+1, \ldots, K$. select $\mathfrak{z}_k$ as the minimal value which fulfills

$$\sup_{\theta^*} \mathrm{E}_{\theta^*}|L\{W^m, \tilde{\theta}_m, \hat{\theta}_m(\mathfrak{z}_1, \ldots, \mathfrak{z}_k)\}|^r \le \frac{k\alpha \mathfrak{r}_r}{K-1} \tag{13}$$

for $m = k+1, \ldots, K$.

Inequality (12) describes the impact of the k Critical Value to the risk, while the factor $\frac{k\alpha}{K-1}$ in (13) ensures that every $\mathfrak{z}_k$ has the same impact. The values of $(\alpha, r, h_1, \ldots, h_K)$ are prespecified hyper parameters for which robustness and sensitivity issues will be discussed in Section 3.

To be more specific, the explicit solution of (9) is in fact a Nadaraya-Watson estimator:

$$\begin{aligned}\tilde{\theta}_k(s) &= \sum_{t,j} \hat{\varepsilon}_{t,j}^2 w(s,t,h_k) \Big/ \sum_{t,j} w(s,t,h_k) \\ &= \sum_{t} \hat{\varepsilon}_{t}^2 w(s,t,h_k) \Big/ \sum_{t} w(s,t,h_k),\end{aligned}$$

with

$$\hat{\varepsilon}_t^2 \stackrel{\text{def}}{=} (J+1)^{-1} \sum_{j=0}^{J} \hat{\varepsilon}_{t,j}^2.$$

From a smoothing perspective we are in a comfortable situation here since the boundary bias is not an issue, as we are dealing with a periodic function $\theta(t) = \theta(t+365)$. We use mirrored observations: assume $h_K < 365/2$, then the observation set, for example for the seasonal variance, is extended to $\hat{\varepsilon}_{-364}^2, \hat{\varepsilon}_{-363}^2, \ldots, \hat{\varepsilon}_0^2, \hat{\varepsilon}_1^2, \ldots, \hat{\varepsilon}_{730}^2$, where

$$\hat{\varepsilon}_t^2 \stackrel{\text{def}}{=} \hat{\varepsilon}_{365+t}^2, -364 \le t \le 0,$$

$$\hat{\varepsilon}_t^2 \stackrel{\text{def}}{=} \hat{\varepsilon}_{t-365}^2, 366 \le t \le 730.$$

Since the location s is fixed, we drop s for simplicity of notation.

The theoretical background for the adaptation procedure can be found in the Appendix.

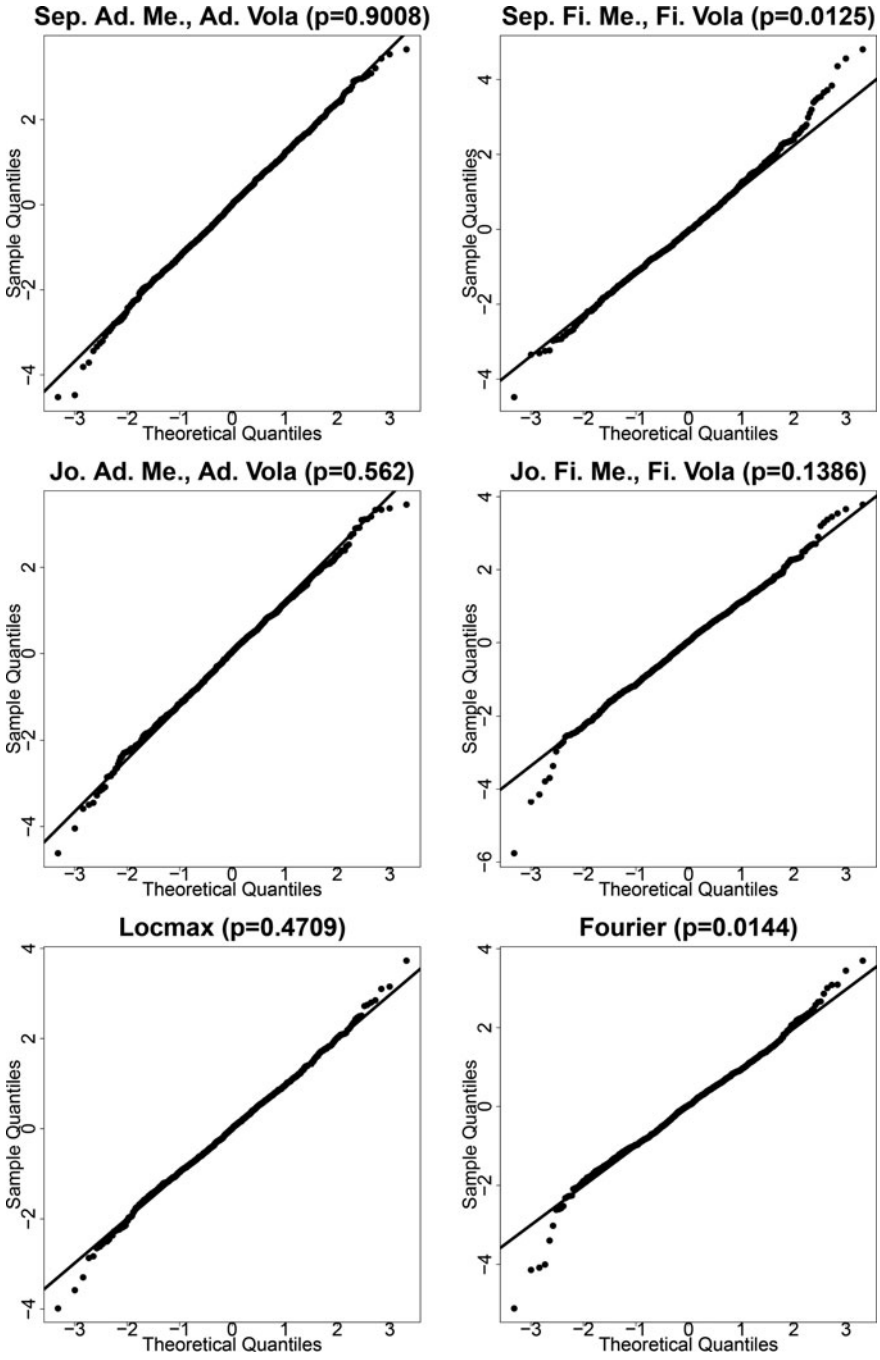


Figure 9. QQ-plot for standardized residuals from Berlin using different methods for the data from 2005–2007 (3 years). Please see Table 3 for a summary of methods.

3. Empirical Analysis

We conduct an empirical analysis of temperature patterns for different cities. The main dataset contains the daily average temperatures for different cities in Europe, Asia, and the U.S. for the period 1900–2011: Atlanta, Beijing, Berlin, Essen, Houston, Kaohsiung, New York, Osaka, Portland, Taipei, and Tokyo. However as different cities have different data history, for a wider study composed of 1000 cities, a history longer than five years cannot be fulfilled. Moreover, the normality results and forecast performance would be worse for longer histories. We therefore use only up to five years' subsamples. For the sake of brevity, we present, from now on, only the results from four cities: Berlin, Kaohsiung, New York, Tokyo, and detail the other results in the online supplementary material. The four cities are from different countries and are quite representative of different types of weather relevant to the interest of weather derivative analysis. Berlin, New York, and Tokyo are cities with weather derivatives that are frequently traded, and Kaohsiung is a coasted city with atypical temperature patterns.

We first check seasonality, intertemporal correlation, and seasonal variation. Table 1 provides the coefficients of the Fourier truncated seasonal function (3) for some cities for different time periods. The coefficient a can be seen as the average temperature, the coefficient b as an indicator for a possible trend within a year. The latter coefficients are stable even when the estimation is done in a window length of 10 years. In the sense of capturing volatility peak seasons, the right panel of Figure 3 visualises the

power of capturing volatility peak seasons by the seasonal local smoother (5) using the quartic kernel over the estimates modeled under Fourier truncated series (3).

After removing the local linear seasonal mean (5) from the daily average temperatures ($X_t = T_t - \Lambda_{t,\mathrm{LLR}}$), we check that X_t is a stationary process with the augmented Dickey-Fuller (ADF) and the KPSS tests. The analysis of the partial autocorrelations and the Akaike information criterion (AIC) suggest that an AR(3) model fits the temperature evolution well. Table 2 presents the results of the stationarity tests. The temperature data and the smoothed seasonal functions are plotted on the left panel in Figure 3. To show the pattern of the squared residuals after seasonal and intertemporal fitting ($\hat{\varepsilon}^2_{t,j}$), we plot the averaged square residuals over years and show the empirical curves on the right panel in Figure 3. Besides, we have also plotted in Figure 3 the smoothed curves by using the Fourier method and the fixed bandwidth local linear method. Furthermore, we check the normality of the final residuals and present the results in the online supplementary material Tables 1– 3 (see there the Fourier method). All seasonal variance estimators lead to residuals that are far from being normally distributed. These facts are of course not an ideal platform for risk neutral pricing (based on standard stochastic financial models). The heavytailedness, as seen in Figure 1, may be attributed to an unsatisfactory extraction of the heteroscedasticity (or mean) function. As a solution we employ a localization scheme.

The adjustment in the smoothing parameter h will provide the localization in time. The bandwidth sequences are selected from six candidates: (1, 2, 3, 4, 5, 6, 7), (1, 2, 3, 5, 7, 10, 13), (3, 5, 7, 9, 11, 13, 15), (3, 5, 8, 12, 17, 23, 30), (5, 7, 10, 14, 19, 25, 32), and (7, 9, 11, 14, 17, 10, 24). These candidates are chosen according to the lowest Anderson–Darling (AD) statistic. The best candidate for the bandwidth sequence is the one which yields a residual distribution closest to the normal one. Smoothing the selected bandwidths gives another adaptive estimator, implemented, but not discussed here, due to space limitations.

The critical values as calibrated from (12) and (13) are given in Figure 4. The left hand side provides Critical Values simulated from a sample of 10,000 observations for a quartic kernel for both mean with $\theta^* = 0$ and variance with $\theta^* = 1$, $r = 0.5$ and different values of significance level α. The Critical Values for different bandwidth sequences are displayed on the right hand side of Figure 4. The critical values, as one observes, are relatively robust to the choice of r and α.

A one year period is considered in the first place for demonstration purposes, while later we show how the results change with different time periods. Figures 5, 6, 7, and 8 present the general results for the different cities under different adaptive localizing schemes for seasonal mean (Me) and seasonal variance (Va): with fixed bandwidth curve (fi), adaptive bandwidth curve (ad), and truncated Fourier (Fourier) for different time intervals. The seasonal mean is estimated jointly over the years, using $\alpha = 0.7$ and power level $r = 0.5$.

The upper panel of each variance plot in Figures 5–8 shows the sequence of bandwidths; the bottom panel displays variance estimation with fixed bandwidth (dashed line), the Fourier truncated method (dotted line), and adaptive bandwidth (solid black line). In all countries, one observes significant differences between the estimates. In particular, in cities like Kaohsiung and New York, one observes more variation of the seasonal variance curves during peak seasons (winter and summer times). The triangles and circles in the bottom panel of each variance plot help us trace the source of the nonnormality over time, since they correspond to ten dots of the upper and lower tails of the QQ-plots of square residuals respectively (see Figure 9 for the Berlin results). The top plots of Figures 5–8 show the mean case. Unlike the seasonal variance function, we do not observe a big variation of smoothness in the mean function. One can see that in all cities, the bandwidths vary over the yearly cycle with a slight degree of nonhomogeneity for Kaoshiung.

An approach to cope with the nonnormality brought in by more observations is to estimate mean functions year by year (SeMe), and then aggregate the residuals for variance estimation. We, therefore, estimate the joint/separate seasonal mean (JoMe/SeMe) and seasonal variance (Va) curves with a fixed bandwidth curve (fi) and an adaptive bandwidth curve (ad). (A summary of the estimation methods can be found in Table 3.) The average over years acts as a smoother when we consider more years. The estimated AR(L) parameters for different cities using a joint/separate mean (JoMe/SeMe) with different bandwidth curves are illustrated in Table 4. The results again show that an AR(3) fits the stylised facts of temperature well.

Kolmogorov–Smirnov (KS), Jarques–Bera (JB), and AD normality tests are taken to test the normality of the corrected residuals (after seasonal mean and variance). For each city, a rejection at 0.05 level is counted as 1 (else 0). The rejection rates over all the cities under different estimation techniques are displayed in Table 5. The results compare different periods (1–5 years) for the robustness of our methods. (Considering data histories longer than 5 years would not give us a better forecast performance and normality test results.) A higher rejection rate would indicate a poorer performance of the relevant method. To make our conclusion more general, we add 988 more cities, which are selected all around the world resulting in a total of 1000 cities, see Figures 10 and 11. The additional data are taken from NCDC Climate Data Online from 2007 to 2012. We observe a superior performance of adaptive methods over the CD method and a truncated Fourier. The JoMe adMe adVa method with one year of history leads to the rejection rate up to of 0.174 which is more than twice smaller than using other methods. Considering more years of history, the rejection rate of the CD method comes close to 1.0 based on all tests and the rejection rate for the truncated Fourier based on the KS test is around 0.25 and based on two other tests, approaches 0.8. In contrary to CD and the truncated Fourier, rejection rates from all the adaptive methods are below 0.2 for all three tests. Moreover, one observes the rejection rate below 0.01 for the KS test for all years of history using the Locave and Locsep methods. SeMe adMe adVa method keeps the rejection rate for 3–5 years of history and JB and AD tests below 0.13. The Locmax procedure has a very stable performance over all the tests and all the history, with rejection rates being bounded by 0.16. Maps with marked locations on which the analysis has been performed using the period of five years of history and most conservative tests namely JB and AD are presented in Figures 10 and 11.

Table 4. AR(L) parameters for Berlin, Kaohsiung, New York, Tokyo using joint/separate mean (JoMe/SeMe) with fixed bandwidth curve (fi), adaptive bandwidth curve (ad) seasonal mean (Me)curve. CAR(L) parameters estimated with the last 2 years.

		AR(p) parameters								CAR(p) parameters								
		1 year			2 years			3 years		4 years			5 years			2 years		
		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\alpha}_1(1)$	$\hat{\alpha}_2(1)$	$\hat{\alpha}_3(1)$
Berlin	JoMe adMe	0.744	− 0.244	− 0.008	0.972	− 0.303	0.060	0.971	− 0.286	0.072	0.960	− 0.252	0.073	0.985	− 0.288	0.101	2.015	1.318	0.202
	JoMe fiMe	0.301	− 0.330	− 0.267	0.908	− 0.283	0.113	0.937	− 0.277	0.113	0.939	− 0.252	0.097	0.976	− 0.289	0.114	2.024	1.337	0.199
	SeMe adMe adVo				0.795	− 0.255	− 0.057	0.807	− 0.293	0.000	0.727	− 0.248	− 0.038	0.768	− 0.290	0.000	2.204	1.664	0.517
	SeMe fiMe fiVo				0.344	− 0.333	− 0.196	0.346	− 0.334	− 0.176	0.294	− 0.332	− 0.175	0.326	− 0.342	− 0.165	2.655	2.644	1.185
	Locave				0.795	− 0.255	− 0.057	0.971	− 0.286	− 0.072	0.960	− 0.252	− 0.073	0.768	− 0.290	0.000	2.204	1.664	0.517
	Locsep				0.795	− 0.255	− 0.057	0.971	− 0.286	− 0.072	0.960	− 0.252	− 0.073	0.768	− 0.290	0.000	2.204	1.664	0.517
	Locmax				0.795	− 0.255	− 0.057	0.971	− 0.286	− 0.072	0.960	− 0.252	− 0.073	0.768	− 0.290	0.000	2.204	1.664	0.517
	Fourier	0.968	− 0.292	0.065	1.017	− 0.309	0.075	1.017	− 0.297	0.074	0.981	− 0.264	0.079	1.000	− 0.297	0.101	1.982	1.273	0.215
Kaohsiung	JoMe adMe	0.601	− 0.132	− 0.152	0.735	− 0.105	− 0.071	0.826	− 0.204	0.000	0.807	− 0.155	0.000	0.785	− 0.139	0.000	2.215	1.569	0.354
	JoMe fiMe	0.507	− 0.164	− 0.200	0.725	− 0.112	− 0.080	0.806	− 0.195	0.000	0.788	− 0.146	0.000	0.773	− 0.139	0.000	2.227	1.593	0.366
	SeMe adMe adVo				0.442	− 0.138	−0,152	0.401	− 0.200	− 0.154	0.412	− 0.159	− 0.132	0.316	− 0.194	− 0.146	2.557	2.253	0.084
	SeMe fiMe adVo				0.504	− 0.152	− 0.210	0.502	− 0.219	− 0.206	0.479	− 0.184	− 0.196	0.457	− 0.189	− 0.190	2.495	2.143	0.857
	Locave				0.442	− 0.138	− 0.152	0.401	− 0.200	− 0.154	0.412	− 0.159	− 0.132	0.316	− 0.194	− 0.146	2.557	2.253	0.848
	Locsep				0.442	− 0.138	− 0.152	0.401	− 0.200	− 0.154	0.412	− 0.159	− 0.132	0.316	− 0.194	− 0.146	2.557	2.253	0.848
	Locmax				0.442	− 0.138	− 0.152	0.401	− 0.200	− 0.154	0.412	− 0.159	− 0.132	0.316	− 0.194	− 0.146	2.557	2.253	0.848
	Fourier	0.820	− 0.116	− 0.020	0.791	− 0.102	− 0.051	0.853	0.185	− 0.000	0.820	− 0.128	0.011	0.802	− 0.126	0.001	2.208	1.520	0.363
New-York	JoMe adMe	0.601	− 0.207	0.000	0.636	− 0.197	0.161	0.728	− 0.212	0.127	0.688	− 0.212	0.165	0.708	− 0.190	0.140	2.292	1.774	0.342
	JoMe fiMe	0.342	− 0.277	− 0.118	0.644	− 0.205	0.138	0.692	− 0.213	0.150	0.695	− 0.222	0.158	0.711	− 0.195	0.139	2.289	1.773	0.345
	SeMe adMe adVo				0.554	− 0.247	0.000	0.495	− 0.282	0.000	0.481	− 0.287	0.000	0.475	− 0.247	− 0.050	2.289	1.773	0.345
	SeMe fiMe fiVo				0.278	− 0.312	− 0.118	0.283	− 0.322	− 0.133	0.299	− 0.326	− 0.120	0.288	− 0.308	− 0.152	2.712	2.732	1.172
	Locave				0.000	0.554	− 0.247	0.495	− 0.282	0.000	0.481	− 0.287	0.000	0.475	0.247	− 0.050	2.525	1.803	0.328
	Locsep				0.000	0.554	− 0.247	0.495	− 0.282	0.000	0.481	− 0.287	0.000	0.475	0.247	− 0.050	2.525	1.803	0.328
	Locmax				0.000	0.554	− 0.247	0.495	− 0.282	0.000	0.481	− 0.287	0.000	0.475	0.247	− 0.050	2.525	1.803	0.328
	Fourier	0.720	− 0.184	0.066	0.763	− 0.193	0.117	0.751	− 0.205	0.134	0.750	− 0.212	0.148	0.756	− 0.190	0.128	2.244	1.678	0.306
Tokyo	JoMe adMe	0.152	− 0.157	− 0.294	0.431	− 0.074	0.000	0.492	− 0.087	0.000	0.541	− 0.092	0.072	0.567	− 0.102	0.083	2.433	1.968	0.452
	JoMe fiMe	0.158	− 0.150	− 0.296	0.452	− 0.074	− 0.054	0.512	− 0.103	0.000	0.564	− 0.097	0.045	0.575	0.105	0.072	2.425	1.745	0.248
	SeMe adMe adVo				0.360	− 0.106	− 0.162	0.330	− 0.177	− 0.136	0.307	− 0.190	− 0.138	0.317	− 0.198	− 0.124	2.639	2.384	0.907
	SeMe fiMe fiVo				0.225	− 0.177	− 0.245	0.265	− 0.211	− 0.184	0.269	− 0.212	− 0.171	0.252	− 0.232	− 0.171	2.774	2.726	1.197
	Locave				0.360	− 0.106	− 0.162	0.333	− 0.177	− 0.136	0.307	− 0.190	− 0.138	0.317	− 0.198	− 0.124	2.639	2.384	0.907
	Locsep				0.360	− 0.106	− 0.162	0.333	− 0.177	− 0.136	0.307	− 0.190	− 0.138	0.317	− 0.198	− 0.124	2.639	2.384	0.907
	Locmax				0.360	− 0.106	− 0.162	0.333	− 0.177	− 0.136	0.307	− 0.190	− 0.138	0.317	− 0.198	− 0.124	2.639	2.384	0.907
	Fourier	0.534	0.038	− 0.039	0.561	− 0.039	− 0.017	0.581	− 0.091	0.044	0.597	− 0.088	0.060	0.615	− 0.096	0.079	2.438	1.916	0.495

Table 5. Rejection rates of the normality at 5% level for 1000 cities with different history, methods of estimation, and normality tests.

	Method	KS	JB	AD		KS	JB	AD
1 Year	JoMe adMe adVa	*0.000*	*0.174*	*0.164*				
	JoMe fiMe fiVa	0.006	0.200	0.270				
	Fourier	0.049	0.378	0.327				
	CD	0.086	0.499	0.426				
2 Years	JoMe adMe adVa	1.000	0.224	0.202	3 Years	0.968	0.354	0.367
	JoMe fiMe fiVa	0.998	0.431	0.390		0.869	0.571	0.533
	SeMe adMe adVa	1.000	0.073	0.043		1.000	*0.044*	*0.072*
	SeMe fiMe fiVa	1.000	0.305	0.261		0.976	0.358	0.367
	Locave	*0.001*	0.057	0.072		*0.004*	0.082	0.118
	Locsep	*0.001*	0.057	0.072		*0.004*	0.082	0.118
	Locmax	0.010	*0.051*	*0.034*		0.024	0.074	0.080
	Fourier	0.109	0.516	0.472		0.180	0.685	0.599
	CD	1.000	0.715	0.642		1.000	0.828	0.769
4 Years	JoMe adMe adVa	0.767	0.480	0.478	5 Years	0.585	0.547	0.539
	JoMe fiMe fiVa	0.608	0.646	0.618		0.483	0.702	0.669
	SeMe adMe adVa	0.975	*0.064*	*0.090*		0.747	*0.081*	*0.124*
	SeMe fiMe fiVa	0.778	0.468	0.433		0.463	0.546	0.506
	Locave	*0.007*	0.129	0.155		*0.009*	0.174	0.210
	Locsep	*0.007*	0.129	0.155		*0.009*	0.174	0.210
	Locmax	0.029	0.135	0.111		0.031	0.157	0.145
	Fourier	0.256	0.766	0.677		0.305	0.816	0.740
	CD	1.000	0.880	0.801		1.000	0.916	0.832

NOTE: Tests for normality are Kolmogorov–Smirnov (KS), Jarque–Bera (JB) and AD. Methods used: joint/separate mean (JoMe/SeMe) with fixed/adaptive (fi/ad) bandwidth for the mean/variance (Me/Va), Locave, Locsep, Locmax, truncated Fourier (Fourier) and CD model. Highlighted in italic are models with the smallest rejection rate for each goodness-of-fit (GoF) test and each history.

Cities marked in blue are those, where the normality at a 5% level cannot be rejected using JB (Figure 10) and AD (Figure 11) tests, otherwise cities are marked in red. One clearly sees dominance of blue marked cities in the Locmax method (in both figures top left map) and the dominance of red marked cities in the other subplots. More detailed results for only 12 original cities can be found in the online supplementary material.

4. Forecast and Comparison

In this section, we compare the forecasting accuracy of the proposed models to the CD model. CD mentions that their point forecasts are always at least as good as the persistence and climatological forecasts, although not so good as the judgmentally adjusted NWP forecast produced by EarthSat for a horizon of eight days. Therefore, a good performance of the technique presented here could potentially suggest that our time series model is relevant for weather derivatives. and

In Figures 12 and 13 we compare the out-of-sample forecast performance between five methods, namely SeMe adMe adVo, Locmax, JoMe adMe adVo, truncated Fourier, and CD. The comparison is provided at different time horizons (1, . . . , 150 days) for Berlin, Kaohsiung, New York and Tokyo using 2 (Figure 12) and 3 (Figure 13) years of history. These figures contain information both on point forecast and interval forecast. The top panel of each plot shows the absolute deviation of the forecasted temperature from the true one, averaged over 10,000 simulation paths. This may be considered as the quality of the point forecast. In these terms, as we see in most cities and over all time horizons, we have at least one localizing method better than the CD method. The lower panel of each plot shows the averaged width of the point-wise confidence interval based on 10,000 sample paths. These curves represent the efficiency of the models. Although the truncated Fourier series method also looks quite competitive in the point forecast, it usually has a very wide confidence interval, which is a sign of low efficiency. Other methods in this context are strictly better. The middle panel shows the coverage of the true temperature by the confidence interval, where larger values represent higher quality. In terms of interval forecast, we can see that from Figure 12 and 13 for most cities, we have at least one model which has better coverage with moderate width of confidence intervals. Moreover, we do not see outperforming behavior of the CD method over proposed adaptive techniques in almost all 12 cities. As a conclusion, we do not claim strict superiority over the CD method in forecasting, but conclude, that both methods are quite competitive.

5. A Temperature Pricing Example

Based on a model for the daily temperature evolution, futures and European options written on temperature indices traded at the Chicago Mercantile Exchange (CME) can be calibrated. Temperature futures are contracts written on different temperature indices measured over specified periods $[\tau_1, \tau_2]$ such as weeks, months, or quarters of a year. Temperature futures allow one party to profit if the realized index value is greater than a predetermined strike level and the other party benefits if the index value is below. The owner of a call (put) option written on futures $F(t, \tau_1, \tau_2)$ with exercise time $t \le \tau_1$ and measurement period $[\tau_1, \tau_2]$ will receive $\max\{F(t, \tau_1, \tau_2) - K, 0\}$ ($\max\{K - F(t, \tau_1, \tau_2), 0\}$), where K denotes the strike level. In other words, in exchange for the payment of the premium, the call (put) option gives the buyer a payoff based upon the difference between the realized index value and the strike level.

The most common temperature indices $I(\tau_1, \tau_2)$ are: Heating Degree Day (HDD), Cooling Degree Day (CDD), Cumulative Averages Temperatures (CAT), or Average Accumulative Temperatures (AAT). The CAT index takes the accumulated average temperature over $[\tau_1, \tau_2]$:

$$\mathrm{CAT}(\tau_1, \tau_2) = \int_{\tau_1}^{\tau_2} T_u du,$$

where $T_u = (T_{u,\max} + T_{u,\min})/2$ denotes the daily average temperature. The measurement period is usually defined in months or season. The HDD index measures the cumulative amount of average temperature below a threshold (typically 18°C or 65°F) over a period $[\tau_1, \tau_2]$: $\int_{\tau_1}^{\tau_2} \max(c - T_u, 0) du$. Similarly, the CDD index accumulates $\max(T_u - c, 0)$. At CME, CAT/CDD futures are traded for European cities, CDD/HDD for the U.S., Canadian, and Australian cities, and AAT for Japanese cities. Note that these temperature indices are the underlying and not the temperature itself. The options at CME are cash settled, that is, the owner of a future receives 20 times the Degree Day Index at the end of the measurement period, in return for a fixed price. At time t, CME trades different contracts $i = 1, \ldots, N$ with measurement period $0 \le t \le \tau_1^i < \tau_2^i$ (usually the length between τ_1^i and τ_2^i is one month). For example, a contract with $i = 7$ is six months ahead from the trading day t. For the U.S. and Europe CAT/CDD/HDD futures, N is usually equal to 7

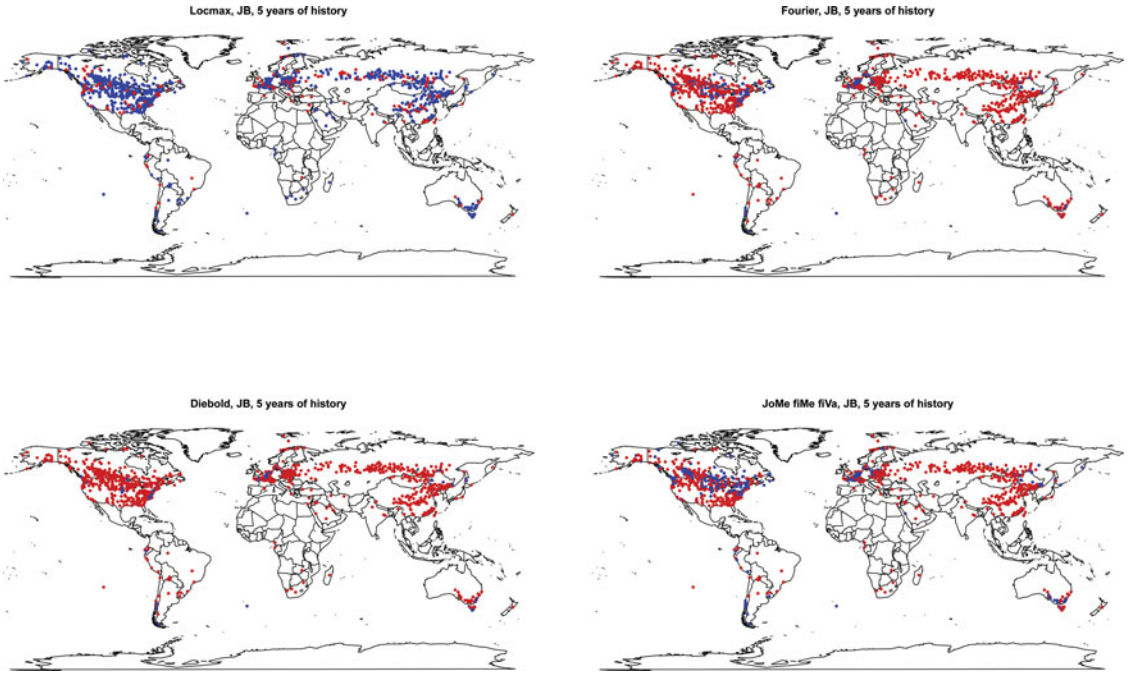


Figure 10. Map of locations where temperature are collected over period 2007–2011. Cities marked in blue do not reject the normality at 5% using the JB test and cities marked red do reject the normality hypothesis. In a clockwise direction, used methods are Locmax, truncated Fourier, CD, and JoMe fiMe fiVa.

(April–November or November–April), while for Asia, $N = 12$ (January–December).

Recall that we adopt the CAR(L) model in (1) for the detrended temperature time series, and the autoregressive process AR(L) in (2) can be seen as a discretely sampled continuous time process (CAR(L)) (1) driven by one dimensional Brownian motion. The detailed demonstration can be found in the Appendix A.2.

The fact that temperature's random factor is close to the normal distribution, as disclosed in the analysis of the residuals before, motivates the use of a Brownian motion as the noise in the Ornstein-Uhlenbeck process. Moreover ACF-plots of the squared residuals presented in the online supplementary material demonstrate the success of the localizing method to explain deterministic variations in temperature data. They do not show signs of stochastic volatility: the squared residuals do not have an exponentially decaying ACF. This contradicts results found in Benth and Benth (2011) and Benth and Benth (2012) and suggests to us that the non-Gaussian shocks found in the literature are the result of model mis-specification. The

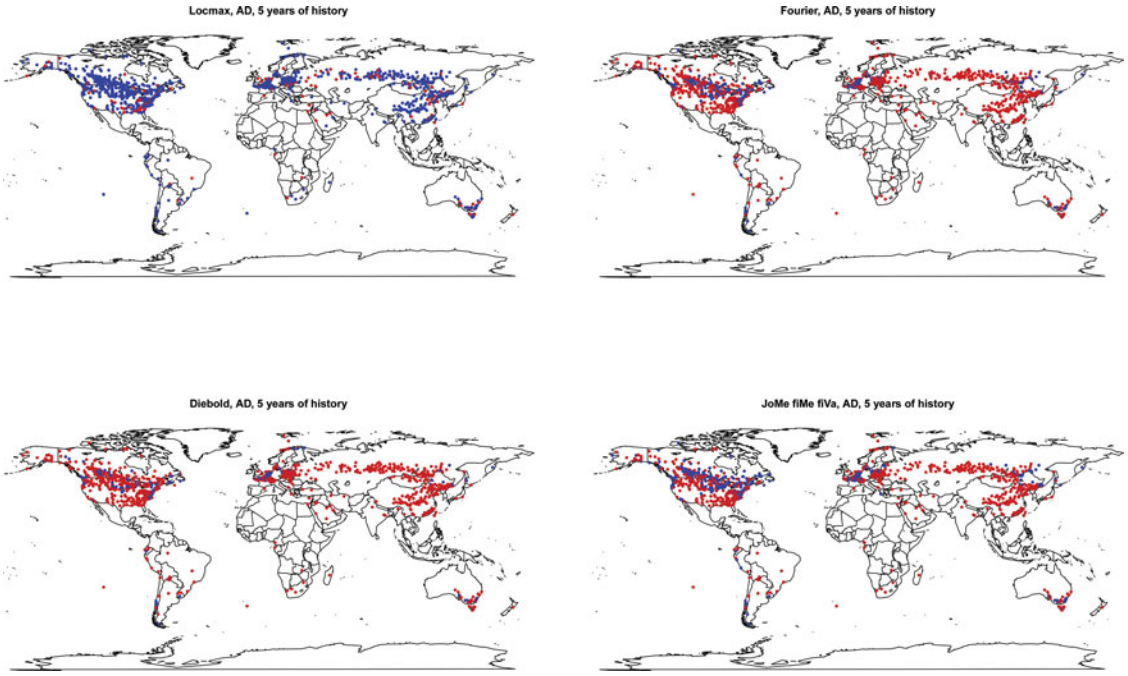


Figure 11. Map of locations where temperature are collected over period 2007–2011. Cities marked in blue do not reject the normality at 5% using the AD test and cities marked red do reject the normality hypothesis. In a clockwise direction, used methods are Locmax, truncated Fourier, CD and JoMe fiMe fiVa.

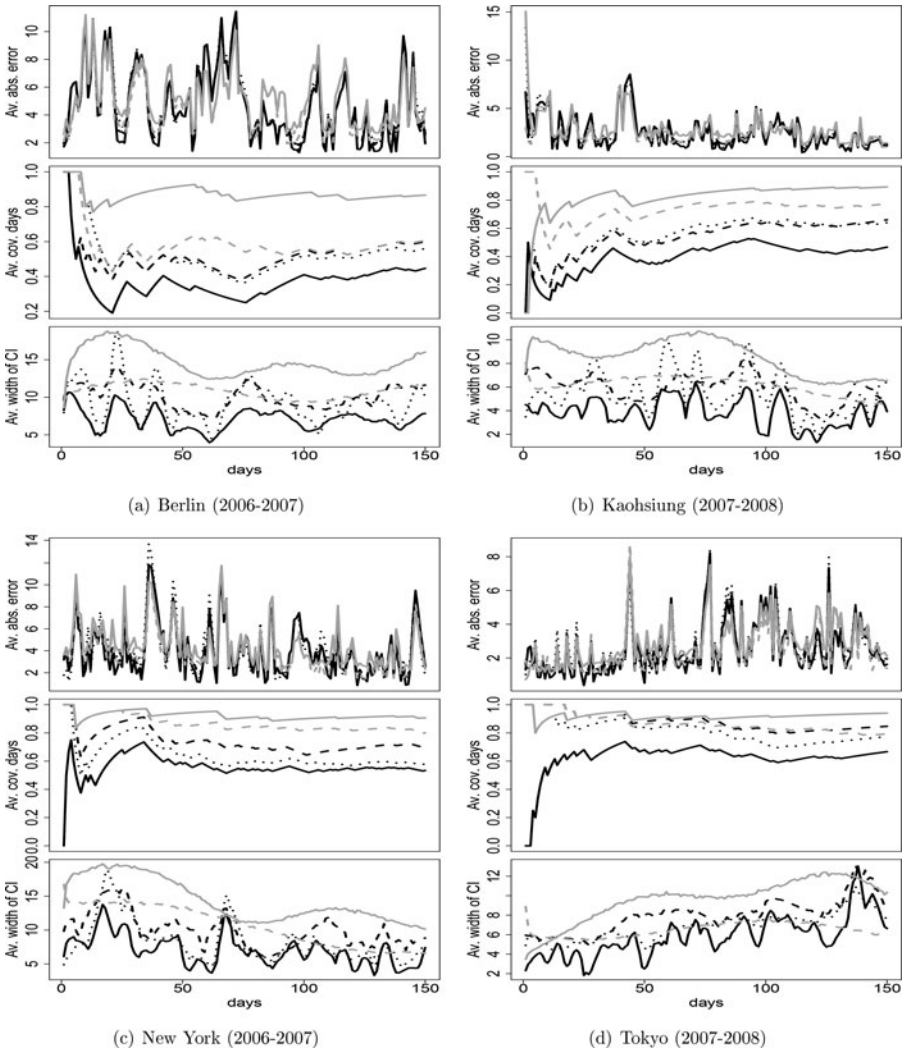


Figure 12. $h = 1, \ldots, 150$ days (X axis) ahead forecast for Berlin, Kaohsiung, New York and Tokyo (left to right, top to bottom); averaged absolute error (Y axis, upper panel), averaged coverage days (Y axis, middle panel), averaged width of the confidence 95% intervals (Y axis, lower panel), SeMe adMe adVo (solid black), Locmax (dashed gray), JoMe adMe adVo (dotted black), truncated Fourier (solid gray), CD (dashed black), fitted using 2 years of historical data and 10,000 samples.

continuous analogue of the CD model is, however, difficult to estimate. Thus the model in (1) is simpler than CD's one and provides a better fit to the data.

The temperature futures price is the risk adjusted index, given today's filtration $\mathcal{F}_t$

$$F_I(t, \tau_1, \tau_2) = \mathsf{E}^{\,Q}\left[I(\tau_1, \tau_2)|\mathcal{F}_t\right], \tag{14}$$

with $I(\tau_1, \tau_2)$ being one of the indices CAT, HDD or CDD. The expectation is computed under a risk neutral pricing probability Q and is equivalent to the physical measure P under which the discounted temperature index is a Q-martingale. To evaluate (14), we need to know the temperature index dynamics under Q. We restrict the class of pricing probabilities to those that can be parameterized via $Q = Q_\lambda$, where equivalent changes of measures are simply associated with changes of drift. Thus, in the modeling of the dynamics of futures prices written on temperature indices, it is natural to define a parameter measuring the market price of risk (MPR) λ_t, which can be calibrated from traded (CAT/CDD/HDD) derivative type contracts. The temperature dynamics in (1) under Q_λ become:

$$d\mathbf{X}_t = (\mathbf{A}\mathbf{X}_t + \mathbf{e}_L\sigma_t\lambda_t)dt + \mathbf{e}_L\sigma_t dB_t^\lambda, \tag{15}$$

where B_t^λ is a Brownian motion for any time before the end of the trading time and a martingale under Q_λ. Then, for $0 \le t \le \tau_1 < \tau_2$, the explicit form of an CAT futures price is given by

$$\begin{aligned} F_{\mathrm{CAT}}(t, \tau_1, \tau_2, \Lambda_t, \sigma_t, \lambda_t) &= \mathsf{E}^{\,Q_\lambda}\left[\int_{\tau_1}^{\tau_2} T_u du|\mathcal{F}_t\right] = \int_{\tau_1}^{\tau_2} \Lambda_u du + \mathbf{a}_{t,\tau_1,\tau_2}\mathbf{X}_t \\ &+ \int_t^{\tau_1} \lambda_u\sigma_u\mathbf{a}_{t,\tau_1,\tau_2}\mathbf{e}_L du \\ &+ \int_{\tau_1}^{\tau_2} \lambda_u\sigma_u\mathbf{e}_1^\top\mathbf{A}^{-1}\left[\exp\{\mathbf{A}(\tau_2 - u)\} - I_L\right]\mathbf{e}_L du, \end{aligned} \tag{16}$$

with $\mathbf{a}_{t,\tau_1,\tau_2} = \mathbf{e}_1^\top\mathbf{A}^{-1}[\exp\{\mathbf{A}(\tau_2 - t)\} - \exp\{\mathbf{A}(\tau_1 - t)\}]$ and I_L the $L \times L$ identity matrix. Similarly one can compute the price dynamics of CDD and HDD, see (Benth, Benth, and Koekebakker 2007). The CAR model (1) provides the analytical formula (16). Note that all constituents except λ_t in the left and right side of (16) are known or estimable (Λ_t and σ_t are out-of-sample estimates as in the previous section), hence the calibration of the MPR from market data turns out to be an inverse problem in terms of λ_t.

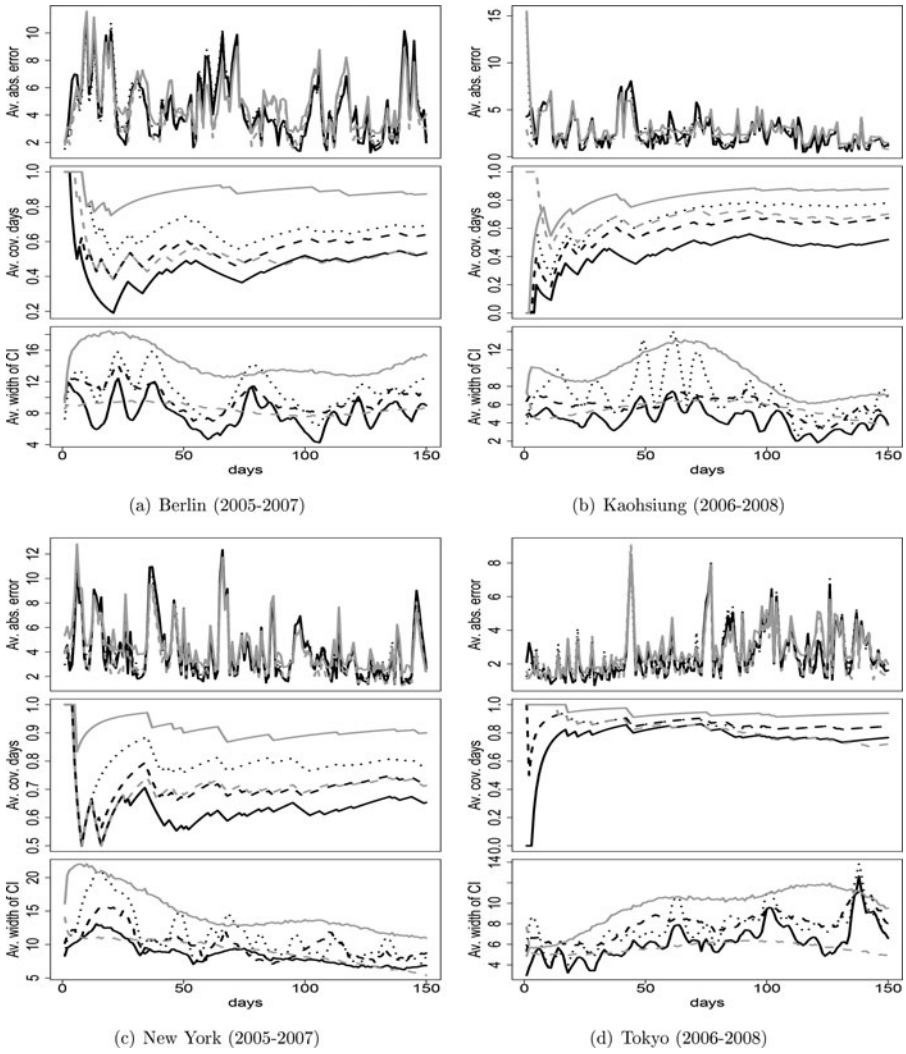


Figure 13. $h = 1, \ldots, 150$ days (X axis) ahead forecast for Berlin, Kaohsiung, New York and Tokyo (left to right, top to bottom); averaged absolute error (Y axis, upper panel), averaged coverage days (Y axis, middle panel), averaged width of the confidence 95% intervals (Y axis, lower panel), SeMe adMe adVo (solid black), Locmax (dashed gray), JoMe adMe adVo (dotted black), truncated Fourier (solid gray), CD (dashed black), fitted using 3 years of historical data and 10,000 samples.

Assuming that the parameterization of the MPR is of a constant form for each observed contract ($\lambda_u = \lambda_{t,\tau_1^i,\tau_2^i}$ in (16) for $u \in [\tau_1, \tau_2]$), one can calibrate the MPR for every combination of (t, τ_1^i, τ_2^i), $i = 1, \ldots, N$ contracts, by inverting the pricing formulas in (16) with the observed CME market prices at time t, $(F_{t,i,\mathrm{CME}})$ with respect to λ as:

$$\hat{\lambda}_{t,\tau_1^i,\tau_2^i} = \arg\min_{\lambda} |F_{\mathrm{CAT}}(t, \tau_1^i, \tau_2^i, \hat{\Lambda}_t, \hat{\sigma}_t, \lambda) - F_{t,i,\mathrm{CME}}|. \quad (17)$$

We name $\hat{\lambda}_{t,\tau_1^i,\tau_2^i}$ as implied MPR. For fixed time t, assuming that λ_t remains the same for different contracts with different maturities, to evaluate the estimation of $\hat{\lambda}_t$ for a particular contract i, the observed price $F_{t,i,\mathrm{CME}}$ for this contract can be excluded for the estimation. We have then the cross-validated estimation by leaving one contract out:

$$\hat{\lambda}_{t,\tau_1^i,\tau_2^i,CV} = \arg\min_{\lambda} \sum_{j=1;\, j\neq i}^{N} \{F_{\mathrm{CAT}}(t, \tau_1^j, \tau_2^j, \hat{\Lambda}_t, \hat{\sigma}_t, \lambda) - F_{t,j,\mathrm{CME}}\}^2. \quad (18)$$

Other specifications of the MPR for temperature derivatives have been explored in Härdle and López Cabrera (2012), where the authors argue that a constant MPR is sufficient for pricing purposes. This might be compared with complete markets, where the MPR is minus the Sharp ratio $(\mu_t - r)/\sigma_t^F$, where μ_t and σ_t^F denote the mean and standard deviation of traded futures, and r is the risk free interest rate. From now on, pricing follows (16) with an MPR from (17), (18) and with Λ_t and σ_t estimated via the localization techniques.

Observe that calibrations in (17), (18) are only valid if a weather derivative market exists, like for example for Berlin and Tokyo. To price temperature derivatives for regions with no weather derivative markets, like Kaohsiung, one can use the implied MPR of traded futures of a neighboring market, for example, Tokyo AAT futures. Thus, by finding a relationship between the MPR and the seasonal variance one can use this as a proxy to price over the counter (OTC) AAT futures for Kaohsiung. This is acceptable since the stylized facts of temperature in Tokyo reveal similarities to that of Kaohsiung. However, generally we are aware of arbitrage opportunities across the two different markets, therefore this approach cannot be generalized for every second weather derivative markets. Considering that the MPR is a risk premium per unit of volatility, one can project the implied MPR on the state variables related to volatility. An

Table 6. RMSE between the weather futures listed at CME and estimated weather futures $F_I(t, \tau_1^i, \tau_2^i, \hat{\Lambda}_t, \hat{\sigma}_t, \hat{\lambda}_{t-1})$ with $\hat{\lambda}_{t-1} = \hat{\lambda}_{t-1,CV}$.

			RMSE between models' prices and F_{CME} $\hat{\lambda}_{t-1,CV}$			
Type	MP	n	AdaptBW	FixedBW	Locmax	Fourier
Berlin-CAT	200704	230	2.868	*2.617*	2.876	9.665
Berlin-CAT	200705	6	79.802	84.078	*79.169*	126.8
Berlin-CAT	200706	58	*2.033*	3.078	2.662	68.262
Berlin-CAT	200707	79	*31.774*	46.633	32.565	45.125
Berlin-CAT	200709	121	*25.17*	39.337	25.485	26.773
Essen-CAT	200804	74	*75.676*	75.686	*75.676*	76.519
Essen-CAT	200805	100	21.871	*21.845*	21.871	22.628
Essen-CAT	200806	79	7.225	*7.131*	7.225	19.15
Essen-CAT	200807	140	*59.392*	59.47	*59.392*	62.318
Essen-CAT	200808	164	*73.511*	73.548	*73.511*	74.469
Essen-CAT	200809	181	6.885	*6.837*	6.885	12.932
London-CAT	200805	100	43.06	*32.377*	40.505	58.495
London-CAT	200806	40	*1.461*	2.56	2.709	6.063
London-CAT	200807	142	*2.467*	2.824	4.745	9.81
London-CAT	200808	163	27.333	27.204	*26.88*	31.23
London-CAT	200809	184	*36.201*	37.255	37.941	41.861
Tokyo-AAT	200903	18	4.922	*1.354*	8.418	26.287
Tokyo-AAT	200904	18	*28.967*	29.401	56.975	76.489
Tokyo-AAT	200905	18	58.553	*54.353*	90.269	77.8
Tokyo-AAT	200906	18	*49.993*	52.228	52.678	*16.35*
Tokyo-AAT	200907	18	24.093	27.72	*21.954*	42.34

NOTE: τ_1^i and τ_2^i are the first and the last day of the measurement period (MP, yyyymm), respectively. Prices are estimated under different estimations schemes ($\hat{\Lambda}_t$, $\hat{\sigma}_t$ under AdaptBW, FixedBW, Locmax, and truncated Fourier). *n* corresponds to the number of trading days for a given MP.

insight into Tokyo's AAT futures, which can be employed for the Kaohsiung case, can be achieved by regressing the averaged implied MPR (17) against the variation:

$$\hat{\lambda}_{\tau_1^i,\tau_2^i} = 4.08 - 2.19\hat{\sigma}^2_{\tau_1^i,\tau_2^i} + 0.28\hat{\sigma}^4_{\tau_1^i,\tau_2^i},$$

where $\hat{\lambda}_{\tau_1^i,\tau_2^i} \overset{\text{def}}{=} (\tau_2^i - \tau_1^i)^{-1} \sum_{t=\tau_1^i}^{\tau_2^i} \hat{\lambda}_{t,\tau_1^i,\tau_2^i}$, $\hat{\sigma}^2_{\tau_1^i,\tau_2^i} \overset{\text{def}}{=} (\tau_2^i - \tau_1^i)^{-1} \sum_{t=\tau_1^i}^{\tau_2^i} \hat{\sigma}_t^2$, $\hat{\sigma}^4_{\tau_1^i,\tau_2^i} \overset{\text{def}}{=} (\tau_2^i - \tau_1^i)^{-1} \sum_{t=\tau_1^i}^{\tau_2^i} \hat{\sigma}_t^4$ and $R^2_{adj} = 0.71$. Plugging the corresponding $\hat{\sigma}_t^2$, $\hat{\sigma}_t^4$ values for Kaohsiung into this equation let us price such a non-CME traded weather derivative via (16).

We compare the prices obtained with localization procedures ('localized' prices) for Λ_t and σ_t (SeMe adMe adVo (AdaptBW), Locmax) with prices estimated under fixed bandwidth (SeMe fiMe fiVo (FixedBW)) and truncated Fourier series.

To judge the performance of the models, we compute the root mean squared errors (RMSE) between the market prices $F_{t,i,\text{CME}}$ (benchmark) and the estimated out-of-sample model prices $F_I(t, \tau_1^i, \tau_2^i, \hat{\Lambda}_t, \hat{\sigma}_t, \hat{\lambda}_{t-1,\tau_1^i,\tau_2^i,CV})$ $(i = 1, \ldots, \text{N})$:

$$\text{RMSE}(\tau_1^i, \tau_2^i) = \sqrt{|\mathfrak{T}|^{-1} \sum_{t\in\mathfrak{T}} \left\{F_I(t, \tau_1^i, \tau_2^i, \hat{\Lambda}_t, \hat{\sigma}_t, \hat{\lambda}_{t-1,\tau_1^i,\tau_2^i,CV}) - F_{t,i,\text{CME}}\right\}^2},$$

in Table 6, where $\mathfrak{T}$ is the set of days when the contract i with the measurement period (τ_1^i, τ_2^i) was traded. The results show smaller RMSE when futures prices are estimated via localization techniques, which in general outperforms the prices based on the truncated Fourier series (Benth). This suggests that our calibrated MPR embeds information on the risk and uncertainty in the market, which is helpful in analyzing market risk. Also, as mentioned before, this information may help to price OTC derivatives in the same market.

These results provide insight on pricing related to the stylized facts (seasonal effect, intertemporal correlation, etc.) of weather data. The role of the terms in the CAT futures price formally confirms this. To illustrate this point, consider, for example, the purchase of a May CAT contract for Berlin on 20070427, which starts measurement at time $\tau_1 = 20,070,501$ and finished at $\tau_2 = 20,070,531$. Setting a constant MPR (for example $\lambda = 0.20$), the first term of (16) is equal to 431.060, the second, third and fourth terms lead to 11.531, 0.8690, and 13.5390, respectively. The seasonal effect in mean Λ_t plays an important role in the level of the futures price, as it explains almost 94% of the price which is 457. Observe that the seasonal volatility σ_t also contributes to the CAT futures price since it enters in the second term (hidden in $\mathbf{X}_t$) and in the last two terms of the CAT pricing formulas. Therefore, as we get closer to the measurement period, temperature variations given by the seasonal variance (σ_t^2) will contribute to the futures prices and clearly display the Samuelson effect that is typical in mean-reverting markets: at any given time, seasonal volatility decreases with time to delivery.

6. Conclusions

We show that temperature risk stochastics are closer to Gaussian when applying adaptive statistical methods for seasonal mean and seasonal variance. This suggests to us that the non-Gaussian shocks found in the literature are truly a result of misspecification. We found that the localization method performs well, and it is robust to the specification given for Λ_t or σ_t. Moreover, intertemporal correlations demonstrate the success of the localizing methods to explain deterministic variations in temperature data. We also observed that the proposed method outperforms the standard estimation methods in most of the cases. Our results provide important insights into how weather is priced at the CME and how the observed prices conform with the stylized facts of weather data. Finally, our adaptive technique on localizing temperature risk is both an excellent temperature modeling tool as well as a novel and more market driven pricer.

Supplementary Materials

In the supplementary materials, we provide the estimation, normality tests and forecast results for the eight cities mentioned but not presented in our article. Also the technical details are provided in the supplementary materials.

References

Benth, F. E., and Benth, S. (2011), "Weather Derivatives and Stochastic Modelling of Temperature," *International Journal of Stochastic Analysis*, 2011, 1–21. [1504]

——— (2012), "A Critical View on Temperature Modelling for Application in Weather Derivatives Markets," *Energy Economics*, 34, 592–602. [1495,1504]

Benth, F. E., Benth, S., and Koekebakker, S. (2007), "Putting a Price on Temperature," *Scandinavian Journal of Statistics*, 34, 746–767. [1491,1494,1495,1505]

Benth, F. E., Härdle, W. K., and López Cabrera, B. (2011), "Pricing Asian Temperature Risk," in *Statistical Tools for Finance and Insurance* (2nd

ed.), eds. P. Cizek, W. K. Härdle, and R. Weron, Heidelberg: Springer Verlag. [1491]

Campbell, S., and Diebold, F. X. (2005), "Weather Forecasting for Weather Derivatives," *Journal of the American Statistical Association*, 100, 6–16. [1491,1494,1495,1499]

Chen, Y., Härdle, W. K., and Pigorsch, U. (2010), "Localized Realized Volatility Modelling," *Journal of the American Statistical Association*, 105, 1376–1393. [1497]

Cízek, P., Härdle, W. K., and Spokoiny, V. (2009), "Adaptive Pointwise Estimation in Time-Inhomogeneous Conditional Heteroschedasticity Models," *The Econometrics Journal*, 12, 248–271. [1497]

Härdle, W. K., and López Cabrera, B. (2012), "The Implied Market Price of Weather Risk," *Applied Mathematical Finance*, 19, 59–95. [1494,1495,1506]

Horst, U., and Mueller, M. (2007), "On the Spanning Property of Risk Bonds Priced by Equilibrium," *Mathematics of Operation Research*, 32, 784–807. [1491]

Karatzas, I., and Shreve, S. (2001), *Methods of Mathematical Finance*, New York: Springer Verlag. [1491]

Spokoiny, V. (2009), "Multiscale Local Change Point Detection With Applications to Value at Risk," *The Annals of Statistics*, 37, 1405–1436. [1492]

Rev Quant Finan Acc (2017) 49:949–971
DOI 10.1007/s11156-016-0613-x

ORIGINAL RESEARCH

Copula-based factor model for credit risk analysis

Meng-Jou Lu[1,2] · **Cathy Yi-Hsuan Chen**[2,3] · **Wolfgang Karl Härdle**[2,4]

Published online: 22 December 2016

Abstract A standard quantitative method to assess credit risk employs a factor model based on joint multivariate normal distribution properties. By extending the one-factor Gaussian copula model to produce a more accurate default forecast, this paper proposes the incorporation of a state-dependent recovery rate into the conditional factor loading and to model them sharing a unique common factor. The common factor governs the default rate and recovery rate simultaneously, implicitly creating their association. In accordance with Basel III, this paper shows that the tendency toward default during a hectic period is governed more by systematic risk than by idiosyncratic risk. Among those considered, the model with random factor loading and a state-dependent recovery rate is shown to be superior in terms of default prediction.

Keywords Factor model · Conditional factor loading · State-dependent recovery rate

JEL classification C38 · C53 · F34 · G11 · G17

✉ Meng-Jou Lu
mangrou@gmail.com

Cathy Yi-Hsuan Chen
cathy1107@gmail.com

Wolfgang Karl Härdle
haerdle@hu-berlin.de

[1] Department of Information Management and Finance, National Chiao Tung University, 1001 Daxue Rd., Hsinchu City 300, Taiwan

[2] Ladislaus von Bortkiewicz Chair of Statistics, C.A.S.E. – Center for Applied Statistics and Economics, Humboldt–Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

[3] Department of Finance, Chung Hua University, 707 WuFu Rd., Hsinchu 300, Taiwan

[4] School of Business, Singapore Management University, 50 Stamford Road, Singapore 178899, Singapore

1 Introduction

The global economy has repeatedly witnessed clusters of default events, such as the burst of the dotcom bubble in 2001 and the global financial crisis from 2007 to 2009. Clusters of default events have been blamed on the role played by systematic risk in leading to default. To reveal this role, numerous studies emphasize the role of systematic risk by employing a factor model (Andersen and Sidenius 2004; Pan and Singleton 2008; Rosen and Saunders 2010). The factor model is a common method of capturing obligors' shared behavior through a joint common factor and of reducing the dimension of dependence parameters, which benefits bond portfolio management. However, it is also relatively common to see certain unrealistic settings in this method, such as constant and linear dependence structures with thin tails of embedded risk factor distribution.

The factor copula model imposes a dependence structure on common factors and on the variables of interest. In measuring credit risk using systematic factors, the factor loading represents the sensitivity of the nth obligor to the systematic factor. All the correlations between obligors thus arise from their dependence on the common factor, and the common factor thus plays a major role in determining their joint dependence. By incorporating factor copula model into credit risk modeling, we can decompose a latent variable into its systematic and idiosyncratic components, which are independent of one another. A latent variable typically acts as a proxy for a firms' assets or liquidation value (Andersen and Sidenius 2004). Default is triggered by company asset values falling below a threshold that corresponds to a fraction of company debt (Merton 1974). In this model, credit risk is measured by a Gaussian random default variable derived from firm asset value that is latent and modeled by a factor copula framework. The implied firm value from the model ideally projects the default time we desire; thus, the lower the firm value, the shorter default the time is.

A constant factor loading assumption embedded in a one-factor Gaussian model is inconsistent with the fact that the loading on common factors varies over time, which hampers the measurement of the dependency structures of obligors. In fact, this observation is at the core of research on the mispricing of structured products (Choroś-Tomczyk et al. 2013, 2014). Longin and Solnik (2001) and Ang and Chen (2002) argue that a "correlation breakdown" structure acts better in the dependence specification. In particular, if we set the factor loading to be constant, we may underestimate default risk as the market turns downward. Our simulation and empirical evidence show that a greater factor loading in a market downturn leads to a higher contribution of common factors on firm value.

In addition to the factor-loading specification, the recovery rate is a critical and essential component in calculating the portfolio loss function. According to Table 1, a state-dependent recovery rate model is suggested since the recovery rate seems to be subject to market conditions, i.e., higher in a bull market and lower in a bear market. Close observation reveals a lower average annual recovery rate in the periods from 1999 to 2002 (internet bubble) and from 2008 to 2009 (US subprime crisis) than in the remaining periods with bullish prospects, as it is assumed that the recovery rate in a bull market should not be lower than in a bear market. Therefore, the recovery rate is likely to vary with market conditions, which resembles the behavior of the default rate. Notably, the market condition is the unique common factor shared by the recovery rate and default rate and causes their time variations.

Table 1 Annual defaulted corporate bond recoveries

Year	Bond					
	Sr. Sec. (%)	Sr. Unsec. (%)	Sr. Sub. (%)	Sub. (%)	Jr. Sub. (%)	All Bonds (%)
1997	75.5	56.1	44.7	33.1	30.6	48.8
1998	46.8	39.5	45.0	18.2	62.0	38.3
1999	36.0	38.0	26.9	35.6	n.a.	33.8
2000	38.6	24.2	20.8	31.9	7.0	25.1
2001	31.7	21.2	19.8	15.9	47.0	21.6
2002	50.6	29.5	21.4	23.4	n.a.	29.7
2003	69.2	41.9	37.2	12.3	n.a.	41.2
2004	73.3	52.1	42.3	94.0	n.a.	58.5
2005	71.9	54.9	32.8	51.3	n.a.	56.5
2006	74.6	55.0	41.4	56.1	n.a.	55.0
2007	80.6	53.7	56.2	n.a.	n.a.	55.1
2008	54.9	33.2	23.3	23.6	n.a.	33.9
2009	37.5	36.9	22.7	45.3	n.a.	33.9

Annual corporate bond recovery rates based on post-default trading price, Moody's 27th annual default study. Sr. Sec., Sr. Unsec., Sr. Sub., Sub., and Jr. Sub. represent senior secured, senior unsecured, senior subordinated, subordinated and junior subordinated, respectively

Andersen and Sidenius (2004) address the fact that both default events and recovery rates are driven by a single factor but with an independence assumption between default and recovery rate, although there are reasons to doubt this assumption. Chen (2010) demonstrates that recovery rates are strongly negatively correlated with default rates (which is given as -0.82). As a consequence, the dependence between them relies on the common factor, which is represented by the macroeconomic state. We claim that the common factor (the market) governs the default rate and recovery rate simultaneously, implicitly creating their association. One of our purposes is to build a tractable model that can reflect the obligors' behavior in reacting to the impact from the market. In addition, we show that systematic risk plays a critical role in credit measurement and prediction, and it contributes more to a firm's credit risk during a market downturn than during a tranquil period. In this sense, the factor loading on the common factor is conditional on market states. This conditional specification enables risk managers to be alerted regarding the deteriorating credit risk conditions when the market turns downward, which prevents underestimating the default probability.

We extend the one-factor Gaussian copula model in two ways. First, to improve the factor loading of Andersen and Sidenius (2004) given a two-point distribution, we apply the state-dependent concept from Kim and Finger (2000) with specific distributions to characterize the correlations in hectic or quiet periods. This concept potentially captures two typical features of equity index distributions: fat tails and a skew to the left. However, for a two-point distribution setting, it is difficult to decide on the threshold level of the two-point distribution and on a time to be chosen arbitrarily. Second, by relaxing the constant recovery rate that is naively presumed by both scholars and practitioners, our state-dependent recovery rate model allows the systematic risk factor to determine loss given default (LGD), as suggested by Amraoui et al. (2012). In addition, it restricts the recovery

rate, as a percentage of the notional is bounded on [0,1] to achieve the tractable and numerically efficient missions. In summary, our contributions include incorporating the state-dependent recovery rate into the conditional factor copula model, and we model them by sharing their unique common factor. The common factor governs the default rate and recovery rate simultaneously, while creating their association implicitly. Our Monte Carlo simulation and empirical evidence appropriately reflect this feature.

We propose four competing default models that have been widely applied to measure credit risk, and we evaluate their performances on the accuracy of forecasting default in the following year. By mapping the various factor copula models developed in the literature to the competing models, this comparison fosters a discussion on model performance. Therefore, to achieve a broader and robust comparison, we group the factor copula models developed in the literature into four competing models: (1) the FC model, i.e., the standard one-factor Gaussian copula model with a constant recovery rate (Van der Voort 2007; Rosen and Saunders 2010); (2) the RFL model, i.e., a one-factor Gaussian copula model in which the factor loadings are tied to the state of the common factor and the recoveries are assumed to be constant (Kalemanova et al. 2007; Chen et al. 2014); (3) the RR model, i.e., a standard one-factor Gaussian copula model in which the recoveries are related to the macroeconomic state (Amraoui and Hitier 2008; Elouerkhaoui 2009; Amraoui et al. 2012); and (4) the RRFL model, i.e., a conditional factor loading specification together with a state-dependent recovery rate, which is the model that we are developing. If the empirical results show that it shows superior performance in predicting default, then the outstanding performance of our refined RRFL model will be clear.

In the FC model, we estimate the Pearson's correlation coefficient between each obligor and the common factor and set the recovery rate as constant. This is a conventional model used to measure capital requirements in the Basel II accord. By relaxing the constant correlation in the RFL model, we suggest that the conditional factor loading plays a significant role in capturing an asymmetric systematic impact from the market. The RR model uses the method proposed by Amraoui et al. (2012) to investigate the effects of the stochastic recovery rate. It allows the LGD function to be driven by the common factor and the hazard rate, while maintaining constant factor loadings. In the RRFL model, we incorporate the conditional factor loading into the state-dependent recovery rate and model them by sharing the unique common factor. To evaluate whether these two specifications significantly improve the default prediction, we use the dataset of daily stock indices of the S&P 500 to represent the market (common factor) and the respective stock prices of the defaulting companies for the period of five years before the default year from the Datastream database. In theory, stock returns should reflect the credit risk information of each firm, based on Merton (1974). Moreover, Xiang et al. (2015) document that strong evidence of time-varying credit risk links to equity markets.

Our default data analysis contains 2008 and 2009 data, as collected by Moody's report. We use Moody's Ultimate Recovery Database (URD), which is the ultimate payoff that obligors can obtain when the defaulting company emerges from bankruptcy or is liquidated rather than the post-default trading price that is proposed by Carty et al. (1998). These authors examine whether the trading price represents a rational forecast of actual recovery and find that it does not. For this period, we employ a state-dependent concept to capture the asymmetric impact from the common risk factor. As a result, both conditional factor loading and state-dependent recovery rates improve the calibration of our default prediction. The conventional factor copula underestimates the impact of systematic risk and portfolio credit loss when the market is in a downturn. We find that incorporating factor loading into the state-dependent recovery rate improves the accuracy of the default

prediction. This result is consistent with the goal of Basel III, which emphasizes the role of systematic risk on overall financial stability and default risk. In our later empirical analysis, we concentrate on senior unsecured bonds because there is a rich data source available.

The remainder of the study is organized as follows. Section 2 describes the goal of Basel III. We present a general framework and the standard one-factor Copula in Sect. 3. Furthermore, we extend the standard one-factor Copula using conditional factor loading and the state-dependent recovery model. Section 4 describes the dataset. In Sect. 5, we offer empirical evidence. Section 6 presents our conclusions.

2 Systematic risk in Basel III

As highlighted by Basel III, several aspects of systemic risk are crucial to the financial markets. First, a bank can trigger a shock throughout a system, and the shock can spill over to its counterparties (Drehmann and Tarashev 2013). Second, procyclicality can also destabilize all the systemic risk. Borrowers cannot offer more funding, as their collateral assets have depreciated due to weak economic conditions. Third, as Basel II focused on minimizing the default probability of individuals, this accord failed to guarantee a stable financial system due to its inattention to systemic risk. The new Basel accord is thus expected to emphasize the role of systemic risk.

The systematic factor is an important driver of systemic risk and likely constitutes a serious threat to systemic fragility (Schwerter 2011; Uhde and Michalak 2010). Tarashev et al. (2010) also distinguish between systemic risk and systematic risk. The former refers to the risk that impedes the financial system, whereas the latter refers to the commonality in the risk exposures of financial institutions. Their model assumes that systemic risk can have systematic and idiosyncratic components. Systemic risk is understandably heightened by systematic risk. A bank is characterized as a systemically important (too-big-to-fail) financial institution; its default would lead to a dramatic impact on systemic risk. This is the very outcome that Basel III attempts to regulate and prevent. In our paper, our model proposes that the contribution of systematic risk is higher than that of the idiosyncratic component and that this dominance is characterized by a higher factor loading on systematic risk during a market downturn. We therefore see that the contribution of systematic risk to credit risk varies with time and market conditions. In this regard, one concern is the interconnection between credit risk and market risk. Notably – and importantly –the points discussed above determine the sufficiency of capital requirements in the banking industry.

To obtain sufficient capital requirements, the recovery rate is one of the determinant variables in the credit risk estimation. Thus, in a recession period, recovery rates tend to decrease while default rates tend to rise. As such, increasing capital requirements under this condition seems advisable. Most early academic studies on credit risk assume that recovery rates are deterministic (Schönbucher 2001; Rosen and Saunders 2010), or they are stochastic but independent of default probabilities (Jarrow et al. 1997; Andersen and Sidenius 2004). Neglecting the stochastic nature of the recovery rate and the interdependence between recovery rates and default rates results in a biased credit risk estimation (Altman et al. 2005).

To adhere to the spirit of Basel III, our study extends the previous literature in two ways. First, we highlight that systematic risk is a predominant factor in a recession period and provide an analysis that measures the proportional contribution of systematic risk against that of an idiosyncratic component. Second, we propose a methodology in which

recovery rates and default rates are correlated by sharing a unique factor, both of which are state-dependent. Our model design, model simulation and empirical results offer several justifications for the goals of Basel III.

3 Methodology

3.1 Default modeling

Recognizing the importance of systematic risk, one-factor Gaussian models have been considered an important tool underlying the internal ratings-based approach (Crouhy et al. 2000; Frey and McNeil 2003) and are thus used to price CDOs (Andersen and Sidenius 2004; Hull and White 2004; Choroś-Tomczyk et al. 2013). These one-factor models reduce the number of correlations estimated from $\frac{N(N-1)}{2}$ in a multivariate Gaussian Model to N, which represents the number of assets. Specifically, we use a non-standardized Gaussian model to represent the deteriorating market condition by presuming a negative mean value together with a higher volatility. The model is based on decomposing a latent variable U_i for obligor i into systematic factor Z and idiosyncratic component ε_i:

$$U_i = \alpha_i Z + \sqrt{1-\alpha_i^2}\varepsilon_i \quad i = 1, \ldots, N \tag{1}$$

where $-1 \leq \alpha_i \leq 1$. Suppose that $Z \sim \mathrm{N}(\mu,\ \sigma^2)$ and ε_i have zero-mean unit variance distributions. In a Gaussian context, Z and ε_i are orthogonal and ε_i is mutually uncorrelated. In an empirical study, U_i is a proxy of respective stock return, which is systematically related to a common factor, Z (Choi and Jen 1991). The distribution of vector U can be described by a copula function that joins two marginals, Z and ε_i. The correlation coefficient ρ_{ij} between U_i and U_j can be described by their α_i and α_j:

$$\rho_{ij} = \frac{\alpha_i \alpha_j \sigma^2}{\sqrt{\alpha_i^2(\sigma^2-1)+1}\sqrt{\alpha_j^2(\sigma^2-1)+1}} \tag{2}$$

where $\sigma_i = \sqrt{\alpha_i^2(\sigma^2-1)+1}, \sigma_j = \sqrt{\alpha_j^2(\sigma^2-1)+1}$. As a consequence, the number of correlations describing the dependency structure is smaller because only N parameters α_i: $i = 1, \ldots, N$ must be estimated. We express the covariance matrices between U_i and U_j using a factor model,

$$\sum_{ij} = \sigma_i^2 \sigma_i^2 \begin{pmatrix} 1 & \rho_{ij} \\ \rho_{ji} & 1 \end{pmatrix} \tag{3}$$

The one-factor Gaussian copula model we consider is used to model the default indicators to time t, $I\{\tau_i \leq t\}$, by projecting U_i into τ_i. U_i here can be viewed as the proxies for a firm's asset and liquidation value (Andersen and Sidenius 2004). In this regard, the lower asset value of the firm is, the shorter the time to default, τ_i. More precisely, $U_i \leq F^{-1}\{P_i(t)\}$ leads to $\tau_i \leq t$, where $P_i(t)$ is a hazard rate and marginal probability that obligor i defaults before t, and $F^{-1}(\cdot)$ donates the inverse cdf of any distribution. The default indicator then can be written as

$$\mathrm{I}\{\tau_i \leq t\} = \mathrm{I}\left[U_i \leq F^{-1}\{P_i(t)\}\right] \tag{4}$$

Given the LGD for each i, $G_i, i = 1, \ldots, N$, we aggregate them as total portfolio loss, L, as follows:

$$L = \sum_{i=1}^{N} G_i \mathrm{I}\{\tau_i \leq t\} = \sum_{i=1}^{N} G_i \mathrm{I}\big[U_i \leq F^{-1}\{P_i(t)\}\big]. \tag{5}$$

3.2 Conditional default model

In accordance with the spirit of Basel III, the systematic latent factor, Z, representing the general economic condition that characterizes the systematic credit risk, influences the default probability $P_i(t)$ and the recovery rate $R_i = 1 - G_i$. Given Z, the conditional default probability may be written as $P_i(Z|S = H, Q)$ and conditional LGD, $G_i(Z|S = H, Q)$, as a function of Z, and it is state-dependent, $S \in \{H, Q\}$. H and Q represent the hectic and quiet periods, respectively.

A higher factor loading, α_i in Eq. (1) has been observed during hectic periods (Longin and Solnik 2001; Ang and Bekaert 2002; Ang and Chen 2002). This observation can be modeled by a regime-switching mechanism, requiring a globally valid time series structure for α_i from t. Avoiding such a structure that may be too rigid, we assume the two asset returns, Z (the common factor proxied by USD S&P 500) and U_i (firm stock price), to have a mixture of bivariate normal distribution (see "Appendix 1") to obtain the estimation of α_i^H and α_i^Q. Given the conditional factor loading, α_i^H,α_i^Q, the conditional default model is defined as follows:

$$U_i|_{S=H} = \alpha_i^H Z + \sqrt{1 - (\alpha_i^H)^2}\varepsilon_i \tag{6}$$

$$U_i|_{S=Q} = \alpha_i^Q Z + \sqrt{1 - (\alpha_i^Q)^2}\varepsilon_i \tag{7}$$

By employing the one-factor Gaussian copula, the state-dependent conditional default probability can be denoted by

$$P(\tau_i \leq t|S) = \Phi\left[\frac{\Phi^{-1}\{P_i(t)\} - \alpha_i^S Z}{\sqrt{1 - (\alpha_i^S)^2}}\right] = P_i(Z|S) \quad S \in \{H, Q\} \tag{8}$$

where $\Phi(\cdot)$ denotes Gaussian distribution. Given $P_i(t)$, if the factor loadings in hectic periods are greater than those during quiet times, say $\alpha_i^H > \alpha_i^Q$, and if the index return of S&P 500 is negative in a bad market condition, both conditions will result in a higher conditional default probability in Eq. (8). From Eq. (8), the systematic risk, Z, and the corresponding factor loading govern the conditional default probability, which is consistent with empirical findings (Andersen and Sidenius 2004; Bonti et al. 2006). Notably, α_i^S is state-dependent instead of a constant setting in the previous literature (Andersen and Sidenius 2004; Amraoui et al. 2012). Ang and Chen (2002) set the probability of both regimes equally ($\omega = 0.5$); however, we instead estimate it from the historical data of the S&P 500 Index return proxied for systematic risk, Z, P(S = H) = ω, P(S = Q) = 1 − ω using expectation–maximization (EM) algorithm.

Likewise, recovery rates can be designed in this manner by incorporating market conditions as the main driver across different states. Based on Das and Hanouna (2009),

recovery rates are negatively correlated with probabilities of defaults and are also driven by market conditions. By relaxing constant recovery rates, we follow Amraoui et al. (2012) and connect recovery rates and default events via a common factor, but we extend their model to a conditional or state-dependent framework. The recovery rate is governed by the state of the economy; in addition, we incorporate a conditional correlation structure, α_i^S, into the stochastic recovery rate model, and set $R_i(Z|S=H,Q)$ of obligor i, in relation to the common factor Z and the marginal default probability P_i. The state-dependent LGD is expressed as

$$G_i(Z|\mathrm{S}=\mathrm{H}) = (1-\bar{R}_i)\frac{\Phi\left[\left\{\Phi^{-1}(\bar{P}_i)-\alpha_i^H Z\right\}\Big/\sqrt{1-(\alpha_i^H)^2}\right]}{\Phi\left[\left\{\Phi^{-1}(P_i)-\alpha_i^H Z\right\}\Big/\sqrt{1-(\alpha_i^H)^2}\right]} \tag{9}$$

$$G_i(Z|\mathrm{S}=\mathrm{Q}) = (1-\bar{R}_i)\frac{\Phi\left[\left\{\Phi^{-1}(\bar{P}_i)-\alpha_i^Q Z\right\}\Big/\sqrt{1-(\alpha_i^Q)^2}\right]}{\Phi\left[\left\{\Phi^{-1}(P_i)-\alpha_i^Q Z\right\}\Big/\sqrt{1-(\alpha_i^Q)^2}\right]} \tag{10}$$

In Eqs. (9, 10), $0 \le \bar{R}_i \le R_i \le 1$ indicates a downward shift of $\bar{R}_i$ to R_i, such that $\bar{R}_i = R_i - v$ and $R_i \ge v > 0$. v is the size of the downward shift. By assuming that the expected loss in name i remains unchanged, we set $(1-R_i)P_i = (1-\bar{R}_i)\bar{P}_i$. Please see the proof in A.1 in Amraoui et al. (2012). $\Phi(\cdot)$ denotes a Gaussian distribution and $\bar{P}_i$ is the adjusted default probability calibrated proposed by Amraoui and Hitier (2008). The LGD function, $G_i(Z|S=H,Q)$, can essentially be obtained under formula (9,10). Numerous studies show that recoveries decline during recessions (Altman et al. 2005; Bruche and González-Aguado 2010). Consistent with the spirit of Eq. (6, 7), we design α_i^H,α_i^Q, and the factor loadings in Eq. (9,10) are therefore conditional and state-dependent. $\bar{R}_i$ is a lower bound for $G_i(Z|S=H,Q)$. Moreover, a partial derivative of the LGD function with respect to Z is less than zero, as shown by property 3.2 in Amraoui et al. (2012), which means that $G_i(Z|S=H,Q)$ is decreasing in Z. Assuming $\alpha_i^H > \alpha_i^Q$ means that a higher factor loading that is typically accompanied by a bad market condition on Z tends to increase LGD. In this regard, "Appendix 2" can be referenced for greater detail. The magnitude of LGD is not only influenced by Z but also sensitive to the factor loading under Z, which is one of our main findings and contributions to the literature. In addition, recovery rates are also linked to the probability of default and are negatively correlated (see Altman et al. 2005; Khieu et al. 2012). With Z, P_i and the estimated conditional factor loading α_i^H,α_i^Q, we obtain the state-dependent recovery rate, $R_i(Z|S=H,Q)$, and state-dependent LGD, $G_i(Z|S=H,Q) = 1 - R_i(Z|S=H,Q)$.

With these two specifications, the conditional default probability $P_i(Z|S=H,Q)$ and conditional LGD, $G_i(Z|S=H,Q)$, conditional expected loss is

$$\mathrm{E}(L_i|\mathrm{Z}) = \omega G_i(Z|\mathrm{S}=\mathrm{H})P_i(Z|\mathrm{S}=\mathrm{H}) + (1-\omega)G_i(Z|\mathrm{S}=\mathrm{Q})P_i(Z|\mathrm{S}=\mathrm{Q}) \tag{11}$$

where $\omega = \mathrm{P}(\mathrm{S}=\mathrm{H})$, 1-$\omega = \mathrm{P}(\mathrm{S}=\mathrm{Q})$. H and Q represent the hectic and quiet periods, respectively. In this paper, by employing the one-factor Gaussian copula model, Eq. (11) is written as

$$\mathrm{E}(L_i|Z) = \omega \mathrm{E}(L_i|Z_{S=H}) + (1-\omega)\mathrm{E}(L_i|Z_{S=Q}) \tag{12}$$

The detail of proof is set forth in "Appendix 3".

3.3 Monte Carlo simulation

In this section, we investigate default prediction performance by establishing a simulation of realistic scenarios. The default probability and recovery rate functions are governed by systematic factors produced by different regimes. Indeed, they are crucial elements in evaluating the accuracy of the default prediction. Our interest is to see whether the designs of conditional factor loadings and state-dependent recovery rates contribute to the default prediction.

3.3.1 One-factor non-standardized Gaussian copula

We simulate a one-factor non-standardized Gaussian copula subject to different states. As described in Eqs. (6) and (7), we generate systematic factor Z by non-standardized Gaussian distribution with different volatilities and independent ε_i' s to reflect the nature of distinct variations exhibited in different market conditions.

Through a mixed bivariate distribution setting in "Appendix 1", the conditional factor loadings, α_i^H and α_i^Q are derived, in the one-factor non-standardized Gaussian copula model. We estimate them from the daily stock returns of the S&P500 and of collected default companies during the crisis (2008–2009) period. The 3-year period prior to the crisis period is the estimation period for the conditional factor loadings. The return of the S&P 500 Index represented as a systematic factor, Z, is presumed to distribute as $\mathrm{N}(-0.03, 3.05)$ estimated in 2008 and 2009, while $\varepsilon_i \sim \mathrm{N}(0,1)$ represents idiosyncratic risk. Z and ε_i generated 10,000 scenarios. Given any of the generated systematic factor random variables, Z, and using Bayes' rule, we calculate the conditional probability that date t belonged to the hectic is $\pi(Z = z)$ using its counterpart, unconditional probability ω, as a formula (13).

$$P(S = H|\mathrm{Z} = \mathrm{z}) = \pi(Z = z) = \frac{\omega\varphi\left(z|\theta^H\right)}{(1-\omega)\varphi\left(z|\theta^Q\right) + \omega\varphi\left(z|\theta^H\right)} \tag{13}$$

where φ^H, θ^Q represent in the hectic (H) and the quiet (Q) periods. $\varphi(\cdot)$ is a normal distribution. Plugging α_i^H, α_i^Q shared with the same simulated Z random variables, conditional $U_i|\mathrm{S}$ is generated as developed in Eqs. (6, 7). These simulated random variables together with the published hazard rates $P_i\,(t)$ ideally produce the simulated default times.

3.3.2 Default time

Projecting U_i simulated from Sect. 3.3.1 to default time, τ_i, as stated in Eq. (4), provides a clue as to whether the firm defaults before time. We set $t = 1$, which represents the time interval of 1 year, so that $\tau_i < 1$ is referred to as a default event in the ith obligor. The hazard rate P_i is the probability of occurrence of the default event within one year. τ_i represents the default time of the ith obligor. More precisely, the expected value of $\mathrm{I}(\tau_i < 1)$ is $\mathrm{P}\,(\tau_i < 1)$ and referred to as P_i, see Franke et al. (2011) Chapter 22, which can be connected to the firm's stock return or firm's value, and U_i leads to $P_i = \mathrm{E}[\mathrm{I}\{U_i < \Phi_i^{-1}(P_i)\}]$, where Φ_i denote the Gaussian cdf of U_i. By applying generated U_i from the conditional factor model into the definition of the survival rate, we have

generated the default time, τ_i, derived from $1 - \exp(-P_i\tau_i) = \Phi(U_i)$ (Hull 2006). To remain in the state-dependent environment, the conditional default time for each obligor is generated by formula (14).

$$\tau_i|S = -\frac{\log\{1 - \Phi(U_i|_S)\}}{P_i} \tag{14}$$

where P_i is the hazard rate or marginal probability that obligor i will default during the first year, conditional on no earlier default, and is obtained from Moody's. It is the cumulative of the default rates during the first year. Equation (14) states that as $U_i|_S$ becomes larger, $\tau_i|$S will become longer. The larger U_i reduces the tendency of default and postpones the default time, $\tau_i|$S.

3.3.3 State-dependent recovery rate simulation

In the third step, we consider a more realistic situation by simulating recovery rates, as described in our settings. The adjusted default probability $\bar{P}_i$ is calibrated using hazard rate P_i from Moody's report. $\bar{R}_i$ is a lower bound for the state-dependent recovery rate [0,1]; therefore, we set $\bar{R}_i = 0$ in the simplest case. With α_i^H,α_i^Q, Z, $\bar{P}_i$, the simulated state-dependent recovery rates are obtained using formula (9, 10).

3.3.4 Loss function

By changing scenarios to quiet and hectic states, we assume the exposure of each obligor is 100 million and calculate the expected loss under the given scenarios corresponding to formula (11).

$$\mathrm{E}(L_i|\mathrm{Z}) = \pi(Z = z)G_i(Z|\mathrm{S} = \mathrm{H})P_i(Z|\mathrm{S} = \mathrm{H}) + (1 - \pi(Z = z))G_i(Z|\mathrm{S} = \mathrm{Q})P_i(Z|\mathrm{S} = \mathrm{Q}) \tag{15}$$

Given the simulated Z random variables, conditional probability $\pi(Z = z)$ naturally provides better information than unconditional probability ω does. By the given formula (15), we compare the theoretical loss amounts across four models with the realized loss values, and evaluate the performance of the default prediction by the mean of square error.

3.3.5 Absolute error

In step 5, the performance of the competing models (FC, RFC, RR, and RRFC) are evaluated to decide which is the best at predicting the default for the following year. Absolute Error (AE) here is linked to prediction performance and is defined as

$$\mathrm{AE} = (\text{actual portfolio loss} - \text{expected portfolio loss}) \tag{16}$$

where the actual portfolio loss is from Moody's. Expected loss is estimated from Eq. (15), although in an unconditional default model, it is computed from formula (5). For each competing model, we generate 10,000 scenarios; then, the mean of the absolute error (referred to as MAE) is calculated. It can be expected that the best one is also included in the minimum AE and MAE.

4 Data

4.1 Financial return data

In this section, we illustrate how to proceed with the financial data. Weiß (2013) proposes the GARCH(1,1) model to describe marginal time-varying volatility in the presence of conditional heteroskedasticity in financial returns. Following Krupskii and Joe (2013), we use the S&P 500 Index and obligors' stock returns following the AR(1)-GARCH(1,1). The model is written as follows:

$$r_{jt} = \mu_j + k_j r_{j,t-1} + \delta_{jt}\epsilon_{jt}$$

$$\delta_{jt}^2 = c_j + a_j r_{j,t-1}^2 + b_j \delta_{j,t-1}^2$$

where r_{jt} are returns and ϵ_{jt} are i.i.d. vectors with Gaussian distribution. By applying the Gaussian copula, the parameters are computed from the GARCH filtered data.

4.2 Data description

We use the list of default companies for 2008 through 2009 published by Moody's annual report since this is a rich source of available data. In total, we obtained 341 defaults with corporate bond recovery rates from Moody's URD covering the period from 1987 to 2007. We focus on senior unsecured bonds because of their wide use in financial contracts, regulatory rules, and the risks associated with measuring for assets under the standardized approach of Basel II (Pagratis and Stringa 2009). We also collected the credit rating of obligors from Moody's to measure the hazard rate. Although there are 94 and 247 defaulting firms in 2008 and 2009, the observations were reduced due to missing stock prices and credit ratings of obligors' bonds. If there were insufficient reported stock prices of defaulting subsidiary companies, we used the stock prices of parent companies instead. In all cases, 31 and 64 sampling firms were collected in 2008 and 2009, respectively.

To estimate the conditional factor loadings of sampled firms, we collect the daily USD S&P 500 return and the respective stock returns of the defaulting companies for a 3-year period prior to the default year from the Datastream database. The USD S&P 500 Index here simply represents common systematic risk. By assuming a mixture of bivariate normal distribution, we estimate the parameters, including factor loadings by EM algorithm. Table 2 presents the results of the EM algorithm.

Table 2 Estimate mixture of normal distribution by employing an EM algorithm

Model	Probability	Mean	STD
Period	2003–2007		
Unconditional	100.00%	−0.01	0.99
Conditional on quiet	21.97%	0.09	0.24
Conditional on hectic	78.03%	−0.03	1.12
Period	2004–2008		
Unconditional	100.00%	0.04	0.99
Conditional on quiet	24.91%	0.19	0.26
Conditional on hectic	75.09%	−0.01	1.14

STD standard deviation

As presented in Table 2, the volatility of the hectic distribution is larger than that of the quiet distribution, and the mean of the hectic distribution is smaller than that of the quiet distribution, reflecting the fat tails and right skew that are consistent with Kim and Finger (2000).

5 Empirical result

5.1 Conditional factor loading estimation

Figures 1 and 2 show that most of the correlation coefficients or factor loadings in the factor copula model during the hectic period are higher than in the quiet period. The proposed correlation structure leads to more accurate and realistic implementations and avoids the underestimation of factor loading in a hectic period or the overestimation in a quiet period. These ideas are well known in statistics and have already been applied to financial questions (Ang and Chen 2002; Patton 2004).

In our approach, we consider this asymmetric correlation structure under real market conditions to implement the conditional default model developed in Sect. 3.2. As shown in

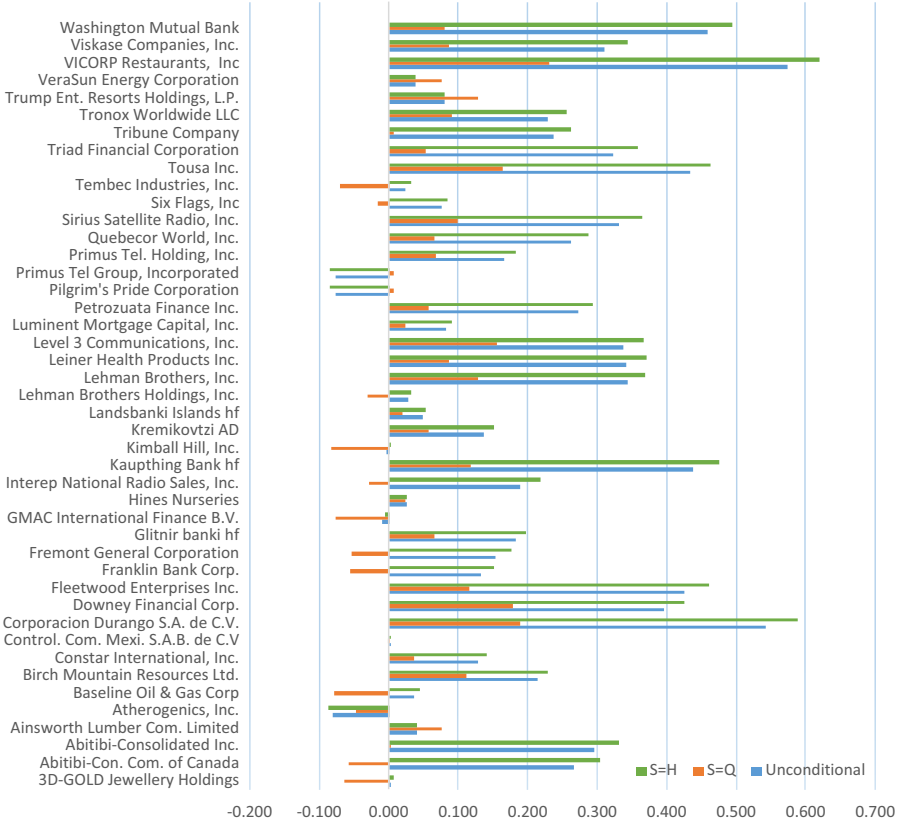


Fig. 1 Conditional and unconditional factor loading comparison in 2008. The estimation of conditional and unconditional factor loading between S&P 500 and default companies in 2008

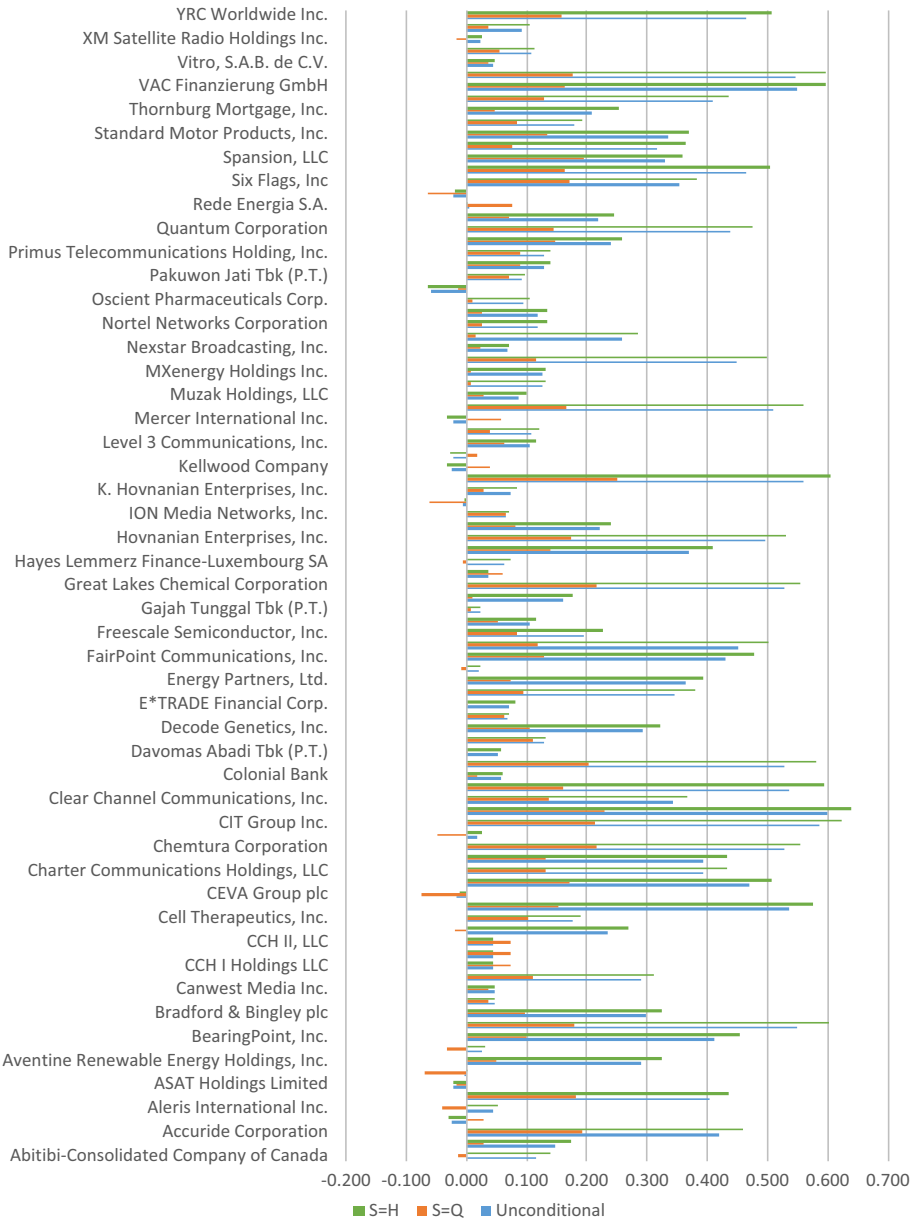


Fig. 2 Conditional and unconditional factor loading comparison in 2009. The estimation of conditional and unconditional factor loading between S&P 500 and default companies in 2009

Figs. 1 and 2, the factor loadings α_i in state H are higher than those in state Q. As factor loadings become higher in state H, the correlation coefficient ρ_{ij} between firm i and j defined in Eq. (2) is expected to increase in this market condition. Therefore, obligors tend to co-move more closely during hectic periods than during quiet periods.

5.2 State-dependent recovery rate estimation

To demonstrate the impact of market conditions measured by Z on the state-dependent recovery rate, we use Fig. 3 to depict the relationship between the state-dependent recovery rate and the S&P 500 (the proxy for systematic factor Z) in blue '*', which developed in Sect. 3.2. It can be observed that as the effect of the systematic factor on the recovery rate is positive, the recovery rate gets higher as Z grows. Because the slope of this curve is influenced by estimated α_i^H,α_i^Q corresponds to formula (9, 10), the slopes behave differently in the four panels but stay monotonically positive. We also depict the stochastic recovery rates in red '+' estimated and simulated through the Amraoui et al. (2012) model, in comparison with blue '*', which is simulated in our model. Taking (c) E*TRADE as an example, compared with the simulated recovery rates based on Eqs. (9) and (10), we note those generated from Amraoui et al. (2012) by assuming constant factor loadings tend to produce higher recovery rates in the market downturn and lower rates in the booming market. This evidence suggests that the recovery rate may be overestimated in a bearish market but underestimated in a bullish market if constant factor loading is assumed. As a consequence, it is highly possible to underestimate credit loss in a bearish market and overestimate it in a bullish market. Similarly, the evidence from (a) Glitnir Banki (b) Lehman Brothers Holdings, Inc. and (d) Idearc, Inc. are comparable and consistent. Notably, the impact of the systematic factor on the recovery rate seems nonlinear, as it is higher in the market downturn but relatively mild in the booming market, and its marginal slope decreases abruptly when the index return decreases; however, the marginal slope decelerates when the index return becomes positive. This simulation result is in accordance with the Moody's report in Table 1. From 2004 to 2006, the annual recovery rates of senior unsecured bond increase slowly. As the crisis begins in August 2007, the recovery rate drops dramatically. By capturing the correlation structure, $\alpha_i^H > \alpha_i^Q$, as shown in (a), (b), (c) and (d), we find this asymmetric pattern, which is more consistent with reality.

With the simulated recovery rates from Eqs. (9, 10), we are more interested in the relation between it and conditional default probability from Eq. (8). As Fig. 4 shows, the simulation result shows the downward trend between default probability and the recovery rate, which is consistent with Altman et al. (2005) and Das and Hanouna (2009). Moreover, it shows that the common factor governs the default rate and recovery rate simultaneously and creates their negative association implicitly. Altman et al. (2005) find that permitting a dependence between default rates and recovery rates yields approximately 29% in the value at risk compared with a model that assumes no dependence between default rates and recovery rates.

5.3 Empirical results of absolute errors

To gauge the conditional factor loading and state-dependent recovery rate approaches for default prediction, we propose four models: (1) the FC model, i.e., the standard one-factor Gaussian copula model with a constant recovery rate developed by Van der Voort (2007) and Rosen and Saunders (2010); (2) the RFL model, i.e., the one-factor Gaussian copula model in which factor loadings are tied to the state of the common factor and the recoveries assumed as constant, as proposed by Kalemanova et al. (2007) and Chen et al. (2014); (3) the RR model, i.e., the standard one-factor Gaussian copula model but with the recoveries related to the macroeconomic state (Amraoui and Hitier 2008; Elouerkhaoui 2009; Amraoui et al. 2012); and (4) the RRFL model, i.e., a conditional factor loading

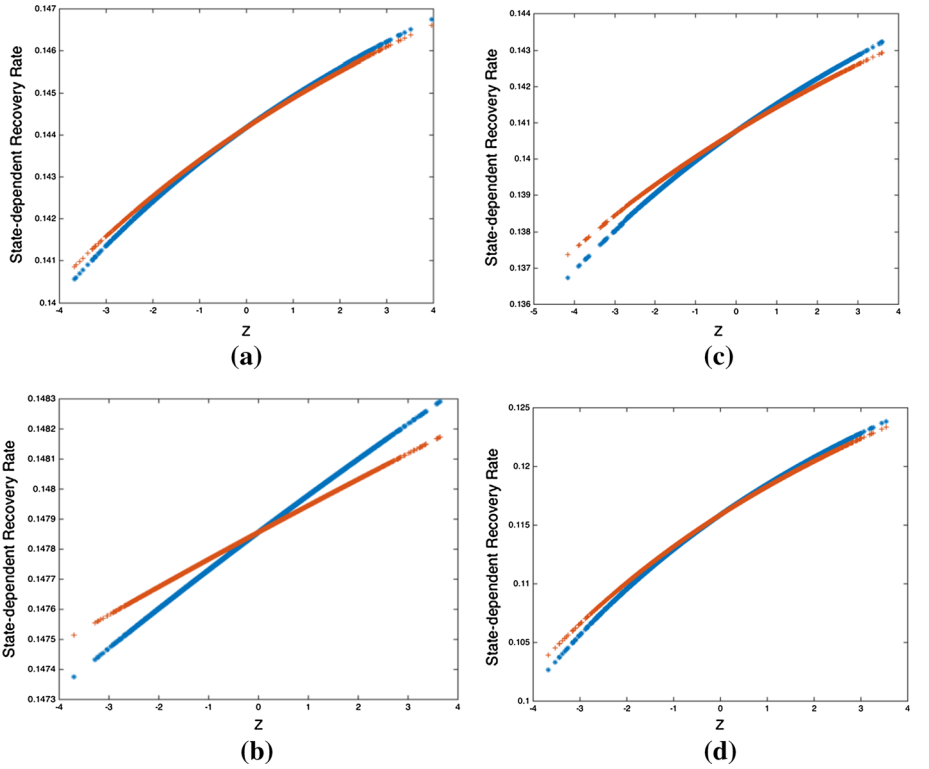


Fig. 3 The relationship between state-dependent recovery rate and index return of S&P 500, *Z*. *Panel* **a** and **b**, '*' in *blue* illustrates the pattern of state-dependent recovery rate of Glitnir banki and Lehman Brothers Holdings, Inc. which incorporate conditional factor loading in 2008. '+' in *red plots* the recoveries proposed by Amraoui et al. (2012). In *panel* **c** and **d**, E*TRADE Financial Corp. and Idearc, Inc. in 2009. **a** Glitnir Banki: $\alpha = 0.183$, $\alpha^Q = 0.066$, $\alpha^H = 0.196$, **b** Lehman Bro: $\alpha = 0.345$, $\alpha^Q = 0.128$, $\alpha^H = 0.370$, **c** E*TRADE: $\alpha = 0.071$, $\alpha^Q = 0.002$, $\alpha^H = 0.082$, **d** Idearc, Inc.: $\alpha = 0.222$, $\alpha^Q = 0.082$, $\alpha^H = 0.239$ (Color figure online)

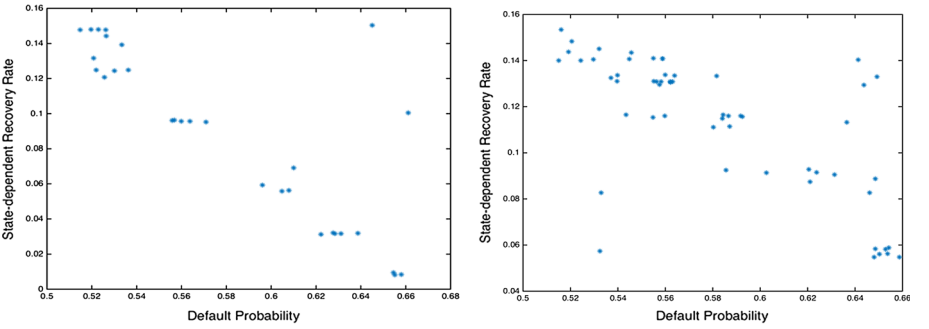


Fig. 4 The relationship between state-dependent recovery rates and default probabilities. By simulating $Z \sim \mathrm{N}(-0.03, 3.05)$, it plots the relationship between the state-dependent recovery rate and default probabilities, given the conditional factor loading. By simulating 10,000 observations, we estimate the default probabilities and state-dependent recovery rate from formula (8) and (9,10). **a** 2008, **b** 2009

specification together with a state-dependent recovery rate. We address the question of whether the two specifications, conditional factor loading and the state-dependent recovery rate model, are meaningful and significant in explaining the gap between expected and actual loss value. To check the predictive ability of the different models, we report the AE and MAE estimated from Sect. 3.3.5.

Table 3 reports the AE between actual portfolio loss and expected portfolio loss constructed by 31 and 64 observations in 2008 and 2009, respectively. A comparison of the four models shows that the estimate of expected portfolio loss in the RRFL model is highest and closest to the corresponding actual loss, which means that the expected portfolio losses may be underestimated by the other three models. In particular, modeling a recovery rate in a stochastic fashion indeed contributes to difficulties in estimating downgrades in credit.

We compare the four competing models of each obligor and choose the best model for achieving the minimum AE and MAE. We find that including the conditional factor loading (RFL model) instead of the Pearson correlation (FC model) does not significantly improve the estimations in 2008 and 2009. Table 3 shows that introducing the state-dependent recovery rate (RR model) leads to a promising improvement over the standard model the (FC model). We interpret this to mean that the setting of a stochastic recovery rate seems necessary, which brings a remarkable improvement to the default prediction, which is consistent with Altman et al. (2005) and Ferreira and Laux (2007). Compared with the RR model, the RRFL model includes conditional factor loading in default

Table 3 The mean of actual portfolio loss, expected portfolio loss and AE, MAE (in million)

	FC	RFL	RR	RRFL
2008				
Actual portfolio loss	2035.02	2035.02	2035.02	2035.02
Expected portfolio loss	1070.57	1085.67	1537.46	1567.66
AE	964.45	949.35	497.56	467.36
MAE	31.11	30.62	16.05	15.08
Expected portfolio loss/actual portfolio loss (%)	52.61	53.35	75.55	77.03
2009				
Actual portfolio loss	3853.10	3853.10	3853.10	3853.10
Expected portfolio loss	2033.25	2064.47	3318.25	3380.69
AE	1819.85	1788.63	534.85	472.41
MAE	28.43	27.95	8.36	7.38
Expected portfolio loss/actual portfolio loss (%)	52.77	53.58	86.12	87.74

This table reports the AE and MAE by comparing the four models: (1) The FC model, i.e., the standard one-factor Gaussian copula model with is a constant recovery rate; (2) the RFL model, i.e., the one-factor Gaussian copula model in which the factor loadings are tied to the state of the common factor and the recoveries assumed to be constant; (3) the RR model, i.e., the standard one-factor Gaussian copula model in which the recoveries are related to the state of the macroeconomic state; and (4) the RRFL model, i.e., a conditional factor loading specification together with a state-dependent recovery rate. This table also presents the difference between actual portfolio loss and expected portfolio loss, which is referred to as AE; when AE is divided by 31 and 64 observations in 2008 and 2009, respectively, it becomes MAE. The percentage represents expected portfolio loss divided by the actual portfolio loss

probabilities and a state-dependent recovery rates function and produces considerably more modest improvements.

We propose two specifications on factor loading and recovery rates across four models. If we assume that default probabilities are a function of two-state correlation constructs but that recovery rates are not, the specification is only identified as concentrated on factor loading. In this case, the recovery rates do not contain information about the state of the business cycle. Conversely, if we assume that recovery rates vary, but factor loading is fixed, then the refinement occurs only by means of variations in the recovery rate. Since the RRFL model with both specifications is superior to the other three competing models, and there is no redundant specification in this study. In this regard, we extend the models proposed by prior studies (Kalemanova et al. 2007; Van der Voort 2007; Amraoui and Hitier 2008; Elouerkhaoui 2009; Amraoui et al. 2012; Rosen and Saunders 2010; Chen et al. 2014), which leads to more accurate default predictions in one year.

5.4 Basel III: relative contribution

Systematic risk has been considered one of the main causes of the 2007–2009 crisis, and Basel III is proposed to control systematic risk (one systemic risk measure) to achieve the goal of overall financial stability. In this section, we highlight the role of systematic risk and its impact on the goals of Basel III. The aim of relative contribution analysis is to investigate the proportional contribution from systematic risk in comparison with that from the idiosyncratic component. By measuring systematic risk, $\alpha_i^S Z$, and idiosyncratic risk, $\sqrt{1-(\alpha_i^S)^2}\varepsilon_i$, $S \in \{H, Q\}$ from formula (6, 7), we depict a scatter plot for simulated systematic risk (horizontal axis) and idiosyncratic risk (vertical axis) in Fig. 5. As shown in the 2D plot for 2008, the 45° line represents the proportion of systematic risk that is equal to that of idiosyncratic risk. If the scatter points are located in the 'A, B, C, D' zones, the contribution of systematic risk to default risk is greater than that of idiosyncratic risk. On the other hand, if the scatter points are settled in the 'a, b, c, d' areas, the contribution of the systematic component is less than that of idiosyncratic risk. For example, the effect of systematic risk on default risk will become larger when point 'Y' moves to point 'X'. Most studies focus on either systematic (King and Khang 2005; Uhde and Michalak 2010) or firm-specific components (Goyal and Santa-Clara 2003; Ferreira and Laux 2007), and a limited number of studies compare the influence of both of them.

By simulating $Z \sim \mathrm{N}(-0.03, 3.05)$, each simulated Z random variable can therefore be mapped into a specific conditional probability of being in a hectic state in Eq. (13). We gather the scatter plots into three groups here. The first group (marked as '+' in red) includes only the simulated Z r.v. with projecting conditional probabilities above the 75% quartile, and indicates that they are generated in distress. The second group (marked as '*' in blue) includes the Z r.v. with projecting conditional probabilities below the 25% quartile to indicate that they are generated in a bullish atmosphere. The third group (marked as 'x' in yellow) collects the rest. With regard to the tranquil scenarios ('blue' points) in 2008, most observations were located in the area in which the relative contribution of idiosyncratic risk is larger than that of the economy-wide component, where credit risk was mainly driven by the idiosyncratic component before the subprime crisis, as reported in Rodríguez-Moreno and Peña (2013), who found that idiosyncratic components were larger than systematic risk before the subprime crisis and were extracted from the CDX-IG-5y using high-frequent measures. At the beginning of the financial crisis, systematic risk skyrocketed. Intuitively, systematic risk increases sharply due to the larger factor loadings

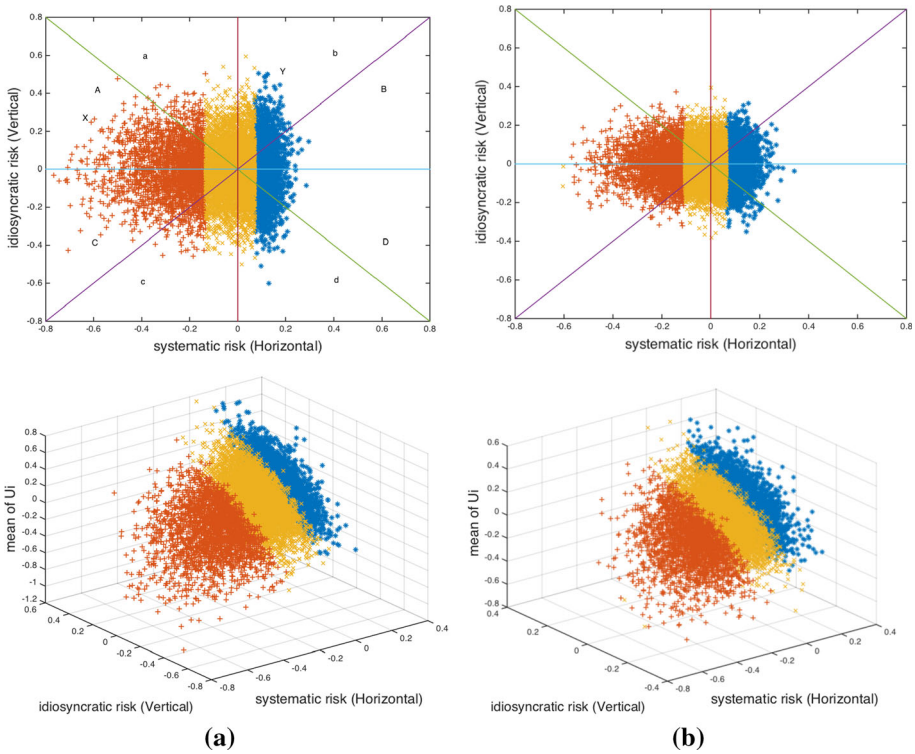


Fig. 5 The 2D and 3D scatters plot of relative contribution. By simulating $Z \sim \mathrm{N}(-0.03, 3.05)$, the 2D graphic illustrates the relationship between the mean of systematic risk, $\alpha_i^S Z$, and idiosyncratic risk, $\sqrt{1-(\alpha_i^S)^2}\varepsilon_i$. Each simulated Z random variable can therefore be mapped into a specific conditional probability of being in a hectic state in Eq. (13). We depict the scatters in three groups here. The first group (marked as '+' in *red*) includes only the simulated Z r.v. with projecting conditional probabilities above the 75% quartile; it indicates that they are generated in distress. The second group (marked as '*' in *blue*) includes the Z r.v. with projecting conditional probabilities below the 25% quartile to indicate that they are generated in a bullish atmosphere. The third group (marked as 'x' in *yellow*) collects the rest. In the 3D plot, observations in hectic periods are marked in *red*. Quiet days are marked in *blue*, otherwise in *yellow*. **a** 2008, **b** 2009 (Color figure online)

when the market is in hectic scenarios. Our result shows that systematic risk was higher than the idiosyncratic component in the hectic scenarios ('red' points) in 2008; in the quiet scenarios, however, firm-specific factors are more important at some points, as noted by Rodríguez-Moreno and Peña (2013). Similarly, it has been shown that the relative contribution of the systematic component explains a higher proportion of obligor asset value in 2009.

More visibly, the 3D plot identifies the relationship among the level of average $U_i|_S$, which is referred to as the mean of firms' value, systematic and idiosyncratic component. Each observation in Fig. 5 reflects its mean of $U_i|_S \quad i = 1, \ldots, N$ in each simulated day in 2008 and 2009, respectively. Figure 5 shows that the points in the hectic period marked as red '+' indicates a negative shock from systematic risk, which lowers the average asset value of obligors; specifically, most observations show the negative impact of systematic shock, which accounts for a substantially larger proportion of firms' value substantially.

Note that it is easy to drive the default event since it lowers the firms' value significantly. On the other hand, the points in quiet days marked as blue '*' indicate a positive shock from the systematic component. However, the negative shock from firm-specific factors may compromise the benefit from economy-wide components that lowers the level of average $U_i|_S$ at some points.

Our model emphasizes the importance of systematic risk, which explains most obligors' default behavior, particularly in hectic periods, which is one of the important features of Basel III (Tarashev et al. 2010; Uhde and Michalak 2010; Schwerter 2011). To be specific, we measure and demonstrate the contribution of overall systematic risk to each asset, and identify the impact direction from systematic and idiosyncratic risk. Moreover, this analysis can be applied to a variety of systematic risk measures. In this sense, portfolio managers should be aware of the systematic risk that can substantially influence the value of portfolios. We propose that the regulatory tool of Basel III could be estimated with such contributions. A related question is how these measures can aid policymakers. The measures in this paper can be used as a tool to prevent systematic crises, and our model can be used as an early warning system that will alert regulators when an individual bank is in trouble and to intervene before a crisis occurs.

5.5 Robustness test

Since Table 3 reports that the expected portfolio loss is far from the actual portfolio loss, we gauge that using bond credit rates as a measure of hazard rate has the disadvantage that they are released annually by Moody's. In this section, we use credit default swap (CDS) spread data as an alternative market-based measure of a company's credit risk. A CDS spread measures a financial swap agreement in which the seller will compensate the buyer in the event of a loan default. Basically, variation in the CDS spread reflects the dynamics of risk condition or hazard rate implicitly. The larger the CDS spread is, the riskier the debtor. Therefore, the hazard rate, $\bar{\kappa}$, for a company can be estimated by the following:

$$\bar{\kappa} = \frac{s}{1 - R}$$

where s is CDS spread. We consider the latest one-year prior to the default year CDS quotes of obligors provided from Datastream. We also use a credit spread, which is the yield on an annual par yield bond issued by the obligors over one-year LIBOR (London Interbank Offered Rate) if the obligor does not have CDS data. Theoretically, the CDS spread is close to the credit spread (Hull and White 2000; Hull et al. 2004). By plugging in the recovery rate, R, obtained from the Moody's report, we compute the average default intensity, $\bar{\kappa}$, per year conditional on no earlier default instead of P_i. Compared with P_i from the Moody's annual report, a CDS spread with active trading activity reflects the market assessments of default risk in a timely fashion. In this regard, the proposed models that incorporate the hazard rate implied in CDS spreads may yield a better prediction.

According to Table 4, the models with a hazard rate implied in a CDS spread seem to perform better than those with a hazard rate from historical bond credit rates. By comparing Tables 3 and 4, generally, a CDS spread as the hazard rate measure reflects more timely information than the bond credit rate does. Table 4 presents the results from a robustness test that shows that the RRFL model outperforms in a robustness test. In both tables, the RRFL model consistently outperforms, which produces the expected portfolio loss most closely to the actual portfolio loss.

Table 4 The actual portfolio loss, expected portfolio loss, AE, and MAE (in million) for robustness

	FC	RFL	RR	RRFL
2008				
Actual portfolio loss	1489.81	1489.81	1489.81	1489.81
Expected portfolio loss	920.68	930.11	1245.14	1258.17
AE	569.13	559.70	244.67	231.64
MAE	22.76	22.39	9.79	9.27
Expected portfolio loss/actual portfolio loss (%)	61.80	62.43	83.58	84.45
2009				
Actual portfolio loss	2707.30	2707.30	2707.30	2707.30
Expected portfolio loss	1776.77	1784.18	2381.91	2402.54
AE	930.52	923.11	325.39	304.76
MAE	22.16	21.98	7.75	7.26
Expected portfolio loss/actual portfolio loss (%)	65.63	65.90	87.98	88.74

This table reports the values of the AE and MAE of four models using the market-based method during 2008 and 2009. This table also shows the actual portfolio loss and the expected portfolio loss of 25 and 42 observations in 2008 and 2009. The percentage represents expected portfolio loss divided by actual portfolio loss

6 Conclusion

This paper proposes a refined factor copula model to assess and predict credit risk. On the basis of our estimated model, we find that systematic risk plays a simultaneously critical role in governing default rates and recovery rates simultaneously. Our simulation results show that recoveries vary with the returns of the S&P 500 and that the impact of systematic factors on the recovery rate is asymmetric by finding a higher factor loading in hectic periods than in tranquil periods. Among the various factor copula models developed in the past and in the current literature as the competing models, the model with conditional random factor loading and a state-dependent recovery rate turns out to be the best performing. In other words, our refined model contributes to studies that have been mapped to three groups of competing models (the FC, RFL, and RR models).

As a response to Basel III, we measure and demonstrate the contribution of overall systematic risk to each firm's value, and we also identify the relative roles of both systematic and idiosyncratic risk. Moreover, this analysis can be applied to a variety of systematic risk measures, and it aids regulators in preventing a systematic crisis. In addition, by investigating the effect of state-dependent recovery rates on the loss function, we suggest that banks should apply this capital requirement issue to ensure its sufficiency.

In further research, we plan to go beyond this study in several ways. First, other copula functions can be modeled to capture various dependence structures. Second, the marginal distribution can be considered in a more general way to capture a fat-tail feature. We will leave these issues for future studies.

Appendix 1: Conditional factor loading

We assume the two asset returns Z (USD S&P 500) and U_i (firm stock price) to have a mixture of bivariate normal distribution:

$$(Z, U_i) \sim \begin{cases} \Phi\left(\begin{bmatrix} \mu_Z^Q \\ \mu_i^Q \end{bmatrix}, \begin{bmatrix} \sigma_Z^Q \sigma_Z^Q & \sigma_Z^Q \alpha^Q \sigma_i^Q \\ \sigma_Z^Q \alpha^Q \sigma_i^Q & \sigma_i^Q \sigma_i^Q \end{bmatrix}\right) \\ \Phi\left(\begin{bmatrix} \mu_Z^H \\ \mu_i^H \end{bmatrix}, \begin{bmatrix} \sigma_Z^H \sigma_Z^H & \sigma_Z^H \alpha^H \sigma_i^H \\ \sigma_Z^H \alpha^H \sigma_i^H & \sigma_i^H \sigma_i^H \end{bmatrix}\right) \end{cases} \tag{17}$$

where volatility in hectic periods is higher than in quiet periods, $(\sigma_i^H)^2 > (\sigma_i^Q)^2$. α^Q and α^H are the correlation coefficients between each obligor and the S&P 500 in quiet and hectic period, as proposed by Kim and Finger (2000), respectively. We estimate the unknown parameters $\omega, \mu_Z^Q, \sigma_Z^Q, \mu_Z^H, \sigma_Z^H$ from the marginal distribution of Z:

$$\begin{cases} \Phi\left[\mu_Z^Q, \sigma_Z^Q \sigma_Z^Q\right] & \mathrm{P}(S = Q) = 1 - \omega \\ \Phi\left[\mu_Z^H, \sigma_Z^H \sigma_Z^H\right] & \mathrm{P}(S = H) = \omega \end{cases} \tag{18}$$

Appendix 2

The assumptions of the bounds on the LGD are $0 \le G_i(Z|S = H, Q) \le (1 - \bar{R}_i)$. Since $0 \le G_i(Z|S = H, Q)$, $\bar{P}_i \le P_i$ and $\alpha_i^H > \alpha_i^Q$,

$$0 < \Phi\left[\frac{\left\{\Phi^{-1}(\bar{P}_i) - \alpha_i^H Z\right\}}{\sqrt{1 - (\alpha_i^H)^2}}\right] \le \Phi\left[\frac{\left\{\Phi^{-1}(P_i) - \alpha_i^H Z\right\}}{\sqrt{1 - (\alpha_i^H)^2}}\right] \tag{19}$$

$$0 < \Phi\left[\frac{\left\{\Phi^{-1}(\bar{P}_i) - \alpha_i^Q Z\right\}}{\sqrt{1 - (\alpha_i^Q)^2}}\right] \le \Phi\left[\frac{\left\{\Phi^{-1}(P_i) - \alpha_i^Q Z\right\}}{\sqrt{1 - (\alpha_i^Q)^2}}\right] \tag{20}$$

which yields

$$G_i(Z|S = H) \le (1 - \bar{R}_i)$$
$$G_i(Z|S = Q) \le (1 - \bar{R}_i).$$

The bounds are strict, which is proven. Moreover, let us assume that $0 < P_i < 1$, $0 \le \bar{R}_i \le R_i \le 1$ and $\alpha_i^H > \alpha_i^Q$, then $G_i(Z|S = H, Q)$ decreases as Z decreases, which can be shown as plugging the $\alpha_i^H > \alpha_i^Q$ by following the proof of property 3.2 in Amraoui et al. (2012).

Appendix 3

We propose that the default probability is conditional on one-factor Z with two states, H (Hectic) and Q (Quiet), as shown in formula (8). By assuming that Z and U_i follow a Gaussian distribution and that the default time is within one year, we also derive the conditional expected loss,

$$
\begin{aligned}
\mathrm{E}(L_i|\mathrm{Z}) &= \mathrm{E}\{(1-R_i)1_{\tau_i<1}|Z\} = \omega G_i(Z|\mathrm{S}=\mathrm{H})P_i(Z|H) + (1-\omega)G_i(Z|\mathrm{S}=\mathrm{Q})P_i(Z|Q) \\
&= \omega(1-\bar{R}_i)\frac{\Phi\left[\frac{\left\{\Phi^{-1}(\bar{P}_i)-\alpha_i^H Z\right\}}{\sqrt{1-(\alpha_i^H)^2}}\right]}{\Phi\left[\frac{\left\{\Phi^{-1}(P_i)-\alpha_i^H Z\right\}}{\sqrt{1-(\alpha_i^H)^2}}\right]}P_i(Z|H) + (1-\omega)(1-\bar{R}_i)\frac{\Phi\left[\frac{\left\{\Phi^{-1}(\bar{P}_i)-\alpha_i^Q Z\right\}}{\sqrt{1-(\alpha_i^Q)^2}}\right]}{\Phi\left[\frac{\left\{\Phi^{-1}(P_i)-\alpha_i^Q Z\right\}}{\sqrt{1-(\alpha_i^Q)^2}}\right]}P_i(Z|Q) \\
&= \omega(1-\bar{R}_i)\frac{\Phi\left[\frac{\left\{\Phi^{-1}(\bar{P}_i)-\alpha_i^H Z\right\}}{\sqrt{1-(\alpha_i^H)^2}}\right]}{P_i(Z|H)}P_i(Z|H) + (1-\omega)(1-\bar{R}_i)\frac{\Phi\left[\frac{\left\{\Phi^{-1}(\bar{P}_i)-\alpha_i^Q Z\right\}}{\sqrt{1-(\alpha_i^Q)^2}}\right]}{P_i(Z|Q)}P_i(Z|Q) \\
&= \omega(1-\bar{R}_i)\Phi\left[\frac{\left\{\Phi^{-1}(\bar{P}_i)-\alpha_i^H Z\right\}}{\sqrt{1-(\alpha_i^H)^2}}\right] + (1-\omega)(1-\bar{R}_i)\Phi\left[\frac{\left\{\Phi^{-1}(\bar{P}_i)-\alpha_i^Q Z\right\}}{\sqrt{1-(\alpha_i^Q)^2}}\right] \\
&= \omega\mathrm{E}(L_i|Z_{S=H}) + (1-\omega)\mathrm{E}(L_i|Z_{S=Q})
\end{aligned}
$$

where $\omega = \mathrm{P}(\mathrm{S}=\mathrm{H})$, $1-\omega = \mathrm{P}(\mathrm{S}=\mathrm{Q})$. Φ indicates the Gaussian cumulative distribution. $\mathrm{E}(L_i|\mathrm{Z}) = (1-\bar{R}_i)\Phi\left[\frac{\left\{\Phi^{-1}(\bar{P}_i)-\alpha_i Z\right\}}{\sqrt{1-\alpha_i^2}}\right]$ has been proven in appendix C.2 of Amraoui et al. (2012). In the same vein, we derive $\mathrm{E}(L_i|Z_{S=H}) = (1-\bar{R}_i)\Phi\left[\frac{\left\{\Phi^{-1}(\bar{P}_i)-\alpha_i^H Z\right\}}{\sqrt{1-(\alpha_i^H)^2}}\right]$ and $\mathrm{E}(L_i|Z_{S=Q}) = (1-\bar{R}_i)\Phi\left[\frac{\left\{\Phi^{-1}(\bar{P}_i)-\alpha_i^Q Z\right\}}{\sqrt{1-(\alpha_i^Q)^2}}\right]$.

References

Altman EI, Brady B, Resti A, Sironi A (2005) The link between default and recovery rates: theory, empirical evidence, and implications. J Bus 78:2203–2228

Amraoui S, Hitier S (2008) Optimal stochastic recovery for base correlation. Working paper, BNP Paribas

Amraoui S, Cousot L, Hitier S, Laurent JP (2012) Pricing CDOs with state-dependent stochastic recovery rates. Quant Financ 12:1219–1240

Andersen LB, Sidenius J (2004) Extensions to the Gaussian copula: random recovery and random factor loadings. J Credit Risk 1:29–70

Ang A, Bekaert G (2002) International asset allocation with regime shifts. Rev Financ Stud 15:1137–1187

Ang A, Chen J (2002) Asymmetric correlations of equity portfolios. J Financ Econ 63:443–494

Bonti G, Kalkbrener M, Lotz C, Stahl G (2006) Credit risk concentrations under stress. J Credit Risk 2:115–136

Bruche M, González-Aguado C (2010) Recovery rates, default probabilities, and the credit cycle. J Bank Financ 34:754–764

Carty V, Hamilton DT, Keenan SC, Moss A, Mulvaney M, Marshella T, Subhas M (1998) Bankrupt bank loan recoveries. Moodys Invest Serv 15:79

Chen H (2010) Macroeconomic conditions and the puzzles of credit spreads and capital structure. J Financ 65:2171–2212

Chen J, Liu Z, Li S (2014) Mixed copula model with stochastic correlation for CDO pricing. Econ Modell 40:167–174

Choi D, Jen FC (1991) The relation between stock returns and short-term interest rates. Rev Quant Financ Acc 1:75–89

Choroś-Tomczyk B, Härdle WK, Okhrin O (2013) Valuation of collateralized debt obligations with hierarchical Archimedean copulae. J Empir Financ 24:42–62

Choroś-Tomczyk B, Härdle WK, Overbeck L (2014) Copula dynamics in CDOs. Quant Financ 14:1573–1585

Crouhy M, Galai D, Mark R (2000) A comparative analysis of current credit risk models. J Bank Financ 24:59–117

Das SR, Hanouna P (2009) Hedging credit: equity liquidity matters. J Financ Intermed 18:112–123
Drehmann M, Tarashev N (2013) Measuring the systemic importance of interconnected banks. J Financ Intermed 22:586–607
Elouerkhaoui Y (2009) Base correlation calibration with a stochastic recovery model. Working paper, Citigroup Global Markets
Ferreira MA, Laux PA (2007) Corporate governance, idiosyncratic risk, and information flow. J Financ 62:951–989
Franke J, Härdle W, Hafner C (2011) Statistics of financial markets: an introduction. Springer, Berlin
Frey R, McNeil AJ (2003) Dependent defaults in models of portfolio credit risk. J Risk 6:59–92
Goyal A, Santa-Clara P (2003) Idiosyncratic risk matters! J Financ 58:975–1007
Hull J (2006) Options, futures, and other derivatives. Pearson Education, India
Hull JC, White AD (2000) Valuing credit default swaps I. J Derivatives 8:29–40
Hull JC, White AD (2004) Valuation of a CDO and an n-th to default CDS without Monte Carlo simulation. J Deriv 12:8–23
Hull J, Nelken I, White A (2004) Merton's model, credit risk, and volatility skews. J Credit Risk 1:05
Jarrow RA, Lando D, Turnbull SM (1997) A Markov model for the term structure of credit risk spreads. Rev Financ Stud 10:481–523
Kalemanova A, Schmid B, Werner R (2007) The normal inverse Gaussian distribution for synthetic CDO pricing. J Deriv 14:80–94
Khieu HD, Mullineaux DJ, Yi HC (2012) The determinants of bank loan recovery rates. J Bank Financ 36:923–933
Kim J, Finger CC (2000) A stress test to incorporate correlation breakdown. J Risk 2:5–19
King THD, Khang K (2005) On the importance of systematic risk factors in explaining the cross-section of coporate bond yield spreads. J Bank Financ 29:3141–3158
Krupskii P, Joe H (2013) Factor copula models for multivariate data. J Multivar Anal 120:85–101
Longin F, Solnik B (2001) Extreme correlation of international equity markets. J Financ 56:649–676
Merton RC (1974) On the pricing of corporate debt: the risk structure of interest rates. J Financ 29(2):449–470
Pagratis S, Stringa M (2009) Modeling bank senior unsecured ratings: a reasoned structured approach to bank credit assessment. Int J Central Bank 5(2):1–39
Pan J, Singleton KJ (2008) Default and recovery implicit in the term structure of sovereign CDS spreads. J Financ 63:2345–2384
Patton AJ (2004) On the out-of-sample importance of skewness and asymmetric dependence for asset allocation. J Financ Economet 2:130–168
Rodríguez-Moreno M, Peña JI (2013) Systemic risk measures: the simpler the better? J Bank Financ 37:1817–1831
Rosen D, Saunders D (2010) Risk factor contributions in portfolio credit risk models. J Bank Financ 34:336–349
Schönbucher PJ (2001) Factor models: portfolio credit risks when defaults are correlated. J Risk Finance 3:45–56
Schwerter S (2011) Basel III's ability to mitigate systemic risk. J Financ Regul Compliance 19:337–354
Tarashev N, Borio C, Tsatsaronis K (2010) Attributing systemic risk to individual institutions. Working paper, BIS No. 308
Uhde A, Michalak TC (2010) Securitization and systematic risk in European banking: empirical evidence. J Bank Financ 34:3061–3077
Van der Voort M (2007) Factor copulas. J Deriv 14:94–102
Weiß GNF (2013) Copula-GARCH versus dynamic conditional correlation: an empirical study on VaR and ES forecasting accuracy. Rev Quant Finance Acc 41:179–202
Xiang V, Chng MT, Fang V (2015) The economic significance of CDS price discovery. Rev Quant Finance Acc. doi:10.1007/s11156-015-0540-2

Econometrics and Statistics 1 (2017) 101–117

Contents lists available at ScienceDirect

Econometrics and Statistics

journal homepage: www.elsevier.com/locate/ecosta

Change point and trend analyses of annual expectile curves of tropical storms

P. Burdejova [a], W. Härdle [a,c], P. Kokoszka [b,*], Q. Xiong [b]

[a] Humboldt-Universität zu Berlin, C.A.S.E. - Center for Applied Statistics and Economics, Spandauer Str. 1, Berlin 10178, Germany
[b] Department of Statistics, Colorado State University, 1877 campus delivery, Fort Collins, CO 80523, USA
[c] Sim Kee Boon Institute for Financial Economics, Singapore Management University, 81 Victoria Street, Singapore 188065, Singapore

ARTICLE INFO

Article history:
Received 20 January 2016
Revised 11 September 2016
Accepted 12 September 2016
Available online 9 November 2016

Keywords:
Change point
Trend test
Tropical storms
Expectiles
Functional data analysis

ABSTRACT

Motivated by the conjectured existence of trends in the intensity of tropical storms, new inferential methodology to detect a trend in the annual pattern of environmental data is developed. It can be applied to any data which form a time series of functions. Other examples include annual temperature or daily pollution curves at specific locations. Within a framework of a functional regression model, two tests of significance of the slope function are derived. One of the tests relies on a Monte Carlo distribution to compute the critical values, the other is pivotal with the chi–square limit distribution. Full asymptotic justification of both tests is provided. Their finite sample properties are investigated by a simulation study. Applied to tropical storm data, these tests show that there is a significant trend in the shape of the annual pattern of upper wind speed levels of hurricanes.

1. Introduction

A great deal of research in environmental and climate sciences has been dedicated to detecting change points and trends in various time series, including those related to temperature, precipitation and wind speed. In a typical setting, a scalar time series $X_1, X_2, \ldots, X_N$ is analyzed. Sometimes several correlated series are considered. Most environmental and climate series exhibit a pronounced annual periodicity which must be removed, or otherwise accounted for, before statements on change–points or trends can be inferred. Sometimes, it is difficult to approximate the periodic component by a Fourier expansion due to the irregular domain and amplitude of observations within a year. The data that motivate this work are tropical storm wind speed data, examples are shown in Figs. 1 and 2. By definition, only storms having the wind speed over 63 kilometers per hour are considered as tropical storms. The onset and end of typhoon and hurricane seasons, as well as their intensity, can change from year to year. We therefore propose to treat the data available for a whole year as a single high–dimensional data object and perform the change point and trend analyses on these objects rather than the scalar observations directly. Such an approach is now relatively well–established in the field of functional data analysis (FDA), the monographs of Horváth and Kokoszka (2012) or Ferraty and Vieu (2006) contain many examples. Methodological foundations of FDA are addressed in Ramsay and Silverman (2005), its mathematical foundations in Hsing and Eubank (2015). While the amount of information available in the data is invariably reduced by various smoothing and dimension reduction methods, the most

* Corresponding author.
E-mail address: Piotr.Kokoszka@colostate.edu (P. Kokoszka).

http://dx.doi.org/10.1016/j.ecosta.2016.09.002

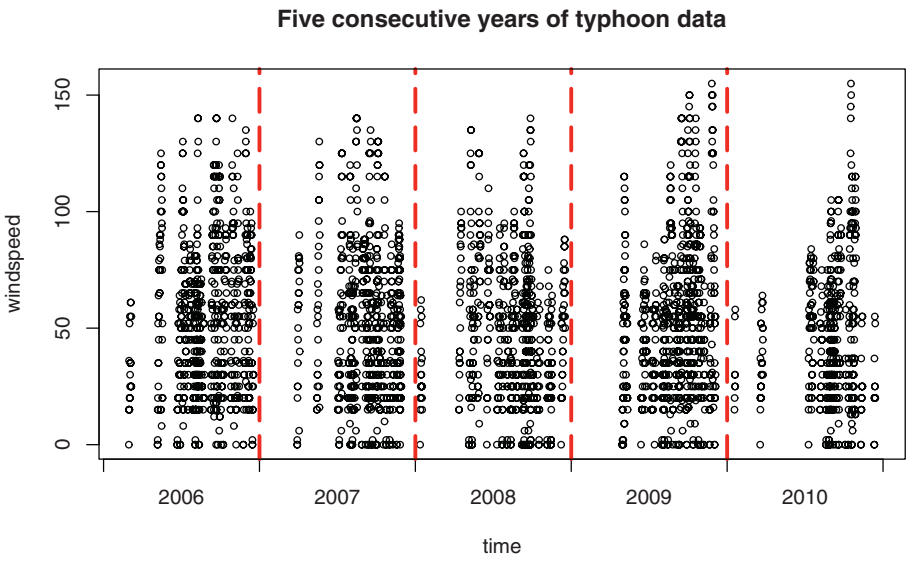


Fig. 1. Five consecutive years (2006–2010) of typhoon data. The dots represent the wind speed measurements. Dashed vertical lines separate the years.

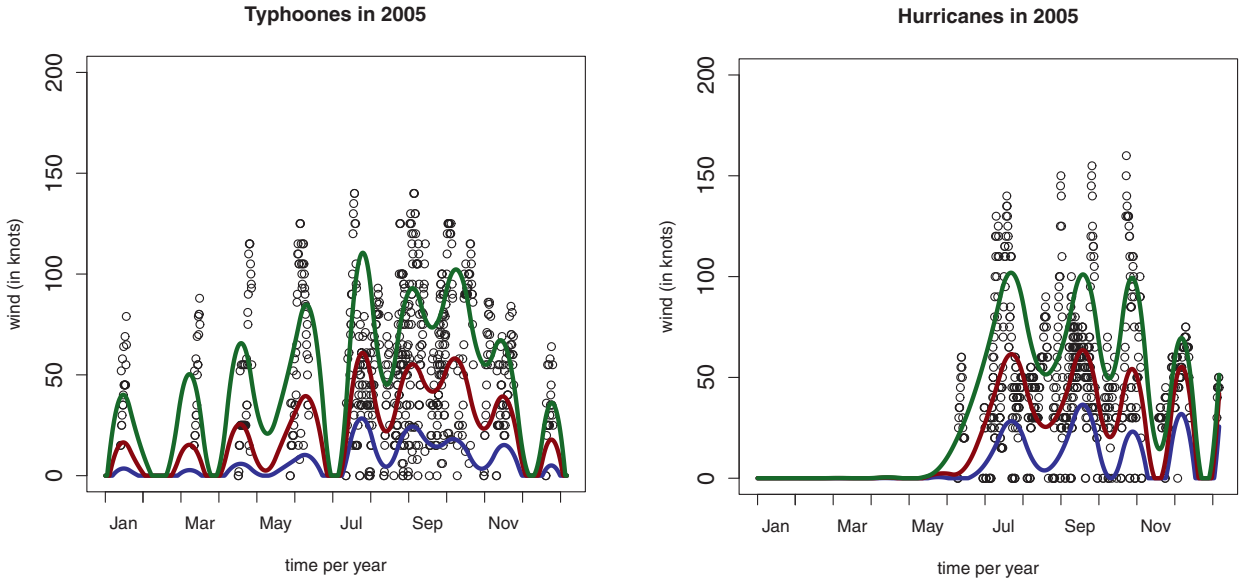


Fig. 2. Typhoons (left) and hurricanes (right) data in 2005 with expectile curves for $\tau = 0.1, 0.5$ and 0.9. The dots represent the wind speed measurements. Generally, a vertical streak of dots represents one tropical storm event. The lines are the estimated expectile curves.

important and relevant features of the data come into focus. In the problems we study in this paper, we are interested in the evolution of the annual pattern of tropical storms strength over several decades, not in specific hourly measurements.

The data objects that this paper studies have the form $X_n(t)$, where n refers to year, and t to time within the year. In the framework of functional data analysis, t is viewed as a continuous argument. The data are observed at a regular or irregular grid, but are converted to functional objects by means of various basis expansions which are defined for every t. We consider a sequence of curves $X_n(t, \tau)$ for several expectile levels $\tau \in (0, 1)$; these are similar to quantile levels. Examples of expectile curves we study are given in Fig. 2.

The index $\tau \in (0, 1)$ has the following interpretation. If $\tau = 0.5$, the curve $X_n(t, \tau)$ describes the median strength of tropical storms throughout the year. If τ is close to 1, the curve $X_n(t, \tau)$ captures the annual pattern of highest wind speeds. If τ is close to zero, it does the same for the lowest wind speeds. We are interested in detecting change points and trends in the functional time series $X_1(\cdot, \tau), X_2(\cdot, \tau), \ldots, X_N(\cdot, \tau)$. For this purpose, we use the existing change point test of Berkes et al. (2009) and develop two trend tests. No trend tests have presently been available for the data structure described

above. These two tests form a methodological contribution to statistics, while the analysis of the expectile curves of tropical storms provides an insight to climate science.

We thus focus not only on the average pattern but on change points and trends in annual curves which describe the behavior at various levels of wind speed. This is illustrated in Fig. 2. The curves in the middle summarize the pattern of average wind speed. These curves will exhibit some evolution from year to year. The curves above them summarize the annual patterns of the highest speeds; they may exhibit a different evolution than the average curves. This issue is well–known in climate research; typically trends in the averages are contrasted with trends in extremes. In our application, no modeling of extreme behavior is required, the expectile curves are within the range of the data points. They provide information of behavior which lies between the typical behavior and the unobservable extreme behavior. Following the work of Smith (1989), evaluation of trends in extremes has attracted a great deal of attention, with respect to change point analysis of extremes, we are aware only of the work of Dierckx and Teugels (2010).

The paper is organized as follows. After reviewing the notion of expectile curves in Section 2, we review in Section 3 the test of Berkes et al. (2009) and present the two trend tests. Section 4 presents the results of a simulation study. The tests are applied in Section 5 to the analysis of expectile curves. The last section contains the details of the asymptotic theory for the trend tests.

2. Expectile curves

In this section we provide some background needed to understand how the expectile curves studied in this paper are constructed. The underlying concept of expectiles was first discussed by Newey and Powell (1987) and further analyzed in several directions, for example Efron (1991) and Rossi and Harwey (2009) focused on time-varying expectiles. Most relevant to our setting is the paper by Schnabel and Eilers (2009), which extended the work of Eilers and Marx (1996). It combined the LAWS (least average weighted squares) algorithm with P-splines in order to estimate expectile curves. Recent applications include Guo and Härdle (2012), Sobotka et al. (2013) and Guo et al. (2015) or more applicable one in finance by Taylor (2008), where Value at risk (VaR) and Expected shortfall (ES) were estimated using expectiles. Expectiles have a similar interpretation as quantiles, but have some desirable properties outlined in the references cited above.

Consider a scatter plot of points $(t_i, x_i), 1 \le i \le I$. In our applications, the t_i correspond to times within a year at which wind speed is measured and x_i to the wind speed. Since the form of the dependence of the x_i on the t_i is unknown, a B-spline expansion is used. We thus assume that

$$x_i \approx g_a(t_i) = \sum_{j=1}^{J} a_j B_j(t_i),$$

and find coefficients $a = (a_1, a_2, \ldots, a_J)$ which minimize

$$S_\tau(a) = (1-\tau)S_-(a) + \tau S_+(a),$$

where

$$S_-(a) = \sum_{x_i \le g_a(t_i)} \{x_i - g_a(t_i)\}^2 \quad \text{and} \quad S_+(a) = \sum_{x_i > g_a(t_i)} \{x_i - g_a(t_i)\}^2.$$

If τ is close to 1, then $S_+(a)$ must be made small. This means that the curve g_a will be above most of the points (t_i, x_i).

Denote a matrix of B-splines differences as D. In order to control the smoothness of curves we can add penalization and minimize

$$S_\tau(a) + \lambda a^\top D^\top D a,$$

with λ as shrinkage parameter chosen by a desired criterion. We chose λ according to AIC criterion. After finding $\hat{a}_j$ using penalized spline estimation, the expectile curve is obtained as $\sum_{j=1}^{J} \hat{a}_j B_j(t_i)$. For our computation we set up $J = 20$. The estimation algorithm is implemented in the R package `expectreg`, see Sobotka et al. (2014). Further details are presented in Schnabel (2011) or Schnabel and Eilers (2013).

3. Change point and trend tests

This section presents the significance tests that will be applied to tropical storm data in Section 5. The change point test described in Section 3.1 was derived by Berkes et al. (2009), it is also described in Chapter 6 of Horváth and Kokoszka (2012). Trend tests introduced in Section 3.2 are new; their full large sample justification is presented in the last section. In both inferential settings, we consider as sequence of curves $X_n(t), t \in [0,1],\ n = 1, 2, \ldots N$. The index n can be identified with year, the index t with time within the year normalized to unit interval. The exposition that follows uses now fairly standard concepts of functional data analysis, including functional principal components (FPC's) and their empirical counterparts (EFPC's), see, for example, Chapter 3 of Horváth and Kokoszka (2012).

Table 1
Critical values of the distribution of K_d, which approximates the distribution of the statistic $\widehat{S}_d$ for large N.

d	5	6	7	8	9	10	11	12
10%	1.2797	1.4852	1.6908	1.8974	2.0966	2.2886	2.4966	2.6862
5%	1.4690	1.6847	1.8956	2.1242	2.3227	2.5268	2.7444	2.9490
1%	1.8667	2.1260	2.3423	2.5893	2.8098	3.0339	3.2680	3.4911

3.1. Change point test

In change point tests, the null hypothesis is that the mean function does not change with year:

$$H_0: \ \mathrm{E}X_1 = \mathrm{E}X_2 = \cdots = \mathrm{E}X_N.$$

The specific value of the mean is not part of the null hypothesis. The alternative is that there is at least one *unknown* change point k^* such that the equality under H_0 fails. The theory and practice of change points tests have been described in many textbooks, for example, Brodsky and Darkhovsky (1993), Csörgő and Horváth (1997), Chen and Gupta (2011), so we do not dwell on the background and move on to the description of the test of Berkes et al. (2009).

The test is based on the normalized differences of estimated mean functions:

$$P_k(t) = \frac{k(N-k)}{N}\{\hat{\mu}_k(t) - \tilde{\mu}_k(t)\},$$

where

$$\hat{\mu}_k(t) = k^{-1}\sum_{i=1}^{k} X_i(t), \quad \tilde{\mu}_k(t) = (N-k)^{-1}\sum_{i=k+1}^{N} X_i(t).$$

Next, we compute the estimated functional principal components $\hat{v}_\ell$ of the curves X_n and calculate the scores

$$\hat{\xi}_{j,n} = \int_0^1 \left\{X_n(t) - \bar{X}_N(t)\right\}\hat{v}_j(t)dt, \quad \bar{X}_N(t) = N^{-1}\sum_{n=1}^{N} X_n(t). \tag{3.1}$$

We find the smallest d such that 85% of the variance is explained and calculate the test statistic

$$\widehat{S}_d = \frac{1}{N^2}\sum_{j=1}^{d}\frac{1}{\hat{\lambda}_j}\sum_{k=1}^{N}\left(\sum_{1\le i\le k}\hat{\xi}_{j,i} - \frac{k}{N}\sum_{1\le i\le k}\hat{\xi}_{j,i}\right).$$

As $N \to \infty$, the statistics $\widehat{S}_d$ converges in distribution to the random variable K_d whose critical values are given Table 1, see Horváth and Kokoszka (2012) for more details.

3.2. Trend tests

Suppose the functions $X_n(t)$ follow the trend model

$$X_n(t) = \alpha(t) + \beta(t)n + \varepsilon_n(t). \tag{3.2}$$

The testing problem in our setting is

$$H_0: \ \beta = 0, \quad \text{vs.} \quad H_A: \ \beta \neq 0.$$

The paper thus focuses on a *linear* trend, which is the most common type of trend considered in atmospheric sciences. The review paper of Kossin et al. (2013) discusses research on linear trends in the context of tropical storms. The assumption of a linear trend makes the development of significance tests easier and leads to readily interpretable results if the null is rejected. More general nonlinear trends can often be displayed using various smoothing methods, but the assessment of their significance and interpretation are difficult due the lack of a simple parameterization. It is however possible to develop tests based on different approaches. Fraiman et al. (2014) propose a permutation test based on the proportion of time t the curve $X_n(t)$ matches the record curve $r_n(t) = \max_{1\le k\le n} X_k(t)$. We are however not aware of other approaches to test the presence of an increasing trend in a sequence of functions. Gromenko and Kokoszka (2013) consider curves $X(\mathbf{s}_k, t)$ defined at spatial locations $\mathbf{s}_k$ and test $H_0: \beta = 0$ in the model $X(\mathbf{s}_k, t) = \alpha + \beta t + \varepsilon(\mathbf{s}_k, t)$.

Before proceeding with the description of our testing approach we state the assumptions on the objects appearing in (3.2).

Assumption 3.1. The error curves ε_n are iid elements of the Hilbert space of square integrable functions with finite second moment: $E\int \varepsilon_n^2(t)dt < \infty$. The functions α and β are deterministic elements of that space: $\int\alpha^2(t)dt < \infty$, $\int\beta^2(t)dt < \infty$.

Assumption 3.1 holds throughout the paper.

A natural approach to testing is based on an estimator of β. If this estimator is small for all $t \in [0, 1]$, there is not enough evidence to reject H_0.

Representing trend model (3.2) as the regression

$$\begin{bmatrix} X_1(t) \\ \vdots \\ X_N(t) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & N \end{bmatrix} \cdot \begin{bmatrix} \alpha(t) \\ \beta(t) \end{bmatrix} + \begin{bmatrix} \varepsilon_1(t) \\ \vdots \\ \varepsilon_N(t) \end{bmatrix},$$

we obtain the least squares estimators

$$\hat{\alpha}(t) = \frac{2}{N(N-1)} \sum_{k=1}^{N} (2N + 1 - 3k) X_k(t) \tag{3.3}$$

and

$$\hat{\beta}(t) = \frac{6}{N(N+1)(N-1)} \sum_{k=1}^{N} (2k - N - 1) X_k(t). \tag{3.4}$$

Our first approach is based on the statistic $\int_0^1 \hat{\beta}^2(t)dt$. To describe its asymptotic distribution additional notation is needed. Introduce the covariance function of the errors $c_\varepsilon(t, s) = \mathrm{E}[\varepsilon_n(t)\varepsilon_n(s)]$. Denote by λ_j, $j = 1, 2, \ldots$ the eigenvalues of c_ε. Next, define the residuals

$$\hat{\varepsilon}_n(t) = X_n(t) - \hat{\alpha}_n(t) - \hat{\beta}_n(t)n \tag{3.5}$$

and denote by $\hat{\lambda}_j$ the eigenvalues of the empirical covariance function

$$\hat{c}_\varepsilon(t, s) = \frac{1}{N} \sum_{n=1}^{N} \hat{\varepsilon}_n(t)\hat{\varepsilon}_n(s). \tag{3.6}$$

Theorem 3.1 describes large sample properties of the suitably normalized statistic $\int_0^1 \hat{\beta}^2(t)dt$.

Theorem 3.1.

(i) *Under* H_0,

$$\widehat{\Lambda}_N = \frac{N^3}{12} \int_0^1 (\hat{\beta}(t))^2 dt \xrightarrow{\mathcal{L}} \Lambda_\infty \overset{def}{=} \sum_{j=1}^{\infty} \lambda_j Z_j^2, \tag{3.7}$$

where $\{Z_j, j \geq 1\}$ *are independent standard normal variables, and the* λ_j *are the eigenvalues of the covariance function* c_ε.

(ii) *Under* H_A,

$$\mathrm{P}\{\widehat{\Lambda}_N > q_N(\alpha)\} \to 1, \quad \text{as } N \to \infty, \tag{3.8}$$

where $q_N(\alpha)$ *is the* $(1-\alpha)$*th quantile of the distribution of* $\Lambda_N = \sum_{j=1}^{N} \hat{\lambda}_j Z_j^2$.

Theorem 3.1 is proven in the last section.

The distribution of Λ_∞ can be approximated by the distribution of

$$\Lambda_N = \sum_{j=1}^{N} \hat{\lambda}_j Z_j^2. \tag{3.9}$$

This leads to the Monte Carlo test whose consistency is claimed in part (ii) of Theorem 3.1. To implement the test, we generate a large number, say $R = 10^4$, of independent replications of Λ_N (the $\hat{\lambda}_j$ are estimated only once, from the original sample). Denote these replications by $\Lambda_{N,r}$, $1 \leq r \leq R$. The *P*-value of the test is computed as the fraction of the $\Lambda_{N,r}$ which are greater than $\widehat{\Lambda}_N$ (computed from the data).

It is also possible to develop a test similar to the test of Berkes et al. (2009) in the sense that a limit distribution is independent of the distribution of the data. In fact, in the trend model, the limit distribution is the usual chi–square distribution. This is stated in Theorem 3.2, in which we use the inner product notation $\langle f, g \rangle = \int_0^1 f(t)g(t)dt$.

Theorem 3.2. *Suppose* $\mathrm{E}\|\varepsilon\|^4 < \infty$ *and*

$$\lambda_1 > \lambda_2 > \ldots > \lambda_q > \lambda_{q+1} > 0. \tag{3.10}$$

(i) Under H_0,

$$\widehat{T}_N = \frac{N^3}{12} \sum_{j=1}^{q} \hat{\lambda}_j^{-1} \langle \hat{\beta}, \hat{v}_j \rangle^2 \xrightarrow{\mathcal{L}} \chi_q^2. \tag{3.11}$$

(ii) If for some $1 \le j \le q$, $\langle \beta, v_j \rangle \neq 0$, *then the test is consistent, i.e.*

$$\mathrm{P}\{\widehat{T}_N > q(\alpha)\} \to 1, \quad \text{as } N \to \infty, \tag{3.12}$$

where $q(\alpha)$ *is the* $(1-\alpha)$*th quantile of the chi–square distribution with* q *degrees of freedom.*

Theorem 3.2 is proven in the last section.

Observe that to establish the consistency, it is not enough to assume $\beta \neq 0$ in L^2. Since the statistic $\widehat{T}_N$ is based on projections on the first q EFPC's, we must assume that the slope function β is not orthogonal to the subspace spanned by the first q FPC's.

Under the assumption of iid error curves ε_n, cf. Assumption 3.1, the functional principal components used in this paper offer an optimal expansion. However, if the Assumption 3.1 is relaxed to allow some form of weak dependence, for example the approximability introduced in Hörmann and Kokoszka (2010), then a different data–driven orthonormal system may offer some advantages. For example, the long–run FPC's of Horváth et al. (2013) or the dynamic FPC's of Hörmann et al. (2015) could be used. These systems however require selections of kernel functions and other tuning parameters, whose selection and impact would need to be studied. We expect that the test statistics could be formulated in an analogous way and their asymptotic distribution would have a similar form to those we derived. Some work in relation to change point tests has been done by Torgovitsky (2016). Theoretical and practical exploration of similar extensions of trend tests is an interesting topic for future research.

4. Finite sample performance of the trend tests

A simulation study validating the change point test of Section 3.1 is reported in Berkes et al. (2009). In this section, we examine the finite sample performance of the trend tests introduced in Section 3.2.

We consider two models for the error functions $\varepsilon_n(t)$. The first is a generic Gaussian model in which we take the $\varepsilon_n(t)$ to be Brownian bridges $B_n(t)$. We represent Brownian bridge as a Fourier series with stochastic coefficients (the Karhunen–Loéve expansion, see Bosq (2000)):

$$B_n(t) = \sqrt{2}\sum_{j=1}^{\infty} Z_{nj}\frac{\sin(j\pi t)}{j\pi} \approx \sqrt{2}\sum_{j=1}^{J} Z_{nj}\frac{\sin(j\pi t)}{j\pi},$$

where $\{Z_j,\ j \ge 1\}$ are independent standard normal random variables. We set $J = 100$ so the trajectories of the B_n have similar smoothness as the typhoon and hurricane expectile curves.

The second model for the ε_n is based more directly on the tropical storm data. We proceed as follows. We consider $\tau = 0.1, 0.5, 0.9$. For each level τ, we compute the sample mean function and the sample functional principal components $\hat{v}_j(t;\tau)$ of the expectile curves $X_n(t,\tau)$. Next we compute the scores $\xi_{jn}(\tau)$ according to (3.1). Denote by $\sigma_j(\tau)$ the standard deviation of the $\xi_{jn}(\tau)$, $1 \le n \le N$, $(N = 65)$. The ε_n are generated as independent realizations of the random function

$$\varepsilon(t;\tau) = \sum_{j=1}^{q} \sigma_j(\tau) Z_j \hat{v}_j(t;\tau), \qquad Z_j \sim \text{ iid } \mathrm{N}(0,1),$$

with q determined from the original expectile curves according to the 85% rule. We thus have four models for the error curves which we refer to as BB, E1, E5, E9. The errors E1, E5, E9 are different depending on whether hurricane or typhoon data are used. The empirical rejection rates are however very similar in both cases. We display the results for the errors based on the hurricane data.

We generate artificial data according to the specification

$$X_n(t) = b\beta(t)n + \varepsilon_n(t).$$

To find empirical size, we set $\beta(t) = \beta_0(t) = 0$. To find empirical power, we use the slope functions

$$\beta_1(t) = -\frac{\cos\left(\frac{t\pi 3}{2}\right)}{100}; \qquad \beta_2(t) = \frac{\sin(t\pi 20)}{100},$$

which are graphed in Fig. 3. The constant b is used to adjust the magnitude of the departure from the null hypothesis. For E1, E5 and E9 error curves we set $b = 20$, for BB errors we use $b = 1$. The different values are used to ensure similar signal to noise ratio for both types of errors.

We consider sample sizes $N = 30, 60, 120$. Empirical rejection rates are shown in Tables 2 and 3. The Monte Carlo test, generally has slightly better size and power, but the pivotal chi–square test performs well too. The chi–square test tends to overreject under H_0 (for $N = 60$ and $N = 120$).

5. Application to typhoon and hurricane data

In this section we apply the tests of Section 3 to annual expectile curves of wind speed data. The data have the form $X_n(t_i)$, where the times t_i are separated by six hours, and the index n stands for year. The value $X_n(t_i)$ is the wind speed in knots (1 kn = 0.5144 m/s). We work with two data sets: typhoons in the West Pacific area over the period 1946–2010,

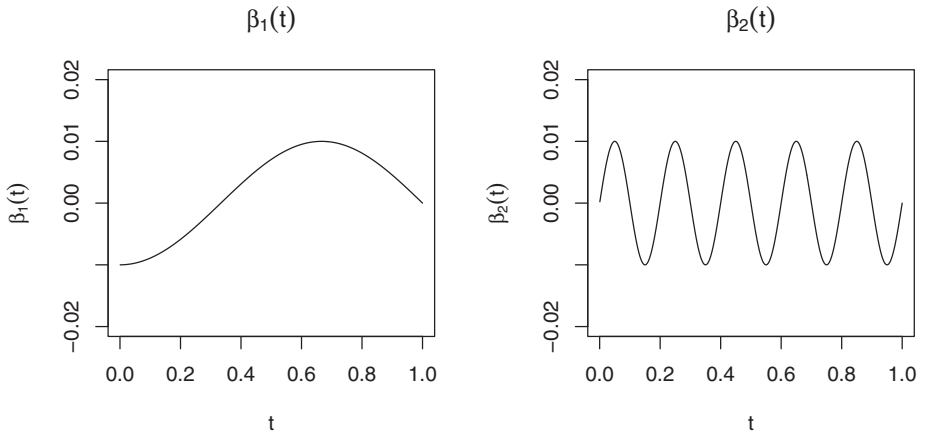


Fig. 3. Slope functions $\beta_1(t)$ (left) and $\beta_2(t)$ (right) used to assess power.

Table 2
Rejection rates of the **Monte Carlo test**. Columns corresponding to β_0 report empirical size, those to β_1 and β_2, empirical power.

BB	β_0	β_1	β_2	E1	β_0	β_1	β_2
$N = 30$	0.055	0.175	0.136	$N = 30$	0.060	0.082	0.078
$N = 60$	0.056	0.967	1.000	$N = 60$	0.045	0.438	0.440
$N = 120$	0.064	1.000	1.000	$N = 120$	0.042	1.000	1.000
E5	β_0	β_1	β_2	E9	β_0	β_1	β_2
$N = 30$	0.042	0.072	0.060	$N = 30$	0.069	0.081	0.091
$N = 60$	0.047	0.435	0.438	$N = 60$	0.058	0.435	0.404
$N = 120$	0.044	1.000	1.000	$N = 120$	0.042	1.000	1.000

Table 3
Rejection rates of the **Chi–square test**. Columns corresponding to β_0 report empirical size, those to β_1 and β_2, empirical power.

BB	β_0	β_1	β_2	E1	β_0	β_1	β_2
$N = 30$	0.064	0.344	0.053	$N = 30$	0.053	0.071	0.089
$N = 60$	0.058	0.995	0.085	$N = 60$	0.058	0.215	0.220
$N = 120$	0.069	1.000	0.238	$N = 120$	0.056	0.975	0.971
E5	β_0	β_1	β_2	E9	β_0	β_1	β_2
$N = 30$	0.047	0.065	0.044	$N = 30$	0.051	0.075	0.085
$N = 60$	0.064	0.249	0.193	$N = 60$	0.065	0.216	0.234
$N = 120$	0.049	0.982	0.898	$N = 120$	0.058	0.929	0.967

and hurricanes across the North Atlantic basin over the period 1947–2011. Both datasets are accessible free of charge at the website of Unisys Weather Information, UNISYS (2015).

Since there are about 1,460 time points t_i per year, we treat time $0 \le t \le T$ within a year as continuous, and the observed curves as functional data. For each year n, we construct expectile curves $X_n(t, \tau)$, for $\tau = 0.1, 0.2, \ldots, 0.9$. Examples of expectile curves we study are given in Fig. 2.

5.1. Change point analysis

The results of the application of the change–point test of Section 3.1 are shown in Table 4. For both data sets and at all levels τ, the test rejects the null hypothesis that the mean pattern does not change. As explained in Section 2, the construction of the expectile curves involves the selection of a smoothing parameter λ. Table 4 shows the results for λ selected by the AIC criterion. To validate our conclusions, we performed the same analysis using λ which is either twice or half of the λ selected by AIC. In both cases, all empirical significance levels remained under 5%.

The change point test shows that for all expectile levels τ, there are statistically significant changes in the annual pattern. It is instructive to complement the above inferential analysis by simple exploratory analysis that reveals some dependence on the level τ. Consider squared norms

$$P_k(\tau) = \int_0^T P_k^2(t, \tau)dt,$$

Table 4
Results of the application of the change point test of Section 3.1 to typhoon (upper panel) and hurricane (lower panel) expectile curves. Usual significance codes are used: ** – significant at 5% level, *** – at 1% level.

τ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
d	10	11	12	12	12	12	12	12	12
$\widehat{S}_d$	3.3522 ***	3.2291 **	3.4317 **	3.4978 ***	3.6564 ***	3.8554 ***	4.0342 ***	4.2317 ***	4.5084 ***
τ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
d	5	5	5	6	6	6	7	7	7
$\widehat{S}_d$	2.7440 ***	3.3993 ***	3.8759 ***	4.4640 ***	4.7141 ***	4.8680 ***	5.0366 ***	4.9247 ***	4.5740 ***

Table 5
P-values for the Monte Carlo trend test based on Theorem 3.1.

τ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Typhoons *P*-value	0.365	0.537	0.545	0.495	0.438	0.381	0.329	0.316	0.269
Hurricanes *P*-value	0.439	0.239	0.133	0.081	0.062	0.047	0.038	0.040	0.055

Table 6
P-values for the chi–square trend test based on Theorem 3.2.

τ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
q	10	11	12	12	12	12	12	12	12
Typhoons *P*-value	0.534	0.705	0.722	0.688	0.587	0.466	0.382	0.371	0.453
q	5	5	5	6	6	6	7	7	7
Hurricanes *P*-value	0.069	0.024	0.015	0.006	0.003	0.003	0.004	0.006	0.035

where the $P_k(t, \tau)$ are the normalized differences $P_k(t)$ introduced in Section 3.1 computed for the expectile level τ. The plot of $P_k(\tau)$ against the year index k shows the magnitude of change of the mean function. We display such plots in Fig. 4. They suggest that the largest changes occur for the expectile levels τ close to one, but it must be kept in mind that they may just reflect the fact that the curves $X_n(t)$ are "larger" for larger τ. By contract, the statistic $\widehat{S}_d$ contains a normalization with the variances $\hat{\lambda}_j$, and is scale invariant.

The change point analysis above shows that the pattern of typhoon and hurricane wind speeds cannot be treated as stable over the sample periods we study. In the next section, we investigate if this instability can be attributed to systematic trends.

5.2. Trend analysis

We now apply the trend tests introduced in Section 3.2 to typhoon and hurricane expectile curves. In the Monte Carlo test based on Theorem 3.1, we use 10^4 replications of the random variable Λ_N defined by (3.9). In the chi–square test based on Theorem 3.2, we determine q as the smallest number which explains at least 85% of the variance of the residual curves $\hat{\varepsilon}_n$ defined by (3.5). The results of the tests are presented in Tables 5 and 6.

For the typhoon data, none of the two tests finds evidence of a trend. For the hurricane data, the Monte Carlo test based on Theorem 3.1 indicates the existence of a trend for expectile levels $\tau = 0.6 - -0.9$ while the chi–square test of Theorem 3.2 for all τ except $\tau = 0.1$. Simulations reported in Section 4 show that the chi–square test tends to overreject for data generating processes (DGP's) of length and error structure similar to the tropical storm expectile curves. We therefore conclude that there is evidence for the existence of a trend for upper expectile functions of hurricane data. The estimated slope functions $\hat{\beta}$ are plotted in Fig. 5. While general shapes look similar, the curves are different for different values of τ, with difference of the order 0.05–0.10 on the same scale as in Figs. 5 and 6.

We conclude the trend analysis by showing in Fig. 7 the dependence on τ of the norm $\|\hat{\beta}\| = \sqrt{\int \hat{\beta}^2(t)dt}$ of the estimated slope function. Even though there is statistical evidence for nonzero slope function only for the upper expectiles of hurricane data, the exploratory analysis of the norms indicates that there is a very clear increasing dependence of the slope on τ. Again, the increasing norms could be attributed to the increasing size of the curve $X_n(t)$, and the plots can be used only as an exploratory tool for comparing the hurricane and typhoon data.

There is not much difference between the size of $\hat{\beta}$, for typhoon and hurricane data, but the $\hat{\beta}$ for hurricanes show a clear pattern with positive mass around July and November, and negative mass in early autumn. For the typhoon curves the pattern of mass accumulation is spread more uniformly throughout the year, with a pronounced negative mass in November. The significance tests we developed provide a statistical justification for these fairly subtle visual differences.

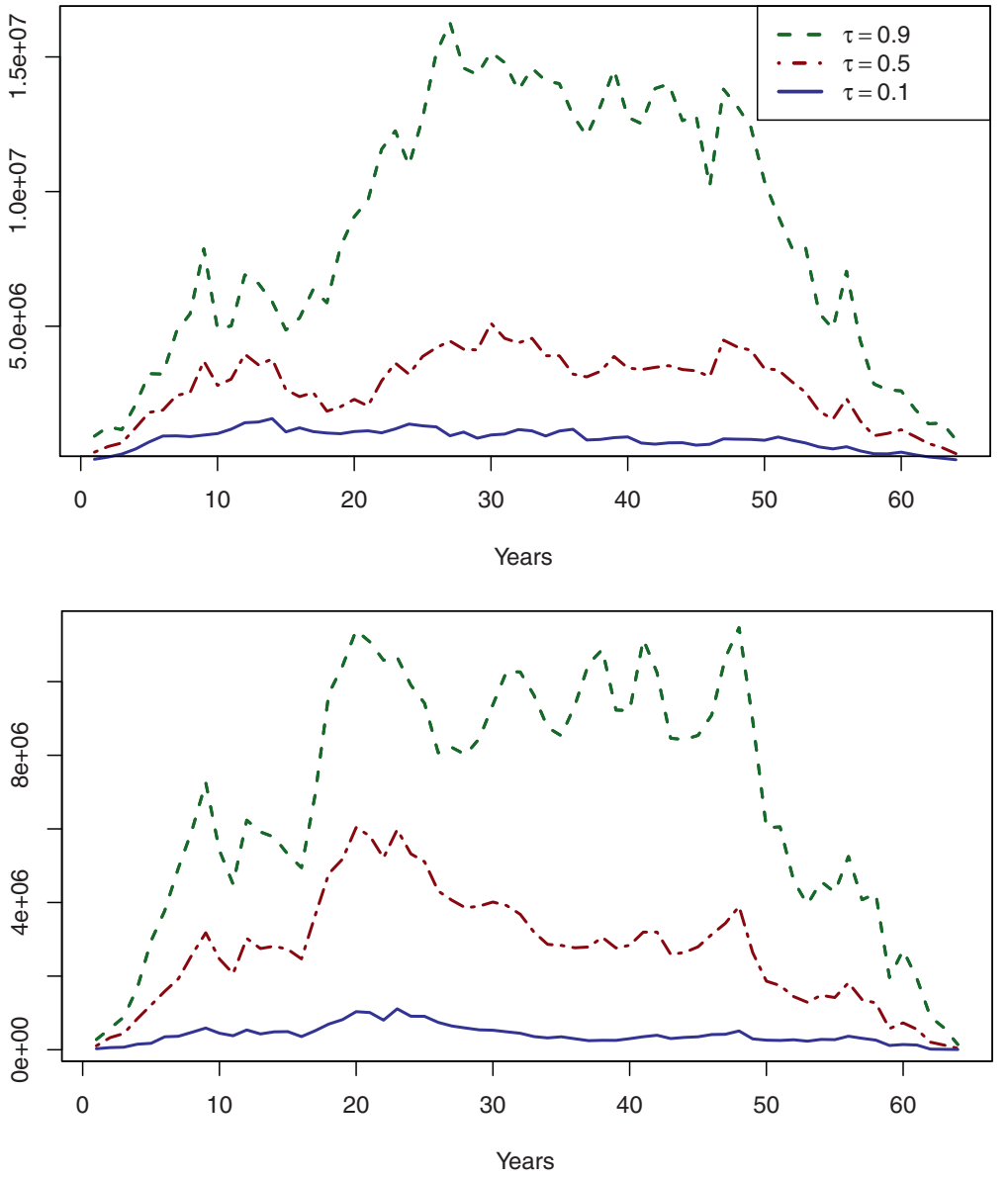


Fig. 4. The squared norms $P_k(\tau)$ showing the magnitude of change in mean annual pattern for expectile curves of typhoons (upper panel) and hurricanes (lower panel). The largest changes occur in the expectile curves corresponding to $\tau = 0.9$.

5.3. Main conclusions of data analysis

The change point tests have shown that the annual pattern of wind speeds for both hurricanes and typhoons cannot be treated as constant, no matter what expectile level is considered. If there is one or two clear–cut change points, their location can be found as the years n for which the curves P_n shown in Fig. 4 attain local maxima. For the tropical storm data, these plots show multiple local maxima indicating that either we must assume many change points or a continuous change, akin to trend. The application of the new trend tests has focused on a question which has however received a fair deal of attention: is there a trend in the intensity of tropical storms. A review of relevant research is not our aim, the paper of Kossin et al. (2013) provides background and references. There are two novel aspects to our approach: (1) focus on the annual curves, (2) separate analysis for each intensity level. Based on sixty years of data, our tests detect a trend in the upper wind speeds of Atlantic hurricanes. Exploratory analysis suggests a similar conclusion for Pacific typhoons, but it cannot be supported by low P-value with the amount of available data. These conclusions are similar to the findings of Kossin et al. (2013) who use different, custom–prepared, data sets. Their P-value for the existence of a trend in North Atlantic is less that 10^{-3}, but for the North–West pacific it is 0.03 (for South Pacific it is 0.09, 0.06 for the South Indian

Ocean). Their analysis is concerned with the trend in the scalar data, not a trend in the annual pattern. They find all trends to be positive. In a sense, such trend coefficients can be viewed as averages of the annual curves like those displayed in Figs. 5 and 6. The hurricane curves indeed have more positive mass, whereas for the typhoon curves the negative mass is larger (the typhoon curves are not statistically different from zero, according to our tests). The slope functions of the hurricanes indicate increasing intensity in summer and late fall, and decreasing intensity in early fall. For typhoons, these curves indicate decreasing intensity in November.

The conclusions of this paper which are supported by significance tests and do not contradict existing research are as follows:

1. The annual pattern of wind speeds of both hurricanes and typhoons has been changing at all wind speed levels over the last 60 years.
2. There is a significant trend in the shape of this pattern for upper wind speed levels of hurricanes.

6. Proofs of Theorems 3.1 and 3.2

Before proceeding to the proofs of Theorems 3.1 and 3.2, we observe that a direct verification shows that

$$c_\beta(t,s) \stackrel{\text{def}}{=} \text{Cov}\{\hat\beta(t), \hat\beta(s)\} = A_N c_\varepsilon(t,s),$$

where

$$A_N = \frac{12}{N(N+1)(N-1)}.$$

The constant A_N is repeatedly used in the proofs of Theorems 3.1 and 3.2.

6.1. Proof of Theorem 3.1

Proof of part (i): Under H_0 $(\beta = 0)$,

$$\hat\beta(t) = A_N \sum_{k=1}^{N} k\varepsilon_k(t) - \frac{1}{2}A_N(N+1)\sum_{k=1}^{N}\varepsilon_k(t).$$

Using the identity

$$\sum_{k=1}^{N} k\varepsilon_k = N\sum_{n=1}^{N}\varepsilon_n - \sum_{k=1}^{N-1}\sum_{n=1}^{k}\varepsilon_n, \tag{6.13}$$

we have

$$\begin{aligned}\hat\beta(t) &= A_N \sum_{k=1}^{N} k\varepsilon_k(t) - \frac{1}{2}A_N(N+1)\sum_{k=1}^{N}\varepsilon_k(t)\\ &= A_N\left(N\sum_{n=1}^{N}\varepsilon_n(t) - \sum_{k=1}^{N-1}\sum_{n=1}^{k}\varepsilon_n(t)\right) - \frac{1}{2}A_N(N+1)\sum_{n=1}^{N}\varepsilon_n(t).\end{aligned} \tag{6.14}$$

To determine the limit behavior of $\hat\beta(t)$, we thus need an invariance principle for the partial sum process:

$$S_N(x,t) = \frac{1}{\sqrt{N}}\sum_{1\le n\le [Nx]}\varepsilon_n(t), \quad 0\le x,t\le 1.$$

A result of this type has recently been established by Berkes et al. (2013). It states that

$$S_N(x,t) \xrightarrow{\mathcal{L}} \Gamma(x,t), \tag{6.15}$$

where $\Gamma(x, t)$ is the two parameter Gaussian process which admits the representation

$$\Gamma(x,t) = \sum_{j=1}^{\infty}\sqrt{\lambda_j}W_j(x)v_j(t), \tag{6.16}$$

where $\{W_j(x),\ 0 \le x \le 1\}$ are independent standard Wiener processes on [0, 1]. The λ_j and the v_j are, respectively, the eigenvalues and the eigenfunctions of the covariance function $c_\varepsilon(t,s) = \text{E}[\varepsilon_n(t)\varepsilon_n(s)]$. In (6.15), and whenever weak convergence of two parameter processes is concerned, $\xrightarrow{\mathcal{L}}$ denotes the convergence in the Skorokhod space $D([0,1], L^2)$.

Since $A_N \sim 12N^{-3}$, (6.14) implies

$$\hat\beta(t) = A_N N^{\frac{3}{2}} S_N(1,t) - A_N N^{\frac{1}{2}}\sum_{k=1}^{N-1} S_N\left(\frac{k}{N}, t\right) - \frac{1}{2}A_N(N+1)N^{\frac{1}{2}}S_N(1,t)$$

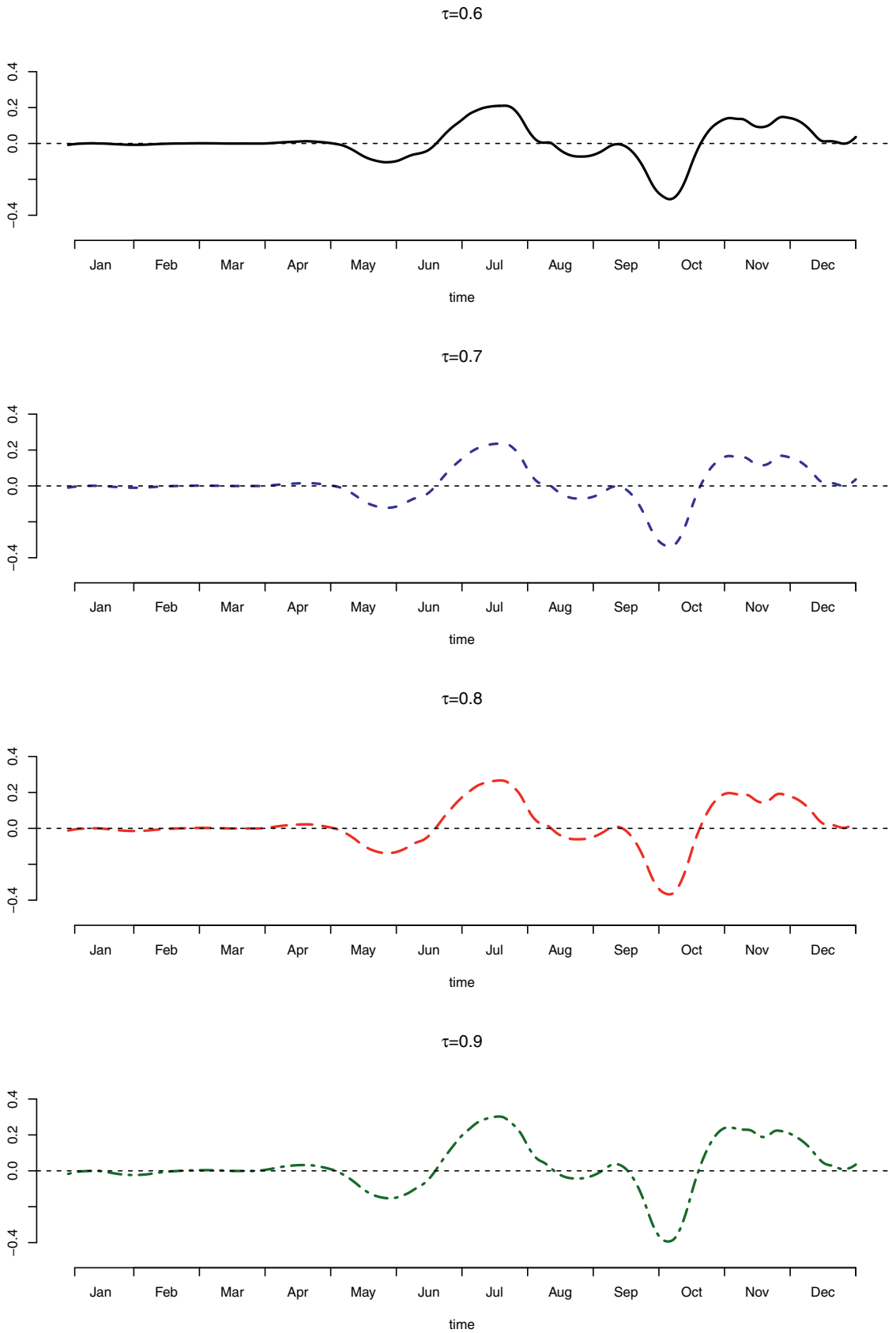


Fig. 5. Estimated slope functions, $\hat{\beta}$, for upper expectile curves of **hurricane** data.

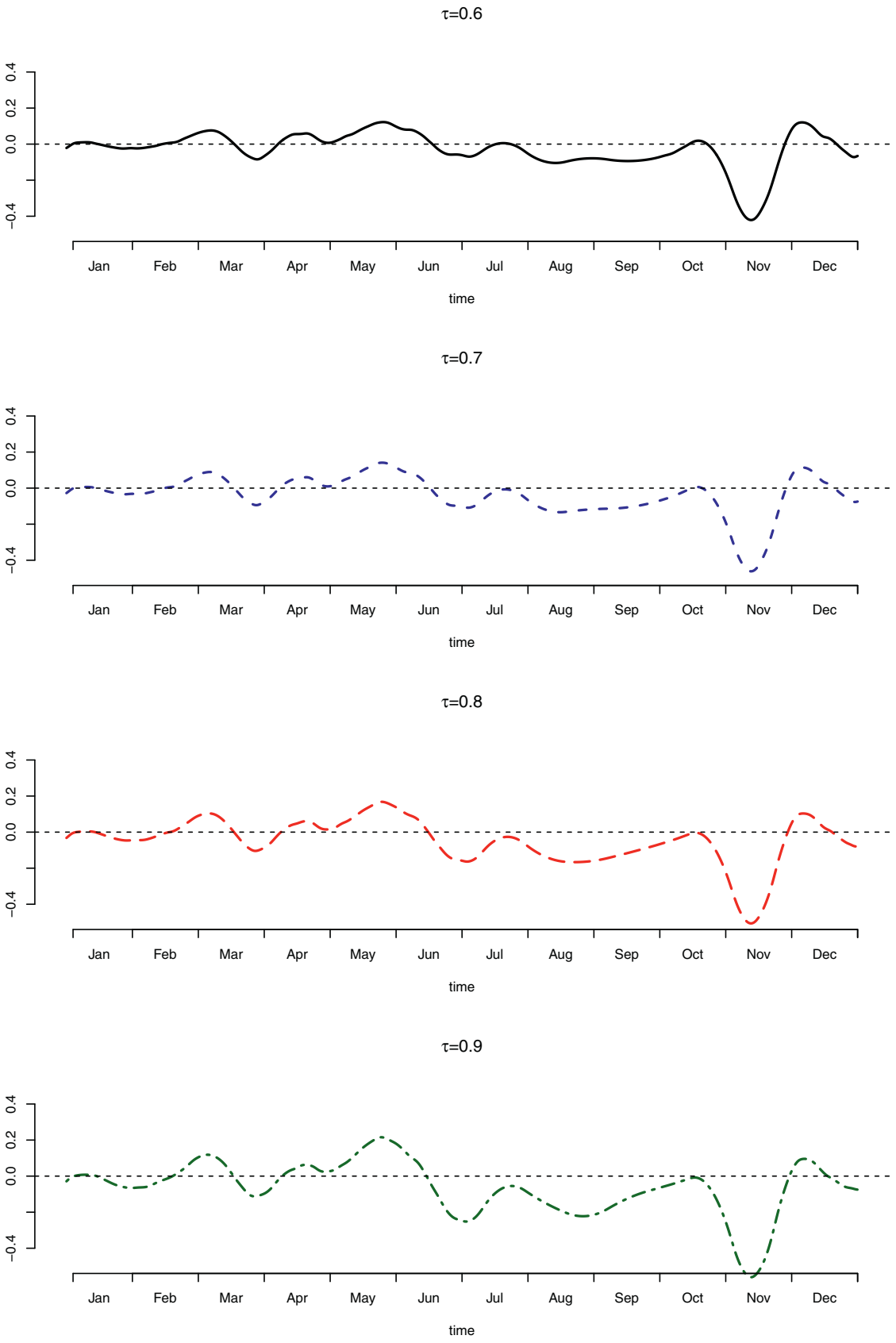


Fig. 6. Estimated slope functions, $\hat{\beta}$, for upper expectile curves of **typhoon** data.

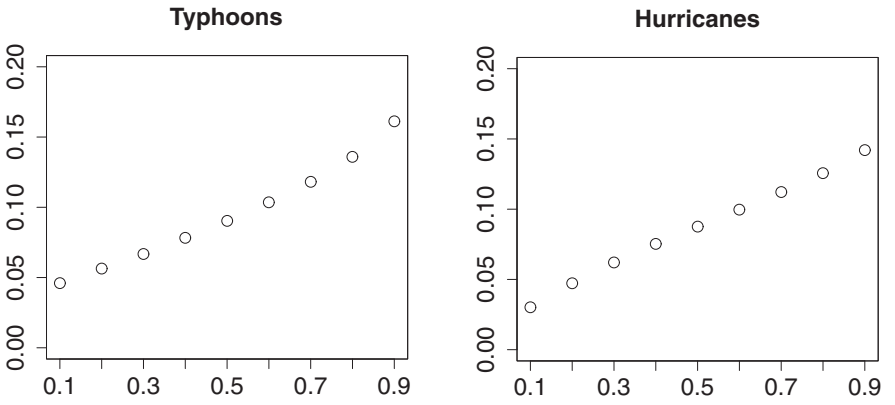


Fig. 7. Norm of the slope function estimate, $\hat{\beta}$, as a function of the expectile level τ; typhoons (left), hurricanes (right).

$$\sim 12N^{-\frac{3}{2}}S_N(1,t) - 12N^{-\frac{3}{2}}\left\{\frac{1}{N}\sum_{k=1}^{N-1}S_N\left(\frac{k}{N},t\right)\right\} - 6N^{-\frac{3}{2}}S_N(1,t)$$
$$= 6N^{-\frac{3}{2}}S_N(1,t) - 12N^{-\frac{3}{2}}\left\{\frac{1}{N}\sum_{k=1}^{N-1}S_N\left(\frac{k}{N},t\right)\right\}.$$

By the continuous mapping theorem and (6.15)

$$\frac{1}{N}\sum_{k=1}^{N-1}S_N\left(\frac{k}{N},t\right) \xrightarrow{\mathcal{L}} \int_0^1 \Gamma(x,t)dx.$$

Thus

$$\frac{N^{\frac{3}{2}}}{6}\hat{\beta}(t) \xrightarrow{\mathcal{L}} \Gamma(1,t) - 2\int_0^1 \Gamma(x,t)dx. \tag{6.17}$$

Using the continuous mapping theorem again, we obtain

$$\frac{N^3}{36}\int_0^1 \{\hat{\beta}(t)\}^2 dt \xrightarrow{\mathcal{L}} \int_0^1 \left\{\Gamma(1,t) - 2\int_0^1 \Gamma(x,t)dx\right\}^2 dt.$$

Set

$$D_j = W_j(1) - 2\int_0^1 W_j(x)dx, \tag{6.18}$$

so that, by (6.16), we have

$$\Gamma(1,t) - 2\int_0^1 \Gamma(x,t)dx = \sum_{j=1}^{\infty}\sqrt{\lambda_j}D_j v_j(t).$$

Then, by Parseval's identity,

$$\int_0^1 \left\{\Gamma(1,t) - 2\int_0^1 \Gamma(x,t)dx\right\}^2 dt = \left\|\sum_{j=1}^{\infty}\sqrt{\lambda_j}D_j v_j\right\|^2 = \sum_{j=1}^{\infty}\lambda_j D_j^2. \tag{6.19}$$

The random variables D_j are independent normal with mean zero and variance

$$\mathrm{Var}[D_j] = \mathrm{E}\left[W(1) - 2\int_0^1 W(x)dx\right]^2$$
$$= \mathrm{E}W^2(1) - 4\mathrm{E}\left[W(1)\int_0^1 W(x)dx\right] + 4\mathrm{E}\left[\int_0^1 W(x)dx\right]^2$$
$$= \frac{1}{3}.$$

We can write $D_j = \frac{1}{\sqrt{3}} Z_j$, where Z_j are standard normal variables. By (6.19)

$$\int_0^1 \left\{ \Gamma(1,t) - 2\int_0^1 \Gamma(x,t)dx \right\}^2 dt = \frac{1}{3}\sum_{j=1}^{\infty} \lambda_j Z_j^2.$$

Thus (3.7) is proven.

Proof of part (ii): The proof follows from several lemmas. It is assumed throughout that H_A holds, i.e. $||\beta|| > 0$. The argument relies on Lemma 6.1 whose proof follows from the relevant definitions, and so is omitted.

Lemma 6.1. *Suppose $\{X_n\}$ and $\{q_n\}$ are sequences of random variables. Suppose further that $\{X_n\}$ diverges to infinity in probability and $\{q_n\}$ is bounded in probability, i.e. for each M, $\lim_{n\to\infty} \mathrm{P}(X_n > M) = 1$ and for each $\varepsilon > 0$, there are M and n_0 such that $\mathrm{P}(q_n > M) < \varepsilon$, if $n > n_0$. Then*

$$\lim_{n\to\infty} \mathrm{P}(X_n > q_n) = 1.$$

Relation (3.8) now follows from Lemmas 6.2 and 6.3.

Lemma 6.2. *The statistic $\widehat{\Lambda}_N$ defined by (3.7) satisfies $\widehat{\Lambda}_N \xrightarrow{\mathrm{P}} \infty$.*

Proof. Decompose $\hat{\beta}(t)$ as

$$\hat{\beta}(t) = \beta(t) + G_N(t), \tag{6.20}$$

where

$$G_N(t) = \frac{1}{2} A_N \sum_{k=1}^{N} (2k - N - 1)\varepsilon_k(t).$$

Observe that $G_N(t)$ is equal to the estimator $\hat{\beta}(t)$ under H_0. Therefore, by (6.17),

$$N^{3/2} G_N(t) \xrightarrow{\mathcal{L}} 6\left\{ \Gamma(1,t) - 2\int_0^1 \Gamma(x,t)dx \right\} \stackrel{def}{=} U(t).$$

Consequently, as $N \to \infty$

$$N^3 \int \hat{\beta}^2(t)dt = \int \left\{ N^{3/2}\beta(t) + N^{3/2} G_N(t) \right\}^2 dt \sim \int \left\{ N^{3/2}\beta(t) + U(t) \right\}^2 dt \xrightarrow{\mathrm{P}} \infty.$$

More precisely,

$$N^{-3}\widehat{\Lambda}_N \sim \frac{1}{12}\int \left\{ \beta(t) + N^{-3/2} U(t) \right\}^2 dt \xrightarrow{\mathrm{P}} \frac{1}{12}\int \beta^2(t)dt.$$

□

Lemma 6.3. *Under H_0, the sequence $\{\Lambda_N\}$ defined by (3.9) is bounded in probability.*

Proof. Since the $\hat{\lambda}_j$ are fixed in the generation of the replications in the Monte Carlo test, the variables Z_j are independent of the $\hat{\lambda}_j$. Therefore, since $\mathrm{E}Z_j^2 = 1$,

$$\mathrm{E}\Lambda_N = \sum_{j=1}^{N} \mathrm{E}\hat{\lambda}_j.$$

The definition of the $\hat{\lambda}_j$ as the eigenvalues of the covariance operator with $\hat{c}_\varepsilon(\cdot,\cdot)$ defined by (3.5) and (3.6) implies that

$$\sum_{j=1}^{N} \hat{\lambda}_j = \frac{1}{N}\sum_{n=1}^{N} \|\hat{\varepsilon}_n\|^2.$$

This is the decomposition of functional sample variance, see details Horváth and Kokoszka (2012), p. 40. Therefore, if we can show that

$$\limsup_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} \mathrm{E}\|\hat{\varepsilon}_n\|^2 < \infty, \tag{6.21}$$

then we can conclude that $\limsup_{N\to\infty} \mathrm{E}\Lambda_N < \infty$, which in turn implies that the sequence $\{\Lambda_N\}$ is bounded in probability.

The decomposition

$$\hat{\varepsilon}_n(t) = \varepsilon_n(t) + \{\alpha(t) - \hat{\alpha}(t)\} + n\{\beta(t) - \hat{\beta}(t)\}, \tag{6.22}$$

implies that for some constant C,

$$\|\hat{\varepsilon}_n\|^2 \le C(\|\varepsilon_n\|^2 + \|\hat{\alpha}-\alpha\|^2 + \|n(\hat{\beta}-\beta)\|^2). \tag{6.23}$$

First note that

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}\|\varepsilon_n\|^2 &= \mathrm{E}\left[\int\left\{\frac{1}{N}\sum_{n=1}^{N}\varepsilon_n^2(t)\right\}dt\right]\\ &= \int\left\{\mathrm{E}\left[\frac{1}{N}\sum_{n=1}^{N}\varepsilon_n^2(t)\right]\right\}dt\\ &= \int \mathrm{E}\varepsilon_1^2(t)dt = \mathrm{E}\|\varepsilon_1\|^2 < \infty.\end{aligned}$$

Next, observe that

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}\|\hat{\alpha}-\alpha\|^2 &= \mathrm{E}\|\hat{\alpha}-\alpha\|^2\\ &= \int\{\mathrm{E}[\hat{\alpha}(t)-\alpha(t)]^2\}dt\\ &= \int \mathrm{E}\left[\frac{2}{N(N-1)}\sum_{k=1}^{N}(2N+1-3k)\varepsilon_k(t)\right]^2 dt\\ &= \frac{2(2N+1)}{N(N-1)}\mathrm{E}\|\varepsilon_1\|^2 \to 0.\end{aligned}$$

Similarly, with $H_N = \frac{(N+1)(2N+1)}{6}$,

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}\|n(\hat{\beta}-\beta)\|^2 &= H_N\mathrm{E}\|\hat{\beta}-\beta\|^2\\ &= H_N\int\left\{\mathrm{E}\left[\hat{\beta}(t)-\beta(t)\right]^2\right\}dt\\ &= H_N\int \mathrm{E}\left[\frac{6}{N(N-1)(N+1)}\sum_{k=1}^{N}(2k-N-1)\varepsilon_k(t)\right]^2 dt\\ &= \frac{2(2N+1)}{N(N-1)}\mathrm{E}\|\varepsilon_1\|^2 \to 0.\end{aligned}$$

Thus (6.21) holds. Therefore $\sup_N \mathrm{E}\Lambda_N =: C_\Lambda < \infty$, and so $\mathrm{P}(\Lambda_N > M) \le M^{-1}C_\Lambda$ can be made arbitrarily small by choosing M sufficiently large. The conclusion follows. □

6.2. Proof of Theorem 3.2

Proof of part (1): Under H_0, by (6.14), (6.16) and consistency of estimated eigenfunctions $\hat{v}_j$, ($\hat{v}_j \xrightarrow{\mathrm{P}} v_j$),

$$\begin{aligned}\left\langle\frac{N^{\frac{3}{2}}}{6}\hat{\beta},\hat{v}_j\right\rangle^2 &\xrightarrow{\mathcal{L}} \left\langle\Gamma(1,\cdot)-2\int_0^1\Gamma(x,\cdot)dx, v_j\right\rangle^2\\ &= \left\langle\sum_{k=1}^{\infty}\sqrt{\lambda_k}W_k(1)v_k - 2\int_0^1\sum_{k=1}^{\infty}\sqrt{\lambda_k}W_k(x)v_k, v_j\right\rangle^2\\ &= \left[\sum_{k=1}^{\infty}\sqrt{\lambda_k}\left\{W_k(1)-2\int_0^1 W_k(x)dx\right\}\langle v_k,v_j\rangle\right]^2\\ &= \lambda_j\left\{W_j(1)-2\int_0^1 W_j(x)dx\right\}^2\\ &= \lambda_j D_j^2 = \frac{1}{3}\lambda_j Z_j^2,\end{aligned}$$

with the random variables D_j defined in (6.18), and Z_j standard normal variables. It follows that

$$\widehat{T}_N = \frac{N^3}{12}\sum_{j=1}^{q}\hat{\lambda}_j^{-1}\langle\hat{\beta},\hat{v}_j\rangle^2 = 3\sum_{j=1}^{q}\hat{\lambda}_j^{-1}\left\langle\frac{N^{\frac{3}{2}}}{6}\hat{\beta},\hat{v}_j\right\rangle^2 \xrightarrow{\mathcal{L}} \sum_{j=1}^{q} Z_j^2 \stackrel{\mathcal{L}}{=} \chi_q^2.$$

Proof of part (ii): We must show that $\widehat{T}_N \xrightarrow{\mathrm{P}} \infty$, if $\langle\beta, v_j\rangle \neq 0$ for some $1 \leq j \leq q$. It is enough to show that

$$\sum_{j=1}^{q}\hat{\lambda}_j^{-1}\langle\hat{\beta},\hat{v}_j\rangle^2 \xrightarrow{\mathrm{P}} \sum_{j=1}^{q}\lambda_j^{-1}\langle\beta, v_j\rangle^2,$$

because the right–hand side is positive. The verification of the above convergence reduces to

$$\|\hat{\beta}-\beta\| \xrightarrow{\mathrm{P}} 0 \tag{6.24}$$

and, for $1 \leq j \leq q$,

$$\|\hat{v}_j - v_j\| \xrightarrow{\mathrm{P}} 0, \qquad \hat{\lambda}_j \xrightarrow{\mathrm{P}} \lambda_j. \tag{6.25}$$

To prove relation (6.24), observe first that by decomposition (6.20),

$$\mathrm{E}\|\hat{\beta}-\beta\| = \mathrm{E}\|G_N\| \leq \{\mathrm{E}\|G_N\|^2\}^{\frac{1}{2}} = \left\{\mathrm{E}\int G_N^2(t)dt\right\}^{\frac{1}{2}}.$$

To calculate the last expected value, we will use the identity

$$\frac{1}{4}A_N\sum_{k=1}^{N}(2k-N-1)^2 = 1,$$

which follows from algebraic manipulations. The independence of the ε_k thus implies that

$$\mathrm{E}\int G_N^2(t)dt = \frac{1}{4}A_N^2\sum_{k=1}^{N}(2k-N-1)^2\mathrm{E}\int\varepsilon_k^2(t)dt = A_N\mathrm{E}\|\varepsilon\|^2 = \mathcal{O}(N^{-3}).$$

By Lemmas 2.2. and 2.3 of Horváth and Kokoszka (2012), relations (6.25) will follow from $\|\hat{c}_\varepsilon - c_\varepsilon\|_{\mathcal{S}} \xrightarrow{\mathrm{P}} 0$, where the subscript $\mathcal{S}$ denotes the Hilbert–Schmidt norm. Proposition 6.1 states that, in fact, $\mathrm{E}\|\hat{c}_\varepsilon - c_\varepsilon\|_{\mathcal{S}}^2 = \mathcal{O}(N^{-1})$. It thus extends a well–known result, e.g. Theorem 2.5. of Horváth and Kokoszka (2012), which states that

$$\mathrm{E}\int\left(\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i(t)\varepsilon_i(s) - \mathrm{E}[\varepsilon(t)\varepsilon(s)]\right)^2 dtds = \mathcal{O}(N^{-1}). \tag{6.26}$$

The covariance function $\hat{c}_\varepsilon$ is defined in terms of the residuals $\hat{\varepsilon}_n$, cf. (3.5) and (3.6). Estimation of the intercept and slope functions introduces many additional terms which are, however, all asymptotically negligible. This is the content of the following proposition whose proof is very long as it requires the examination of 16 cross–terms. The proof is therefore not presented her, but is available upon request.

Proposition 6.1. *Suppose model (3.2) holds and* $\mathrm{E}\|\varepsilon\|^4 < \infty$. *Then the sample covariance function* $\hat{c}_\varepsilon$, *defined by (3.5) and (3.6), satisfies* $\mathrm{E}\|\hat{c}_\varepsilon - c_\varepsilon\|_{\mathcal{S}}^2 = \mathcal{O}(N^{-1})$.

Acknowledgment

The authors gratefully acknowledge financial support from the Deutsche Forschungsgemeinschaft via the International Research Training Group IRTG 1792 “High Dimensional Non Stationary Time Series” and the Collaborative Research Center CRC 649 “Economic Risk”, Humboldt-Universität zu Berlin. Kokoszka and Xiong were partially supported by the National Science Foundation grant DMS–1462067 “FRG: Collaborative Research: Extreme Value Theory for Spatially Indexed Functional Data”

References

Berkes, I., Gabrys, R., Horváth, L., Kokoszka, P., 2009. Detecting changes in the mean of functional observations. J. R. Stat. Soc. (B) 71, 927–946.
Berkes, I., Horváth, L., Rice, G., 2013. Weak invariance principles for sums of dependent random functions. Stoch. Process. Appl. 123, 385–403.
Bosq, D., 2000. Linear Processes in Function Spaces. Springer.
Brodsky, B.E., Darkhovsky, B.S., 1993. Nonparametric Methods in Change–Point Problems. Kluwer.
Chen, J., Gupta, A.K., 2011. Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance. Birkhäuser.
Csörgő, M., Horváth, L., 1997. Limit Theorems in Change-Point Analysis. Wiley.
Dierckx, G., Teugels, J.L., 2010. Change point analysis of extreme values. Environmetrics 21, 661–686.
Efron, B., 1991. Regression percentiles using asymmetric squared loss. Statistica Sinica 1, 93–125.

Eilers, P.H.C., Marx, B.D., 1996. Flexible smoothing with B–splines and penalties. Stat. Sci. 11, 89–121.
Ferraty, F., Vieu, P., 2006. Nonparametric Functional Data Analysis: Theory and Practice. Springer.
Fraiman, R., Justel, A., Liu, R., Llop, P., 2014. Detecting trends in time series of functional data: a study of antarctic climate change. Can. J. Stat. 42, 597–609.
Gromenko, O., Kokoszka, P., 2013. Nonparametric inference in small data sets of spatially indexed curves with application to ionospheric trend determination. Comput. Stat. Data Anal. 59, 82–94.
Guo, M., Härdle, W.K., 2012. Simultaneous confidence bands for expectile functions. AStA Adv. Stat. Anal. 96, 517–541.
Guo, M., Zhou, L., Huang, J.Z., Härdle, W.K., 2015. Functional data analysis of generalized regression quantiles. Stat. Comput. 25, 189–202.
Hörmann, S., Kidziński, L., Hallin, M., 2015. Dynamic functional principal components. J. R. Stat. Soc.(B) 77, 319–348.
Hörmann, S., Kokoszka, P., 2010. Weakly dependent functional data. Annals Stat. 38, 1845–1884.
Horváth, L., Kokoszka, P., 2012. Inference for Functional Data with Applications. Springer.
Horváth, L., Kokoszka, P., Reeder, R., 2013. Estimation of the mean of functional time series and a two sample problem. J. R. Stat. Soc. (B) 75, 103–122.
Hsing, T., Eubank, R., 2015. Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators. Wiley.
Kossin, J.P., Olander, T.L., Knapp, K.R., 2013. Trend analysis with a new global record of tropical cyclone intensity. J. Climate 26, 9960–9976.
Newey, W.K., Powell, J.L., 1987. Asymmetric least squares estimation and testing. Econometrica 55, 819–847.
Ramsay, J.O., Silverman, B.W., 2005. Functional Data Analysis. Springer.
Rossi, G.D., Harwey, A., 2009. Quantiles, expectiles and splines. J. Econom. 152, 179–185.
Schnabel, S.K., 2011. Expectile Smoothing: New Perspectives on asymmetric Least Squares. An Application to Life Expectancy, Ph.D. thesis. Utrecht University.
Schnabel, S.K., Eilers, P.H.C., 2009. Optimal expectile smoothing. Comput. Stat. Data Anal. 53, 4168–4177.
Schnabel, S.K., Eilers, P.H.C., 2013. Simultaneous estimation of quantile curves using quantile sheets. AStA Adv. Stat. Anal. 97, 77–87.
Smith, R.L., 1989. Extreme value analysis of environmental time series: An application to trend detection in ground–level ozone. Stat. Sci. 4, 367–377.
Sobotka, F., Kauermann, G., Schulze-Waltrup, L., Kneib, T., 2013. On confidence intervals for semiparametric expectile regression. Stat. Comput. 23, 135–148.
Sobotka, F., Schnabel, S., Schulz-Waltrup, L., Eilers, P., Kneib, T., Kauermann, G., 2014. R package: expectreg. Version: 0.39, published: March 05, 2014.
Taylor, J., 2008. Estimating value at risk and expected shortfall using expectiles. J. Financ. Econom. 6, 231–252.
Torgovitsky, L., 2016. Hilbert Space valued signal Plus Noise Models: Analysis of Structural Breaks Under High Dimensionality and Temporal Dependence. Ph.D. thesis. Universität zu Köln.
UNISYS, 2015. Data in atlantic and west pacific. UNISYS Weather Information Systems, http://weather.unisys.com/hurricane/index.php, Accessed: February 20, 2015.

Journal of Business & Economic Statistics

ISSN: 0735-0015 (Print) 1537-2707 (Online) Journal homepage: http://amstat.tandfonline.com/loi/ubes20

Analysis of Deviance for Hypothesis Testing in Generalized Partially Linear Models

Wolfgang Karl Härdle & Li-Shan Huang

To cite this article: Wolfgang Karl Härdle & Li-Shan Huang (2017): Analysis of Deviance for Hypothesis Testing in Generalized Partially Linear Models, Journal of Business & Economic Statistics, DOI: 10.1080/07350015.2017.1330693

To link to this article: https://doi.org/10.1080/07350015.2017.1330693

Analysis of Deviance for Hypothesis Testing in Generalized Partially Linear Models

Wolfgang Karl HÄRDLE
Center for Applied Statistics and Economics, Humboldt University 10099, Berlin, Germany (*haerdle@wiwi.hu-berlin.de*)

Li-Shan HUANG
Institute of Statistics, National Tsing Hua University, 30013, Taiwan (*lhuang@stat.nthu.edu.tw*)

In this study, we develop nonparametric analysis of deviance tools for generalized partially linear models based on local polynomial fitting. Assuming a canonical link, we propose expressions for both local and global analysis of deviance, which admit an additivity property that reduces to analysis of variance decompositions in the Gaussian case. Chi-square tests based on integrated likelihood functions are proposed to formally test whether the nonparametric term is significant. Simulation results are shown to illustrate the proposed chi-square tests and to compare them with an existing procedure based on penalized splines. The methodology is applied to German Bundesbank Federal Reserve data.

KEY WORDS: ANOVA decomposition; Integrated likelihood; Local polynomial regression.

1. INTRODUCTION

Generalized linear models (McCullagh and Nelder 1989) are a large class of statistical models for relating a response variable to linear combinations of predictor variables. The models allow the response variable to follow probability distributions in the exponential family such as the Binomial and Poisson, generalizing the Gaussian distribution in linear models, though a major limitation is the prespecified linear form of predictors. Generalized partially linear models (Green and Silverman 1994; Carroll et al. 1997; Härdle et al. 2004) allow for a nonparametric component for a continuous covariate while retaining the ease of linear relationships for the remaining variables. It is more flexible than the conventional linear approach and is a special case of generalized additive models (Hastie and Tibshirani 1990; Wood 2006) which include multiple nonparametric components. Härdle, Mammen, and Müller (1998) applied the generalized partially linear model to 1991 East–West German migration data to model the probability of migration with a nonlinear relationship to household income and linear relationships to other covariates such as age, gender, and employment status. Wood (2006, p. 248) gave an example of modeling the daily total deaths in Chicago in the period 1987–2000 as a Poisson distribution with a nonlinear trend of time and linear effects of daily temperature and daily air-pollution levels of ozone, sulfur dioxide, and pm10. An illustrating finance example in Section 6 of this article is on bankruptcy prediction for firms, known as rating or scoring, from a set of financial ratio variables. The logistic partially linear model is used to model the probability of default with a nonlinear relationship to the account payable turnover ratio, which is a short-term liquidity measure, and linear relationships to some selected financial ratios.

In applying generalized partially linear models to data, inference tools to examine whether the nonparametric term is significant are of interest. For example, in Härdle, Mammen, and Müller (1998), the nonlinear estimated function of household income showed a saturation in the intention to migrate for higher income households and the question was whether the overall income effect was significant statistically. As analysis of deviance was developed for generalized linear models (McCullagh and Nelder 1989), it is natural to ask whether one can extend it for generalized partially linear models. Though Hastie and Tibshirani (1990) briefly discussed analysis of deviance for generalized additive models, they noted that "the distribution theory, however, is undeveloped" and "informal deviance tests with some heuristic justification" were adopted. The present article fills the gap by establishing local and global analysis of deviance expressions for generalized partially linear models and developing associated tests for checking whether the nonparametric term is significant. Li and Liang (2008) addressed assessing the significance of the nonparametric term in the local polynomial settings by extending generalized likelihood ratio tests (Fan, Zhang, and Zhang 2001), which have asymptotic chi-square distributions. Wood (2013) discussed approximate p-values for testing significance of smooth components of semiparametric generalized additive models by Wald-type tests based on penalized splines. We remark that testing in the generalized partially linear models is relatively less developed as compared to the special case of partially linear models under the Gaussian distribution (Härdle et al. 2004). Hence, there is a need for developing analysis of

Journal of Business & Economic Statistics
XXXX 2017, Vol. 0, No. 0
DOI: 10.1080/07350015.2017.1330693
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/jbes.

deviance tools parallel to those in generalized linear models for applications of generalized partially linear models.

Based on the local polynomial approach (Fan and Gijbels 1996) and assuming a canonical link in generalized partially linear models, we propose local and global expressions for analysis of deviance, with the latter obtained by integrating the corresponding local likelihood quantities. This mimics the "integrated likelihood" approach discussed by Lehmann (2006) and Severini (2007). Though the idea of local likelihood has been around for some time (Hastie and Tibshirani 1987; Loader 1999), we are not aware of using the integrated likelihood approach to combine the information of local likelihood in the smoothing literature. Then, integrated likelihood ratio tests with asymptotic chi-square distributions are proposed to check whether the nonparametric term is significant. Our work extends the classic analysis of deviance to generalized partially linear models with theoretical justifications, and generalizes the work of Huang and Chen (2008) and Huang and Davidson (2010) in a special case of the Gaussian distribution to distributions in the canonical exponential family.

The organization of this article is as follows. Section 2 outlines the analysis of deviance for nested hypotheses in parametric generalized linear models by Simon (1973). In Section 3, we propose local and global analysis of deviance for nonparametric models in Theorem 1 for the simpler case with the nonparametric term as the only predictor. By combining local likelihood through integration, as a by-product, new estimators for the canonical parameter and response mean are given in equation (12), and Theorem 2 shows that the integrated likelihood quantities are asymptotically global likelihood quantities with the new estimators. Theorem 3 proposes integrated likelihood ratio tests with asymptotic chi-square distributions for testing whether the nonparametric term is significant. Section 4 presents an extension of Theorems 1–3 to generalized partially linear models as Theorems 4 and 5. In Section 5, we illustrate the potential usefulness of the new tests with simulated data and compare with tests by Wood (2013) in the R package `mgcv`. Section 6 applies the methodology to 2002 German Bundesbank Federal Reserve data and Section 7 gives some concluding remarks and directions for future research.

2. PRELIMINARIES

We first describe generalized linear models based on McCullagh and Nelder (1989). Let $(x_1, y_1), \ldots, (x_n, y_n)$ be independent data pairs with the conditional density of Y given covariate $X = x$ from a one-parameter exponential family:

$$L(y; \theta(x)) = \exp\left[\frac{y\theta(x) - b\{\theta(x)\}}{a(\phi)} + c(y, \phi)\right], \qquad (1)$$

where $a(\cdot) > 0$, $b(\cdot)$, and $c(\cdot)$ are known functions, ϕ is known or a nuisance parameter, and θ is the canonical parameter with the conditional mean of response, $\mathbf{E}(Y \mid X = x) = \mu(x) = b'\{\theta(x)\}$. A transformation of mean $G\{\mu(x)\}$ may be modeled linearly by $G\{\mu(x)\} = b_0 + b_1 x$, where $G(\cdot)$ is called the "link" function and estimates of b_0 and b_1 are obtained by maximum likelihood. If $G(\cdot) = (b')^{-1}(\cdot)$, then G is the canonical link function that links θ to the linear predictor. For simplicity, G is the canonical link function throughout the article and the dependence of θ on covariates is often suppressed if no ambiguities result.

Let $\ell(y; \theta) = \log L(y; \theta)$, $\hat{\theta} = G(\hat{\mu})$ denote the fitted value of θ with corresponding $\hat{\mu}$, and $\tilde{\theta} = G(y)$ when the fitted value equals the observed y. The deviance D (McCullagh and Nelder 1989), measuring the discrepancy between data $\mathbf{y} = (y_1, \ldots, y_n)^\top$ and fitted $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)^\top$, is

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2\sum_i \{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\}. \qquad (2)$$

In the Gaussian case, G is the identity link and $D = \sum_i (y_i - \hat{\mu}_i)^2$, which is the residual sum of squares in linear models. Let us now turn to testing hypotheses about $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$. Assume that $D_j = \inf_{\boldsymbol{\theta} \in \Omega_j} D$, $j = 1, 2$, with $\Omega_2 \subseteq \Omega_1$. The analysis of deviance usually refers to comparing two nested parametric models and inference may be based on the difference $D_2 - D_1$, which is simply the log-likelihood ratio statistic with an asymptotic χ^2 distribution. The conventional analysis of deviance is generally not parallel to the analysis of variance in linear models, in the sense that the former does not have all the sum-of-squares quantities.

An attempt to mimic analysis of variance for (1) can be based on the Kullback–Leibler (KL) divergence of two probability distributions with means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$:

$$\mathrm{KL}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = 2\mathbf{E}_{\boldsymbol{\mu}_1}\left[\ell\{\mathbf{y}; G(\boldsymbol{\mu}_1)\} - \ell\{\mathbf{y}; G(\boldsymbol{\mu}_2)\}\right],$$

where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are treated as fixed values and $\mathbf{E}_{\boldsymbol{\mu}_1}$ is the conditional expectation with respect to $\mathbf{y}$ with $\boldsymbol{\mu} = \boldsymbol{\mu}_1$. Simon (1973) showed that for nested hypotheses $\Omega_2 \subset \Omega_1 \subset \mathbb{R}^n$ with $\mathbb{R}^n$ corresponding to the parameter space for an exact fit of $\tilde{\theta}$ and θ parameterized linearly in Ω_1 and Ω_2,

$$\mathrm{KL}(\mathbf{y}, \boldsymbol{\mu}_2) = \mathrm{KL}(\mathbf{y}, \boldsymbol{\mu}_1) + \mathrm{KL}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) \qquad (3)$$

when plugging in the values of maximum likelihood estimates for $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. In other words, (3) shows that the KL divergence exhibits the Pythagorean property. For the Gaussian distribution, (3) reduces to the analysis of variance decomposition in linear models when $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ correspond to the linear fit and the intercept-only model, respectively, and the terms in (3) becomes total, residual, and regression sums of squares, respectively.

A linear form of x may be restrictive and one may consider a nonparametric approach:

$$G\{\mu(x)\} = m(x). \qquad (4)$$

Fan, Heckman, and Wand (1995) discussed estimating $m(\cdot)$ by maximizing a locally weighted likelihood with a local polynomial approximation. Based on Taylor's expansion at x, $\theta_i \approx \beta_0 + \beta_1(x_i - x) + \cdots + \beta_p(x_i - x)^p \equiv \theta_i(x)$. This approximation is plugged in the locally weighted log-likelihood at x,

$$\ell_x(\mathbf{y}; \boldsymbol{\theta}_x) \equiv \sum_i \ell\{y_i; \theta_i(x)\}K_h(x_i - x), \qquad (5)$$

where $\boldsymbol{\theta}_x = (\theta_1(x), \ldots, \theta_n(x))^\top$, $K(\cdot)$ is usually a density function being symmetric at 0, h is the bandwidth determining the neighborhood size, and $K_h(\cdot) = K(\cdot/h)/h$. Then, $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)^\top$ maximizing $\ell_x(\mathbf{y}; \boldsymbol{\theta}_x)$ is solved and $j!\hat{\beta}_j$ estimates $m^{(j)}(x)$, $j = 0, \ldots, p$, which is $\theta^{(j)}(x)$ with the canonical link. Fan, Heckman, and Wand (1995) derived asymptotic properties of $\hat{\beta}_j(x)$'s and adopted $G^{-1}\{\hat{\beta}_0(x)\}$ as an estimate for $\mu(x)$.

A further extension to generalized partially linear models is

$$G\{\mu(\mathbf{z}, x)\} = \mathbf{z}^\top \alpha + m(x), \tag{6}$$

where $\mathbf{z}$ is a K-dimensional covariate vector. Without loss of generality, the intercept in (6) is embedded in $m(\cdot)$. When α is unknown, estimation of α can be done via a two-step maximum likelihood procedure that updates the linear and nonparametric estimates iteratively, as discussed by Carroll et al. (1997), p. 479.

3. NONPARAMETRIC ANALYSIS OF DEVIANCE

This section focuses on (4). We start by deriving a local analysis of deviance expression for model (4) in the following by adapting (3) for locally weighted likelihood. Let $\hat\theta_i(x) = \hat\beta_0 + \cdots + \hat\beta_p(x_i - x)^p$, the resulting local polynomial estimate of θ_i at x, $\hat{\boldsymbol\theta}_x = (\hat\theta_1(x), \ldots, \hat\theta_n(x))^\top$, $\hat\mu_x(x_i) = G^{-1}\{\hat\theta_i(x)\}$, and $\hat{\boldsymbol\mu}_x = (\hat\mu_x(x_1), \ldots, \hat\mu_x(x_n))^\top$. As the $\hat\beta_j$'s maximize $\ell_x(\mathbf{y}; \boldsymbol\theta_x)$, the following equations hold:

$$\begin{aligned}\sum_i y_i(x_i - x)^j K_h(x_i - x) &= \sum_i \hat\mu_x(x_i)(x_i - x)^j K_h(x_i - x), \\ &\qquad j = 0, \ldots, p, \\ \sum_i y_i \hat\theta_i(x) K_h(x_i - x) &= \sum_i \hat\mu_x(x_i)\hat\theta_i(x) K_h(x_i - x).\end{aligned} \tag{7}$$

The last equation indicates that $(\mathbf{y} - \hat{\boldsymbol\mu}_x)$ is orthogonal to $\hat{\boldsymbol\theta}_x$ in the locally weighted inner product space with weights $K_h(x_i - x)$. Hence, the fact of residuals being orthogonal to fitted values in ordinary linear models now becomes the fact of local residuals $(\mathbf{y} - \hat{\boldsymbol\mu}_x)$ being orthogonal to locally fitted canonical parameters $\hat{\boldsymbol\theta}_x$ in a kernel-weighted space. For $\ell_x(\mathbf{y}; \hat{\boldsymbol\theta}_x)$, an expression mimicking (2) for local deviance at x is therefore

$$\begin{aligned} d_x(\mathbf{y}, \hat{\boldsymbol\mu}_x) &= 2\{\ell_x(\mathbf{y}; \tilde{\boldsymbol\theta}) - \ell_x(\mathbf{y}; \hat{\boldsymbol\theta}_x)\} \\ &= 2\sum_i [y_i\{\tilde\theta_i - \hat\theta_i(x)\} - b(\tilde\theta_i) + b(\hat\theta_i(x))] K_h(x_i - x), \end{aligned} \tag{8}$$

where $\tilde{\boldsymbol\theta} = (\tilde\theta_1, \ldots, \tilde\theta_n)^\top$ with $\tilde\theta_i = G(y_i)$, same as those defined around (2). Though (8) is a natural definition for local likelihood, we are not aware of a similar quantity to (8) in the literature. Published work focuses on global deviance by taking (2) with $G^{-1}\{\hat\beta_0(x_i)\}$ as the estimate, and strictly speaking, the resulting deviance expression is not based on maximized likelihood as $\hat\beta_1, \ldots, \hat\beta_p$ are ignored. In comparison, the deviance (8) makes use of all coefficients $\hat\beta_0, \ldots, \hat\beta_p$ from maximizing local likelihood. Then, (3) is adapted to form a local analysis of deviance expression, and a global expression may be obtained by integrating local quantities, as given in the following theorem.

Theorem 1. Suppose that conditions (A1) and (A2) in the online Appendix hold. Under model (4), the following results hold when using local polynomial approximations of pth order.

(a) For a grid point x in the support of covariate X, a local analysis of deviance expression is

$$d_x(\mathbf{y}, \bar y) = d_x(\mathbf{y}, \hat{\boldsymbol\mu}_x) + d_x(\hat{\boldsymbol\mu}_x, \bar y), \tag{9}$$

where $\bar y$ is the sample mean of $\mathbf{y}$, $d_x(\mathbf{y}, \bar y)$ is (8) with $\hat{\boldsymbol\mu}_x$ and $\hat{\boldsymbol\theta}_x$ replaced by $\bar y$ and $G(\bar y)$ respectively, and

$$\begin{aligned} d_x(\hat{\boldsymbol\mu}_x, \bar y) &\equiv 2\mathbf{E}_{\hat{\boldsymbol\mu}_x}\left[\ell_x\left(\mathbf{y}; \hat{\boldsymbol\theta}_x\right) - \ell_x\{\mathbf{y}; G^{-1}(\bar y)\}\right] \\ &= 2\left[\ell_x\left(\mathbf{y}; \hat{\boldsymbol\theta}_x\right) - \ell_x\left\{\mathbf{y}; G^{-1}(\bar y)\right\}\right]. \end{aligned} \tag{10}$$

(b) A global analysis of deviance expression is obtained by integrating the local quantities in (9) over the support of covariate X:

$$\int d_x(\mathbf{y}, \bar y)dx = \int d_x(\mathbf{y}, \hat{\boldsymbol\mu}_x)dx + \int d_x(\hat{\boldsymbol\mu}_x, \bar y)dx, \tag{11}$$

where $\int d_x(\mathbf{y}, \bar y)dx = \mathrm{KL}(\mathbf{y}, \bar y) = D(\mathbf{y}, \bar y)$ under a boundary condition in (A1) that the weights $\int K_h(x_i - x)dx = 1$, $i = 1, \ldots, n$.

Theorem 1 provides elegant local and global analysis of deviance expressions that mimic the classic case (3) (McCullagh and Nelder 1989; Simon 1973) and shows that the Pythagorean property of the KL divergence holds under model (4) with local polynomial fitting. It is straightforward to show (9) based on (7) and (10) and hence the proof is omitted. Alternatively, the proof in Simon (1973) for (3) can be adapted with kernel weights to show (9). The local expression (9) has an interpretation that the null deviance at point x, $d_x(\mathbf{y}, \bar y)$, can be decomposed into two parts, the residual deviance after fitting a locally weighted polynomial at x, $d_x(\mathbf{y}, \hat{\boldsymbol\mu}_x)$, and the model deviance at x, $d_x(\hat{\boldsymbol\mu}_x, \bar y)$. Equality (9) holds in finite-sample cases, similar to (3). The global analysis of deviance (11) extends the above interpretation to a fitted curve by local polynomials: the residual deviance $\int d_x(\mathbf{y}, \hat{\boldsymbol\mu}_x)dx$ is a measure of the lack of fit of fitting (4), whereas the null deviance $\int d_x(\mathbf{y}, \bar y)dx$ is such a measure for a reduced model that only includes the intercept. The quantities in (11) are weighted integrals (see (5)), which may be approximated by the Riemann sum in practice, and an analysis of deviance table based on (11) is formed, similar to the parametric framework. For a special case of Normal distribution with an identity link, (9) and (11) become the local and global analysis of variance decompositions respectively in Huang and Chen (2008). For the boundary condition in Theorem 1(b), if $K(\cdot)$ has a support $[-1, 1]$ and $\{x_i, i = 1, \ldots, n\}$ has a range of $[a, b]$, then a boundary-corrected kernel $[\int K_h(x_i - x)dx]^{-1}K_h(x_i - x)$ may be used for x_i in $[a, a + h)$ and $(b - h, b]$ to ensure that the integrated kernel weights equals to 1.

As a by-product, the above derivations give rise to new "global" estimators for θ_i's and μ_i's:

$$\theta_i^* = \int \hat\theta_i(x)K_h(x_i - x)dx \qquad \text{and} \qquad \mu_i^* = G^{-1}(\theta_i^*). \tag{12}$$

They are different from local estimates at x_i: $\hat\beta_0(x_i)$ and $G^{-1}\{\hat\beta_0(x_i)\}$. The asymptotic properties of θ_i^* and μ_i^* for "interior" points x_i with $p = 1$ and 3 are discussed in Proposition 1. The reason $p = 1$ and 3 only is due to their simpler asymptotic bias expressions of $\hat\beta_0(x)$ than those of $p = 0$ and 2; see Theorems 1a and 1b in Fan, Heckman, and Wand (1995). The "interior" region is defined as follows. For a kernel function with support $[-1, 1]$, if the convex support of x_is is $[a, b]$,

then define the interior region as $[a + 2h, b - 2h]$. This definition is narrower than the conventional $[a + h, b - h]$, since for x_i in $[a + h, a + 2h) \bigcup (b - 2h, b - h]$, the corresponding θ_i^* and μ_i^* in (12) involve $\hat{\beta}_j(x)$ with x in $[a, a + h) \bigcup (b - h, b]$, $j = 0, \ldots, p$.

Proposition 1. Suppose that conditions (A1)–(A5) in the online Appendix hold. Assume that $h \to 0$ and $nh^3 \to \infty$ as $n \to \infty$. Then for interior points x_i with $p = 1$ and 3,

(a) the order of the asymptotic bias of θ_i^* is smaller than the conventional order $h^{(p+1)}$; that is, the $h^{(p+1)}$ term of the bias of θ_i^* is zero;
(b) the asymptotic variance of θ_i^* is of order $n^{-1}h^{-1}$;
(c) similarly, the order of the bias of μ_i^* is smaller than the conventional order $h^{(p+1)}$ and the asymptotic variance of μ_i^* is of order $n^{-1}h^{-1}$.

The proof for the Proposition is given in the online Appendix. There has been some research aimed at finding new ways of reducing bias of basic kernel smoothers, for example, Kosmidis and Firth (2009). In the Gaussian case with an identity link, Huang and Chan (2014) showed that the bias of θ_i^* for interior points is of order $h^{2(p+1)}$ for $p = 0, 1, 2, 3$, which is consistent with intuition that the higher the p, the smaller the order of the bias. The derivation of explicit bias expressions of θ_i^* in exponential family is technically challenging, since the second-order expansions of the bias of $\hat{\beta}_j(x)$ for (1) with (4) have not been addressed in the literature. We thus focus on analysis of deviance, while the issue of bias reduction may be studied in a future article.

Theorem 1 involves integrating local likelihood quantities to form a global analysis of deviance expression and hence it is of interest to explore how integrated local likelihood $\int \ell_x(\mathbf{y}, \hat{\boldsymbol{\theta}}_x)dx$ behaves as a global likelihood function. The following theorem shows that integrated local likelihood is asymptotically a global likelihood $\ell(\mathbf{y}; \boldsymbol{\theta}^*)$ with estimate $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_n^*)^\top$ and that the integrated deviance quantities $\int d_x(\mathbf{y}, \hat{\boldsymbol{\mu}}_x)dx$ and $\int d_x(\hat{\boldsymbol{\mu}}_x, \bar{y})dx$ are asymptotically KL-divergence measures with estimate $\boldsymbol{\mu}^* = (\mu_1^*, \ldots, \mu_n^*)^\top$.

Theorem 2. Under model (4), assume that conditions (A1)–(A5) in the online Appendix hold, and $h \to 0, nh^3 \to \infty$ as $n \to \infty$. For $p = 1$ and 3,

(a) the integrated likelihood function is asymptotically

$$\int \ell_x(\mathbf{y}; \hat{\boldsymbol{\theta}}_x)dx = \ell(\mathbf{y}; \boldsymbol{\theta}^*) + \mathcal{O}\left(h^{(p+1)}\right), \tag{13}$$

where the elements of $\boldsymbol{\theta}^*$ are defined in (12);
(b) the integrated deviance quantities are asymptotically

$$\int d_x(\mathbf{y}, \hat{\boldsymbol{\mu}}_x)dx = \mathrm{KL}(\mathbf{y}, \boldsymbol{\mu}^*) + \mathcal{O}\left(h^{(p+1)}\right) \quad \text{and}$$

$$\int d_x(\hat{\boldsymbol{\mu}}_x, \bar{y})dx = \mathrm{KL}(\boldsymbol{\mu}^*, \bar{y}) + \mathcal{O}\left(h^{(p+1)}\right), \tag{14}$$

where the elements of $\boldsymbol{\mu}^*$ are defined in (12);
(c) from (11) and (14),

$$\mathrm{KL}(\mathbf{y}, \bar{y}) = \mathrm{KL}\,(\mathbf{y}, \boldsymbol{\mu}^*) + \mathrm{KL}(\boldsymbol{\mu}^*, \bar{y}) + \mathcal{O}\left(h^{(p+1)}\right), \tag{15}$$

which shows that the classic analysis of deviance holds asymptotically with $\boldsymbol{\mu}^*$.

The proof of Theorem 2 is given in the online Appendix and it uses some results stated in the Proposition for $p = 1$ and 3. Hence Theorem 2 is limited to $p = 1$ and 3 only for the same reason described before Proposition 1. The integrated local likelihood $\int \ell_x(\mathbf{y}, \hat{\boldsymbol{\theta}}_x)dx$ in (13) is a weighted integral of local likelihood with fitted local polynomials. In the literature, the idea of integrated likelihood was mentioned in Lehmann (2006), and Severini (2007) discussed integrated likelihood functions to eliminate nuisance parameters in parametric settings. To our knowledge, (13) and (15) have never been raised in the nonparametric regression literature. The convention was to plug in $\hat{\beta}_0(x_i)$ in (1) for θ_i; as $\hat{\beta}_0(x_i)$'s are not maximum likelihood estimates globally, the KL-type additivity (3) would not hold. In contrast, (15) shows that the classic analysis of deviance holds asymptotically by using the local additivity in (9). Two topics for further investigation are to apply Theorems 1 and 2 to develop bandwidth selection and residual diagnostic procedures. For example, bandwidth selection may be based on cross-validating the deviance or minimizing the corrected Akaike information criterion (AICc; Hurvich, Simonoff, and Tsai 1998), both with close connection to KL divergence. In Section 5, we explore adapting the AICc criterion with the integrated deviance for bandwidth selection empirically.

Based on integrated local likelihood, we next develop an integrated likelihood ratio test for examining the significance of a nonparametric fit, parallel to chi-square tests in parametric settings (McCullagh and Nelder 1989). Under model (4), the intercept term is embedded in $m(\cdot)$ and hence testing significance of $m(\cdot)$ becomes testing whether $m(\cdot)$ equals to a constant.

Theorem 3. Under the conditions of Theorem 2, for testing $H_0 : m(x) = a_0$ with a_0 a constant versus $H_a : m(x)$ is not a constant function, when estimating $m(\cdot)$ by pth order local polynomials with $p \geq 0$, the test statistic

$$2\left\{\int \ell_x(\mathbf{y}; \hat{\boldsymbol{\theta}}_x)dx - \ell(\mathbf{y}; \hat{a}_0)\right\} \tag{16}$$

is asymptotically distributed according to a χ^2-distribution with degrees of freedom (df) $\mathrm{tr}(H_p^*) - 1$, where $\hat{a}_0$ is the maximum likelihood estimate under H_0 and H_p^* is the smoothing matrix for local pth order polynomial regression defined in Huang and Chen (2008) in the case of the Normal distribution.

More explicitly, H_p^* depending on x_i's, bandwidth h, and the kernel function $K(\cdot)$, is

$$H_p^* = \int W X_p \left(X_p^\top W X_p\right)^{-1} X_p^\top W dx, \tag{17}$$

where W is an n-dimensional diagonal matrix with $K_h(x_i - x)$ as its diagonal elements, and X_p is the $n \times (p + 1)$ design matrix with the $(j + 1)$th column $((x_1 - x)^j, \ldots, (x_n - x)^j)^\top$, $j = 0, \ldots, p$. The dependence of W and X_p on x is suppressed and the integration in (17) is performed element by element in the resulting matrix product. In Theorem 3, the χ^2-distribution is allowed to have a non integer degree of freedom, since the χ^2-distribution is a special case of the gamma distribution. The asymptotic order of $\mathrm{tr}(H_1^*)$ in the case of local linear regression $p = 1$ is of order h^{-1} (Huang and Chen 2008, p. 2093). We name the χ^2-test in Theorem 3 as an integrated likelihood ratio test since the test statistic can be expressed as integrated likelihood

ratio:

$$\int \ell_x(\mathbf{y}; \hat{\boldsymbol{\theta}}_x)dx - \ell(\mathbf{y}; \hat{a}_0) = \int \sum_i \left[\ell\{y_i; \hat{\theta}_i(x)\} - \ell(y_i; \hat{a}_0)\right] K_h(x_i - x)dx$$

under the boundary condition in (A1). In other words, under model (4), the test statistic (16) integrates the differences in local deviances between a nonparametric fit (8) and an intercept-only reduced model and it is distributed asymptotically as a chi-squared distribution with the difference in degrees of freedom of the two models. This interpretation makes (16) more compelling than the generalized likelihood ratio test in Li and Liang (2008), since their work does not have a connection to deviance.

4. ANALYSIS OF DEVIANCE FOR PARTIALLY LINEAR MODELS

We extend the results in Section 3 to generalized partially linear models (6). Denote $\breve{\mu}_x(x_i) = G^{-1}\{\breve{\theta}_i(x)\}$ where $\breve{\theta}_i(x) = \mathbf{z}_i^\top \breve{\alpha} + \breve{\beta}_0 + \cdots + \breve{\beta}_p(x_i - x)^p$ with estimates $\breve{\alpha}$ and $\breve{\beta}_j$'s under (6). To avoid confusion with the notation in Section 3, from now on $\breve{\mu}_x(x_i)$, $\breve{\theta}_i(x)$, $\breve{\boldsymbol{\theta}}_x$, $\breve{\boldsymbol{\mu}}_x$, $\breve{\alpha}$, $\boldsymbol{\mu}^{**}$, and $\boldsymbol{\theta}^{**}$ denote the estimates under (6). Since $\breve{\beta}_j$'s maximize the local likelihood, the equations in (7) continue to hold with $\breve{\boldsymbol{\theta}}_x$ and $\breve{\boldsymbol{\mu}}_x$ under (6). The interpretation that $(\mathbf{y} - \breve{\boldsymbol{\mu}}_x)$ is orthogonal to $\breve{\boldsymbol{\theta}}_x$ in the locally weighted inner product space with weights $K_h(x_i - x)$ continues to hold under (6). An additional equation from estimating α by maximum likelihood is

$$\sum_i y_i z_{ik} = \sum_i z_{ik} \int \breve{\mu}_x(x_i) K_h(x_i - x)dx, \qquad k = 1, \ldots, K, \tag{18}$$

where z_{ik} denotes the value of the kth covariate for the ith observation. From (18), we observe that the column vector with entries $(y_i - \int \breve{\mu}_x(x_i)K_h(x_i - x)dx)$, $i = 1, \ldots, n$, is orthogonal to the column space spanned by $\mathbf{z}$. Moreover, it can be shown that $\int \breve{\mu}_x(x_i)K_h(x_i - x)dx = \mu_i^{**} + \mathcal{O}(h^{p+1} + n^{-1/2})$ and hence $(\mathbf{y} - \boldsymbol{\mu}^{**})$ is asymptotically orthogonal to the column space spanned by $\mathbf{z}$.

Theorems 1 and 2 are extended to generalized partially linear models (6) in the following as Theorem 4(a) and 4(b) respectively when α is estimated by maximum likelihood. We develop local and global analysis of deviance expressions for (6) in Theorem 4(a) and Theorem 4(b) shows that the integrated likelihood quantities are asymptotically global likelihood quantities with $\boldsymbol{\theta}^{**}$ and $\boldsymbol{\mu}^{**}$.

Theorem 4. For model (6), assume that Conditions (A) in the online Appendix hold, and $h \to 0$, $nh^3 \to \infty$ as $n \to \infty$.

(a) The local and global analysis of deviance (9) and (11), respectively, hold with $\breve{\mu}_x(x_i)$ and $\breve{\boldsymbol{\mu}}_x$ when α is estimated by maximum likelihood.

(b) Assume that α is estimated with a root-n rate. For $p = 1$ or 3, the expression in (13) holds with $\breve{\boldsymbol{\theta}}_x$ and $\boldsymbol{\theta}^{**}$ except the $\mathcal{O}(h^{p+1})$ term replaced by $\mathcal{O}(h^{p+1} + n^{-1/2})$. Similarly, (14) and (15) hold with $\breve{\mu}_x$ and $\boldsymbol{\mu}^{**}$ and the $\mathcal{O}(h^{p+1})$ terms replaced by $\mathcal{O}(h^{p+1} + n^{-1/2})$.

(c) When the same kernel function and bandwidth are used in (4) and (6), the nonparametric model (4) is nested in (6). Then, the difference in local residual deviance from fitting (6) to (4) can be expressed as

$$d_x(\mathbf{y}, \hat{\boldsymbol{\mu}}_x) - d_x(\mathbf{y}, \breve{\boldsymbol{\mu}}_x) = d_x(\breve{\boldsymbol{\mu}}_x, \hat{\boldsymbol{\mu}}_x). \tag{19}$$

The proofs of Theorem 4(a) and 4(b) are analogous to Theorems 1 and 2, respectively, and are thus omitted. We briefly outline the proof for Theorem 4(c). Based on (7) under (6), we have $\sum_i y_i(x_i - x)^j K_h(x_i - x) = \sum_i \breve{\mu}_x(x_i)(x_i - x)^j K_h(x_i - x)$, $j = 0, \ldots, p$. Then multiplying the jth equation by $\hat{\beta}_j$ and summing them up, $\sum_i\{y_i - \breve{\mu}_x(X_i)\}\hat{\theta}_i(x)K_h(X_i - x) = 0$ is obtained and (19) is proved.

In a special case of the Gaussian distribution with an identity link, Theorem 4(a) becomes the local and global analysis of variance for partially linear models, which was discussed in Huang and Davidson (2010, Sec. 3.1). Theorem 4(c) implies that the local residual deviance for fitting (6) is the local residual deviance for fitting (4) minus a term due to the parametric component. That is, the difference of local residual deviances between (6) and (4) is a KL-divergence measure $d_x(\breve{\boldsymbol{\mu}}_x, \hat{\boldsymbol{\mu}}_x)$, and the local KL-divergence is additive between nested models (6) and (4). A similar interpretation holds at a global scale after integrating the local contributions of (19):

$$\int d_x(\mathbf{y}, \breve{\boldsymbol{\mu}}_x)dx = \int d_x(\mathbf{y}, \hat{\boldsymbol{\mu}}_x)dx - \int d_x(\breve{\boldsymbol{\mu}}_x, \hat{\boldsymbol{\mu}}_x)dx.$$

Analogous to Theorem 3, testing whether $m(\cdot)$ is significant under model (6) becomes testing whether $m(\cdot)$ is significantly different from a constant and an integrated likelihood ratio test is proposed in the following theorem.

Theorem 5. For model (6), assume that Conditions (A) in the online Appendix hold and the data matrix Z for covariates $\mathbf{z}$ is orthogonal to $\mathbf{x} = (x_1, \ldots, x_n)^\top$ and the intercept column. For testing $H_0 : m(x) = a_0$ with a_0 a constant versus $H_a : m(x)$ is not a constant function, when estimating $m(\cdot)$ by pth order local polynomials with $p \geq 0$, the test statistic

$$2\left\{\int \ell_x(\mathbf{y}; \breve{\boldsymbol{\theta}}_x)dx - \ell(\mathbf{y}; \hat{\boldsymbol{\alpha}}_0)\right\} \tag{20}$$

is asymptotically distributed according to a χ^2-distribution with df $\text{tr}(H_p^*) - 1$, where $\hat{\boldsymbol{\alpha}}_0$ is the maximum likelihood estimate under parametric H_0.

The assumption that Z is orthogonal to $\mathbf{x}$ in Theorem 5 is required for mathematical convenience, in the sense that the corresponding off-diagonal elements of the local Fisher information is 0, for ease of deriving the asymptotic χ^2-distribution of the test statistic (20). It will be seen in the simulations that the performance of the proposed tests remain reasonable when this assumption is violated. The integrated likelihood ratio tests in Theorems 3 and 5 depend on the bandwidth h, like the other nonparametric tests. Analogous to Theorem 3, the test statistic (20) has an interpretation of integrating the differences in local deviances between a fitted generalized partially linear model and a parametric reduced model. We remark that the proposed tests are different from those in Hastie and Tibshirani (1990) and Li and Liang (2008). The proposed test statistics use integrated likelihood that combines all maximized local

likelihoods from fitting local polynomials. Some existing methods use only $G^{-1}(\hat{\beta}_0(x_i))$'s, and strictly speaking, the resulting expression is not based on maximizing likelihood as $\hat{\beta}_1, \ldots, \hat{\beta}_p$ are ignored; this fact was mentioned before Theorem 1 as well.

Some work in the literature—for example Härdle, Mammen, and Müller (1998)—has considered testing whether m in (6) is significantly different from a linear trend, $G(\mu) = \mathbf{z}^\top \alpha + a_0 + a_1 x$. The extension of Theorem 5 to testing a linear trend is nontrivial and will be pursued in future work, since the variance function of y is allowed to be a function of the mean of y in exponential family (1). In a special case of the Gaussian distribution with a constant variance, analysis-of-variance F-type tests for checking linear trends were derived by Huang and Davidson (2010).

5. SIMULATION RESULTS

We examine the empirical Type I errors and power for the proposed tests in Theorems 3 and 5. Local linear smoothing $p = 1$ with the Epanechnikov kernel is used throughout this section. We first describe the algorithm for calculating the test statistic (20) in Theorem 5 for testing $H_0 : m(x) = a_0$ under model (6), while that for (16) under model (4) is similar. The algorithm adapted from Carroll et al. (1997) is given as follows:

Step 0 (initialization). Fit a parametric generalized linear model to obtain initial values $\breve{\alpha}^{(0)}$.

Step 1. For a set of grid points on the data range of x_i's, given a value of h, maximize the local likelihood with $\breve{\alpha}^{(r)}$ to obtain $\breve{\beta}_0^{(r)}, \ldots, \breve{\beta}_p^{(r)}$ for each grid point. Then with $\breve{\theta}_i^{(r)}(x)$'s, calculate a locally weighted average as in (12) to obtain $\breve{\theta}_i^{**(r)}$, $i = 1, \ldots, n$.

Step 2. Maximize the global likelihood with $\boldsymbol{\theta}^{**(r)} = (\breve{\theta}_1^{**(r)}, \ldots, \breve{\theta}_n^{**(r)})^\top$ to update $\breve{\alpha}^{(r+1)}$.

Step 3. Continue Steps 1 and 2 until convergence. The test statistic (20) is calculated by integrating the final local likelihoods and taking its difference to the global likelihood under H_0.

The simulation study focuses on logistic regression in Examples 1–4 as we wish to evaluate the proposed methods to analyze the German Bundesbank data in Section 6, while Example 5 is on Poisson regression. The integrated likelihood in the test statistics (16) and (20) are approximated discretely by taking 201 equally-spaced points on [0, 1] in Examples 1 and 2, and 301 equally spaced points on $[-0.5, 1]$ in Examples 3–5. For x_i's that fall in conventional boundary regions, analogous approximations are used for calculating $\int K_h(x_i - x)dx$ for boundary correction in condition (A1). In addition to implementing the proposed tests with a fixed h, we also try selecting the bandwidth by AICc (Hurvich, Simonoff, and Tsai 1998) and by the idea of Horowitz and Spokoiny (2001) (HS). The AICc criterion is adapted with the integrated deviance:

$$\mathrm{AICc}(h) = \log(D^*/n) + 2\left(\mathrm{tr}\left(H_1^*\right) + 1\right) / \left(n - \mathrm{tr}\left(H_1^*\right) - 2\right),$$

where D^* denotes the integrated deviance $\int d_x(\mathbf{y}, \hat{\boldsymbol{\mu}}_x)dx$ in (14) under model (4) or $\int d_x(\mathbf{y}, \breve{\boldsymbol{\mu}}_x)dx$ under model (6). The HS idea is to select the bandwidth that maximizes the test statistic. Critical values for the proposed tests are taken from the χ^2-distribution with 5% significance level and 5000 simulated datasets are generated. The `gam` function in the `mgcv` R package (Wood 2013) provides a chi-square test of zero effect of a smooth term and we include it for comparison, with default 10 spline basis functions and the penalty estimated by REML.

Example 1. $\mathrm{logit}(p) = -1 + a\cos(2\pi x)$, $a = 0, 0.5, 0.75, 1$. We first check the χ^2-approximation under H_0 when $a = 0$ for both a fixed design, x equally-spaced on [0, 1], and a random design, $x \sim U(0, 1)$, with sample sizes n =50, 100, and 200. The values of bandwidth h =0.1, 0.12, 0.15, 0.17, 0.2, 0.25, and 0.3 are chosen so that they are roughly equally spaced on a logarithm scale and they correspond to smoothing with about 20%–60% data. For both AICc and HS, the bandwidth among the 7 values that satisfies the criterion is selected. The results with h =0.1, 0.15, 0.2, and 0.25, are chosen to present in Table 1, from under-smoothing slightly to over-smoothing slightly. The results with varying h by AICc and HS are also given in Table 1. It appears that when n =50, the χ^2-approximation for (16) under H_0 is not good as the empirical Type I errors are all above 0.05 for either a fixed or random design. For this reason, we do not consider the case of n =50 further. As suggested by two reviewers, a bootstrap alternative for calculating the sample critical values for $n = 50$ may be considered for future research.

When $a = 0$ and $n = 100$, the empirical Type I errors are mostly reasonable except for a small h =0.1, h_{AICc}, and h_{HS}, with rates ranging about 7–10%. In this case with a random design, h_{AICc} tends to select the largest bandwidth 0.3, 92.12% of 5000 simulations, since the true model under H_0 is a constant, and when AICc happens to select a small bandwidth such as 0.1, it often leads to rejecting H_0. For h_{HS}, it behaves differently since it attempts to optimize the power; when $a = 0$ and $n = 100$, the empirical proportions of h_{HS} on the 7 values of $h = 0.1, \ldots, 0.3$ are 44.58%, 5.8%, 4.96%, 4.08%, 5.04%, 4.84%, and 30.70%, respectively. Therefore, the inflated Type I errors of h_{HS} are somewhat expected. In Horowitz and Spokoiny (2001), the critical values were based on resampling from the finite-sample null distribution, while we use the asymptotic χ^2-distribution. When $n = 200$, the empirical type-I errors for the proposed tests are around 0.05 with a fixed bandwidth, and slightly above 0.05 for h_{AICc} and h_{HS}. The performance of h_{AICc} when $n = 200$ is closer to H_0 than that of $n = 100$. The `gam` function performs consistently around 0.05 under H_0 regardless of the sample sizes. When n =100 with a fixed design, the df $(\mathrm{tr}(H_1^*) - 1)$ are 10.29, 6.84, 5.11, and 4.07, respectively, for h =0.1, 0.15, 0.2, and 0.25, respectively, and the average estimated degrees of freedom (edf) for `gam` is 1.33 with a range [1.00, 6.85]. Quantile–quantile plots (qqplots) of 5000 test statistics (16) for n =100 with a fixed design against the χ^2-quantiles with the corresponding df are shown in Figure 1 for h =0.1, 0.15, 0.2, and 0.25, indicating satisfactory approximations of the χ^2-distribution. The qqplots of n =200 with a fixed bandwidth (not shown) are similar to those of n =100.

When a =0.5, 0.75, and 1, with a random design, we examine the performance under alternatives. The percent of rejection is given in Table 2. For the proposed tests, we observe that the rejection rate increases as the value of the bandwidth

Table 1. Percent of rejection under H_0 in Example 1 with $a = 0$

	$h = 0.1$	$h = 0.15$	$h = 0.2$	$h = 0.25$	h_{AICc}	h_{HS}	gam
$n = 50$ fixed design	12.80	8.74	6.70	6.02	10.02	15.86	3.22
$n = 50$ random design	14.10	9.34	7.20	6.24	10.38	17.26	3.06
$n = 100$ fixed design	7.38	5.30	4.80	4.52	7.18	9.96	3.99
$n = 100$ random design	7.94	5.84	4.70	4.36	7.28	10.48	4.08
$n = 200$ fixed design	4.96	4.20	4.18	3.96	5.56	7.36	4.62
$n = 200$ random design	5.22	4.38	4.42	4.20	6.00	7.76	5.04

The empirical Type I errors of the test statistic (16) are close to 0.05 for $n = 100$ and 200 with a fixed $h \geq 0.15$. The performance of h_{AICc} when $n = 200$ is closer to H_0 than that of $n = 100$, and h_{HS} has larger Type I errors as it attempts to optimize the power. The gam function performs consistently around 0.05 under H_0 regardless of the sample sizes.

increases when a =0.75 and 1, and h_{AICc} and h_{HS} are more powerful than those with a fixed h. Under alternatives, the proposed tests are more powerful than gam except the case with $n = 200$ and h =0.1. When n =100, the average dfs for (16) of $a = 0.5$, 0.75, and 1 are similar to those of a =0, since H_1^* in (17) does not involve the response y. When $n = 100$, the average edf of gam increases as a increases, 1.80, 2.30, and 2.64 for a =0.5, 0.75, and 1, respectively. The behavior for df and edf of n =200 is similar to that of $n = 100$.

Example 2. $\text{logit}(p) = -2 + f_k(x)$, $k = 0, 1, 2$, where $x \sim U(0, 1)$ and functions $f_0(x) = 8x(1 - x)$, $f_1(x) = \exp(2x)$, and $f_2(x) = 2 \times 10^5 x^{11}(1 - x)^6 + 10^4 x^3(1 - x)^{10}$, are taken from Wood (2013). Wood (2013) considered an additive model with $\text{logit}(p) = -5 + f_0(x_0) + f_1(x_1) + f_2(x_2)$, while we use those functions in the univariate case separately. The results are shown in Table 3, with the proportions of rejection nearly 100% for both tests in cases of f_1 and f_2. For f_0, a quadratic trend, our test is more powerful than gam except the case with $n = 200$ and h =0.1.

Example 3. $\text{logit}(p) = b_1 z_1 + b_2 z_2 + a \exp(-16x^2)$, $a = 0$, 1, 2, 3, where z_1 is first generated as binary taking values -1 and 1 with equal probabilities, z_2 and x are first generated from a bivariate normal distribution with mean 0, variances 0.5 and 1 respectively, and correlation 0.3. Then, x is transformed to have a uniform distribution on $(-0.5, 1)$. To satisfy the conditions in Theorem 5, z_1 and z_2 are then made orthogonal to $\mathbf{x}$ and the intercept vector. After the orthogonized z_1 and z_2 are obtained, $b_1 = 0.1$, $b_2 = -0.1$. To understand how restrictive

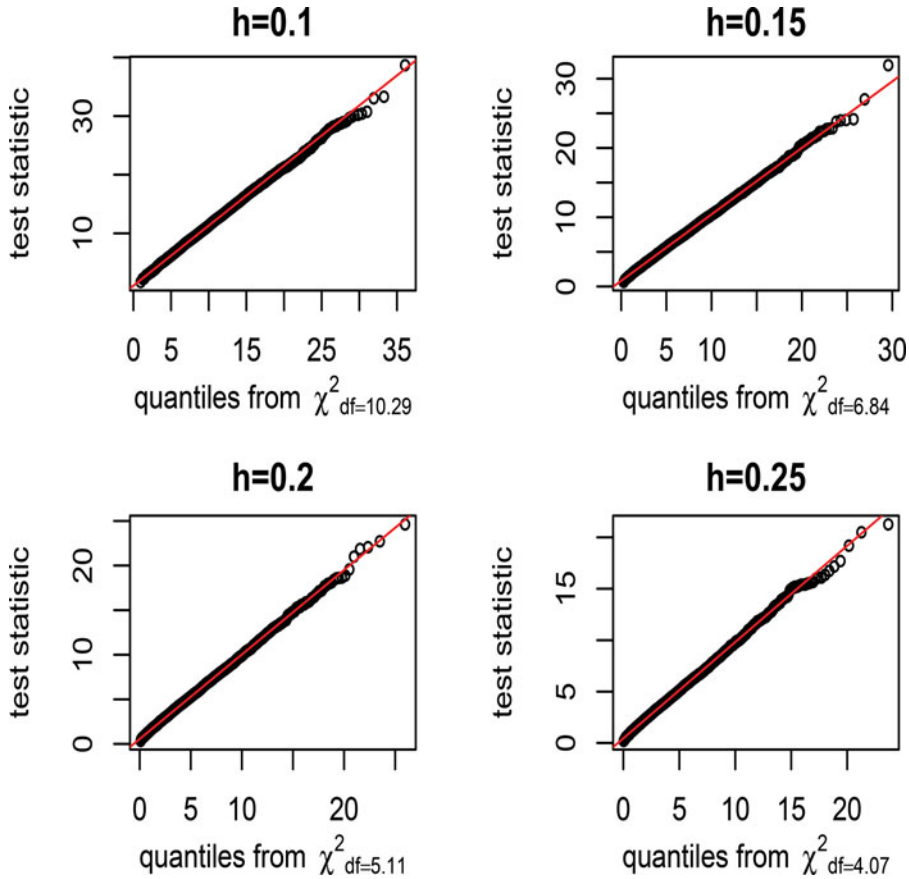


Figure 1. Quantile–quantile plots of 5000 integrated likelihood ratio test statistics (16) under H_0 in Example 1 for $n = 100$ with h =0.1, 0.15, 0.2, and 0.25, against quantiles from χ^2-distribution with df 10.29, 6.84, 5.11, and 4.07, respectively.

Table 2. Percent of rejection under H_1 in Example 1

	$h = 0.1$	$h = 0.15$	$h = 0.2$	$h = 0.25$	h_{AICc}	h_{HS}	gam
$n = 100, a = 0.5$	20.12	18.78	19.32	20.38	24.26	28.58	13.96
$n = 100, a = 0.75$	37.22	38.38	40.96	43.42	47.14	51.26	31.96
$n = 100, a = 1$	60.56	63.72	67.96	70.60	73.52	75.78	58.86
$n = 200, a = 0.5$	29.18	33.04	36.30	38.96	42.28	45.08	30.82
$n = 200, a = 0.75$	61.18	67.40	71.68	75.00	77.66	78.84	66.58
$n = 200, a = 1$	88.58	92.38	94.28	95.56	96.34	96.64	92.36

The rejection rate for the test statistic (16) increases as the value of the bandwidth increases when a =0.75 and 1, and h_{AICc} and h_{HS} are more powerful than those with a fixed h. The test statistic (16) is more powerful than gam except the case with $n = 200$ and $h = 0.1$.

the orthogonality assumption in Theorem 5 is, we also examine the performance of (20) with the original non-orthogonalized values of z_1 and z_2 and same values of b_1 and b_2.

The values of bandwidth are 0.15, 0.2, 0.25, 0.3, and 0.4, so that they are roughly equally-spaced on a logarithm scale. The percent of rejection is given in Table 4 for h =0.2, 0.25, 0.3, 0.4, h_{AICc}, and h_{HS}, with the nonorthogonalized version in brackets. The case of $h = 0.15$ is not presented due to its inflated Type I errors: when $a = 0$, $n = 100$, and $h = 0.15$, the percent of rejection is 8.72 and 8.28 for the orthogonalized and non orthogonalized version respectively. From Table 4, we observe that when $a = 0$, the empirical Type I errors are reasonable except h_{AICc} and h_{HS}. Together with the observations in Example 1 under H_0, we may imply that optimizing the bandwidth by some criterion may lead to inflated Type I errors for our test in the case of logistic regression. When $a = 1$ and 2, our test is more powerful than the gam test, while for $a = 3$, the performance of the two tests are close. Under alternatives, h_{HS} is the most powerful, while h_{AICc} also performs well, supporting AICc as a bandwidth-selection criterion. The rejection rates for (20) are quite close whether Z and $\mathbf{x}$ are orthogonal or not, suggesting that this assumption may be relaxed in practice. When $n = 200$ and $a = 0$, the average df $(\text{tr}(H_1^*) - 1)$ corresponding to $h = 0.2, \ldots, 0.4$ are 7.69, 6.14, 5.10, and 4.81, respectively, and again they stay about the same between different values of a. When $n = 200$, the average edf for gam is 1.34, 2.22, 3.89, and 4.73 for a =0, 1, 2, 3, respectively. The df and edf of $n = 100$ are similar to those of $n = 200$.

Example 4. $\text{logit}(p) = b_1 z_1 + b_2 z_2 + a\cos(2\pi x)$, $a = 0.5$, 1, 1.5, where the data generation scheme of z_1, z_2, and x is identical to Example 3, and b_1 and b_2 are the same as Example 3. This example adopts a nonlinear function of x similar to that of Example 1 with a different range of x. The same values of h as Example 3 are used and hence the dfs are analogous to Example 3, omitted for brevity. Table 5 shows that (20) is more powerful than gam when $a = 0.5$ and 1.0, and our test with h_{HS} and h_{AICc} continues to perform well in this example. Again, we observe that for the proposed tests, the rejection rates are quite close whether Z and $\mathbf{x}$ are orthogonal or not.

Example 5.

$$\log(\mu) = b_1 z_1 + b_2 z_2 + a\exp(-16x^2), \quad a = 0, 1, 2, \tag{21}$$

and

$$\log(\mu) = b_1 z_1 + b_2 z_2 + a\cos(2\pi x), \quad a = 0.5, 1.5. \tag{22}$$

This example is for Poisson regression with the canonical log link, while the functional form for $\theta = \log(\mu)$ and data generation scheme follow those of Examples 3 and 4. From Table 6, we observe that when $a = 0$ in (21) with $n = 100$ and $h = 0.15$, the Type I error is reasonable, in contrast to the logistic regression case. The performance of $h = 0.4$ is close to that of $h = 0.3$ and therefore not presented in Table 6. We observe that our test using a fixed h is more powerful with a larger bandwidth when $a = 1$ in (21) and $a = 0.5$ in (22), and the empirical power is comparable between (20) and gam tests.

6. APPLICATION TO GERMAN BUNDESBANK DATA

Banking throughout the world is based on credit, or on trust in the debtor's ability to fulfill his/her debt obligation. However, facing increasing pressure from markets and regulators, banks have based their risk analysis, increasingly, on statistical techniques to judge or predict corporate bankruptcy. This is known as rating or scoring. Its main purpose is to estimate the financial

Table 3. Percent of rejection for Example 2

	$h = 0.1$	$h = 0.15$	$h = 0.2$	$h = 0.25$	h_{AICc}	h_{HS}	gam
$n = 100, f_0$	47.94	51.26	54.72	57.46	60.20	64.38	38.70
$n = 100, f_1$	99.86	100	100	100	100	100	98.90
$n = 100, f_2$	100	100	100	100	100	100	99.74
$n = 200, f_0$	76.10	82.00	85.38	87.84	89.34	90.18	80.82
$n = 200, f_1$	100	100	100	100	100	100	100
$n = 200, f_2$	100	100	100	100	100	100	100

The rejection rates are nearly 100% for (16) and gam in cases of f_1 and f_2. For f_0, a quadratic trend, (16) is more powerful than gam except the case with $n = 200$ and $h = 0.1$.

Table 4. Percent of rejection for Example 3

	$h = 0.2$	$h = 0.25$	$h = 0.3$	$h = 0.4$	h_{AICc}	h_{HS}	gam
$n = 100, a = 0$	6.40	5.78	5.34	4.96	8.10	11.04	4.26
	[6.22]	[5.64]	[4.98]	[4.74]	[7.70]	[10.46]	[4.46]
$n = 100, a = 1$	20.50	20.54	20.64	20.76	23.24	28.40	13.36
	[20.98]	[20.32]	[20.36]	[20.12]	[22.46]	[28.24]	[13.04]
$n = 100, a = 2$	69.14	70.76	72.00	72.30	73.50	77.14	57.72
	[68.36]	[70.26]	[71.54]	[71.74]	[72.84]	[76.62]	[57.08]
$n = 100, a = 3$	96.00	96.92	97.38	97.42	97.54	98.22	92.34
	[95.50]	[96.56]	[96.88]	[97.00]	[97.16]	[97.86]	[91.46]
$n = 200, a = 0$	5.56	5.24	5.20	4.78	6.80	8.88	5.14
	[5.62]	[5.08]	[4.90]	[4.62]	[6.60]	[8.62]	[5.00]
$n = 200, a = 1$	34.86	36.94	38.04	38.46	39.94	44.24	29.26
	[34.22]	[36.06]	[37.58]	[38.48]	[39.94]	[43.76]	[29.32]
$n = 200, a = 2$	95.92	96.58	96.92	97.16	97.36	97.70	93.50
	[95.68]	[96.56]	[97.06]	[97.20]	[97.32]	[97.54]	[93.26]
$n = 200, a = 3$	100	100	100	100	100	100	99.96
	[100]	[100]	[100]	[100]	[100]	[100]	[99.98]

When $a = 0$ with a fixed $h >= 0.2$, the empirical Type I errors of (20) are close to 0.05. When $a = 1$ and 2, (20) is more powerful than the gam test, while for $a = 3$, the performance of the two tests are close. Under alternatives, h_{HS} is the most powerful, while h_{AICc} also performs well. The rejection rates for (20) are quite close whether Z and $\mathbf{x}$ are orthogonal or not (the non orthogonalized version in brackets).

status of a company and, if possible, to estimate the probability of a company default on its debt obligations within a certain period. Logistic regression is probably the most commonly used technique to model the probability of default and logistic partially linear models may also be advantageous because of its flexibility, in allowing for the possibly nonlinear effects of one continuous covariate.

We apply the methodology to the German Bundesbank Data in year 2002. The data provided by CRC 649, Humboldt-Universität zu Berlin, contained 6123 companies of which 186 were insolvent. Each firm is described by 28 financial ratio variables, $x1, \ldots, x28$, and those of insolvent firms were collected two years prior to insolvency. To ensure the value of some variables as the denominator should not be zero when calculating the ratios, 2079 firms were retained with 92 insolvent. Though removing almost two-thirds of the sample may seem excessive, we did not intend to analyze the majority of firms in the database. The focus was to investigate (i) differences between the financial ratios of the solvent and insolvent firms, and (ii) how the nonlinear effects improve parametric logistic fitting.

Based on support vector machines and for a much larger data sample spanning from 1996 through to 2002, Chen, Härdle, and Moro (2011) selected $x24$ (accounts payable/sales) measuring account payable turnover, as the best predictor, and subsequently selected $x3$ (operating income/total assets) measuring profitability, $x15$ ((cash and cash equivalents)/total assets) measuring liquidity, $x12$ (total liabilities/total assets) measuring leverage, $x26$ (increase (decrease) inventories/inventories) measuring percentage of incremental inventories, $x22$ (inventories/sales) measuring inventory turnover, $x5$ ((earnings before interest and tax)/total assets) and $x2$ (net income/sales) measuring net profit margin. For year 2002 data, we found that $x3$ and $x5$ have a large sample correlation coefficient 0.95 and thus $x5$ is removed from our analysis and we further include $x25$ (log(total assets)) measuring firm size, as it is shown to be an important variable on predicting the probability of bankruptcy

Table 5. Percent of rejection for Example 4

	$h = 0.2$	$h = 0.25$	$h = 0.3$	$h = 0.4$	h_{AICc}	h_{HS}	gam
$n = 100, a = 0.5$	20.16	19.66	19.62	18.56	22.38	26.78	8.60
	[20.02]	[19.54]	[19.22]	[18.60]	[22.40]	[27.24]	[8.98]
$n = 100, a = 1$	62.76	64.36	65.46	64.68	67.08	71.54	38.86
	[61.60]	[63.48]	[64.08]	[63.38]	[65.88]	[70.58]	[38.34]
$n = 100, a = 1.5$	94.58	95.18	95.60	95.18	95.72	96.48	85.24
	[94.00]	[94.98]	[95.22]	[94.72]	[95.18]	[96.14]	[84.90]
$n = 200, a = 0.5$	30.60	32.36	33.36	32.58	34.90	39.00	18.58
	[30.80]	[32.24]	[32.88]	[32.54]	[34.94]	[38.88]	[18.66]
$n = 200, a = 1$	91.60	93.10	93.92	93.68	94.18	95.04	83.60
	[90.94]	[92.42]	[93.50]	[93.04]	[93.48]	[94.48]	[82.84]
$n = 200, a = 1.5$	100	100	100	100	100	100	99.86
	[99.98]	[100]	[100]	[100]	[100]	[100]	[99.88]

The test statistic (20) is more powerful than gam under alternatives. The rejection rates are quite close whether Z and $\mathbf{x}$ are orthogonal or not (the non-orthogonalized version in brackets).

Table 6. Percent of rejection for Example 5

	$h = 0.15$	$h = 0.2$	$h = 0.25$	$h = 0.3$	h_{AICc}	h_{HS}	gam
$n = 100$, $a = 0$ in (21)	5.36	4.80	4.18	3.94	7.06	7.62	4.88
	[5.48]	[4.64]	[4.40]	[4.26]	[7.14]	[7.68]	[4.86]
$n = 100$, $a = 1$ in (21)	41.26	44.48	47.24	48.42	51.98	54.48	37.58
	[41.36]	[44.52]	[47.12]	[48.90]	[52.60]	[54.98]	[37.24]
$n = 100$, $a = 2$ in (21)	99.98	99.98	99.98	100	99.98	100	99.94
	[99.98]	[99.98]	[99.98]	[100]	[99.98]	[100]	[99.96]
$n = 200$, $a = 0$ in (21)	4.98	4.56	4.48	4.32	6.92	7.38	5.14
	[5.02]	[4.66]	[4.74]	[4.60]	[7.06]	[7.62]	[5.32]
$n = 200$, $a = 1$ in (21)	74.10	78.64	81.48	82.88	85.02	85.86	75.04
	[73.62]	[78.46]	[80.84]	[82.04]	[84.12]	[84.92]	[74.14]
$n = 200$, $a = 2$ in (21)	100	100	100	100	100	100	100
	[100]	[100]	[100]	[100]	[100]	[100]	[100]
$n = 100$, $a = 0.5$ in (22)	26.20	28.36	30.02	30.90	34.72	36.58	17.32
	[25.72]	[27.56]	[29.40]	[30.18]	[34.06]	[35.66]	[16.96]
$n = 100$, $a = 1.5$ in (22)	99.92	100	100	100	100	100	99.98
	[99.94]	[99.94]	[99.98]	[100]	[100]	[100]	[99.94]
$n = 200$, $a = 0.5$ in (22)	51.86	56.74	60.22	62.24	64.90	65.88	43.94
	[51.18]	[56.22]	[60.18]	[61.92]	[64.44]	[65.38]	[43.90]
$n = 200$, $a = 1.5$ in (22)	100	100	100	100	100	100	100
	[100]	[100]	[100]	[100]	[100]	[100]	[100]

When $a = 0$, the Type I errors of (20) are reasonable for this Poisson regression example. The test statistic (20) using a fixed h is more powerful with a larger bandwidth when $a = 1$ in (21) and $a = 0.5$ in (22), and the empirical power is comparable between (20) and gam tests.

in the literature (see, e.g., Lopez 2004). In summary, there are eight predictors, $x2$, $x3$, $x12$, $x15$, $x22$, $x24$, $x25$, and $x26$, and a binary response. See Chen, Härdle, and Moro (2011) for detail descriptions about the data.

Since $x24$ was selected as the most important predictor by Chen, Härdle, and Moro (2011), we model its effects nonparametrically, while retaining linear trends for the remaining predictors in a logit model. The variable $x24$ measuring account payable turnover is a short-term liquidity measure for quantifying the rate at which a firm pays off its suppliers. Generally speaking, "the firms with higher account payable turnover will have less ability to convert their accounts into sales, have lower revenues, and go bankrupt more readily" (Chen, Härdle, and Moro 2011). However, this measure is specific to different industries; every industry has a slightly different standard. Further examination of $x24$ indicates that most values lie in $[0, 0.5]$ with only 15 observations in $(0.5, 20.52)$. If those 15 observations are excluded, then the sample size becomes 2064, in which 91 are insolvent. An alternative approach, suggested by a reviewer, is taking logarithm of $(x24 + 0.001)$ (0.001 is added since $x24$ includes 0's) and retaining the sample size $n = 2079$.

Local linear smoothing with the Epanechnikov kernel is used. The values of bandwidth for $x24$, 0.125, 0.1, and 0.08, are equally spaced on a logarithmic scale, corresponding to df 4.94, 5.92, and 7.17, respectively. The bandwidth that minimizes AICc is 0.125 and $h_{\text{HS}} = 0.1$. The curves for $m(x24)$ with pointwise confidence intervals based on empirical Fisher information matrices are shown in Figure 2 with $h = 0.125$ and 0.1, and the proposed test for testing $H_0 : m(x24)$ is a constant, gives highly significant p-values $< 10^{-14}$ for all three values of the bandwidth, indicating significance of $m(x24)$ in predicting probability of default . A linear logistic model gives a positive slope 10.37 for $x24$ with a highly significant p-value $< 10^{-15}$. Since Chen, Härdle, and Moro (2011) interpreted the linear trend as higher default probability with high turnover, we attempt to interpret the seemingly nonlinear curves in Figure 2 as follows. Taking the curve with $h = 0.1$ in Figure 2, when $x24$ increases from 0.1 to 0.3, the estimate increases about 1.895, which means the odds ratio for a firm with $x24 = 0.3$ to become insolvent is $\exp(1.895) = 6.653$ times relative to that for a firm with $x24 = 0.1$. On the other hand, between $x24 = 0.3$ and 0.4, the estimate decreases by an amount of -0.607, implying that the odds ratio for a firm with $x24 = 0.4$ to become insolvent is $\exp(-0.607) = 0.545$ times relative to that for a firm with $x24 = 0.3$. Thus our analysis gains new insight suggesting that a German firm is likely to go bankrupt when it has a higher turnover for roughly 97.5% of firms (0.3 is approximately 97.5-percentile of $x24$), but

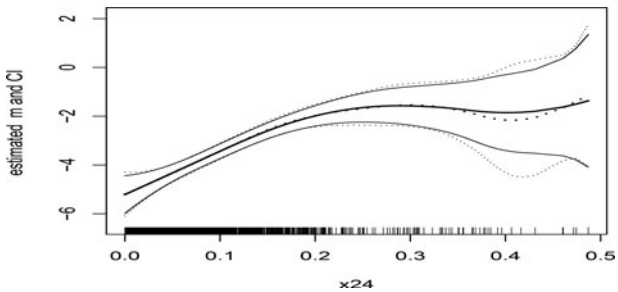


Figure 2. Plot of the nonlinear trends of $x24$ in predicting the probability of bankruptcy using bandwidth $h = 0.125$ (solid line) and $h = 0.1$ (dashed line) with 95% pointwise confidence intervals for the 2002 German Bundesbank Data.

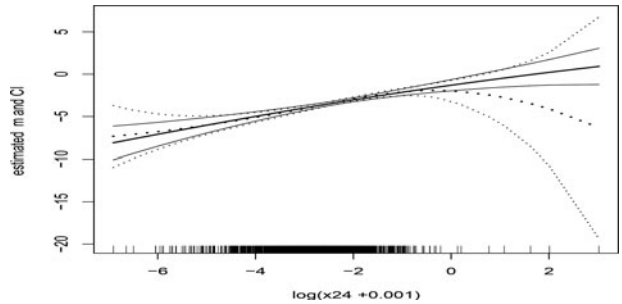


Figure 3. Plot of the trends of log$x24$ in predicting the probability of bankruptcy using bandwidth $h = 6$ (solid line) and $h = 3.375$ (dashed line) with 95% pointwise confidence intervals for the 2002 German Bundesbank Data.

for those firms with $0.3 < x24 < 0.4$ (approximately 97.5- to 99-percentile), the default probability decreases as $x24$ increases.

For smoothing on $\log(x24 + 0.001) = \log x24$ with $n = 2079$, the values of h are 6, 4.5, and 3.375, corresponding to df 2.32, 2.83, and 3.69, respectively. The bandwidth that minimizes AICc is 3.375 and $h_{\mathrm{HS}} = 6$. The curves for $\log x24$ with $h = 3.375$ and 6 shown in Figure 3 have a linear tendency for $\log x24 < -0.2$ ($x24 < 0.8$), and the confidence intervals corresponding to $h = 3.375$ imply some uncertainty near the right-hand end points. The tests for $H_0 : m(\log x24)$ is a constant, give highly significant p-values of $< 1.5 \times 10^{-12}$. If using a linear trend for $\log x24$, the slope is 0.927 with a significant p-value of 2.0×10^{-12}. The analysis using $\log x24$ implies that a German firm is likely to go bankrupt when it has a high turnover in the log scale, but the linear trend is uncertain for those with $x24 > 0.8$. Hence, an interpretation in Chen, Härdle, and Moro (2011) that a German firm is likely to go bankrupt when it has high turnover may not be entirely correct; the effects of $x24$ on the probability of bankruptcy may be nonlinear for those with large turnovers, as shown in Figures 2 and 3.

7. DISCUSSION

We develop local and global analysis of deviance expressions and associated integrated likelihood ratio tests for generalized partially linear models with canonical links based on fitting local pth order polynomials. Though the idea of nonparametric analysis of deviance is not new (Hastie and Tibshirani 1990), the work in this article provides theoretical justifications that connect to the classic framework. Theorems 2 and 4(b) are restricted for $p = 1$ and 3 only, while Theorems 1, 3, 4(a)(c), and 5 are for a nonnegative integer p. As a by-product, new estimators for the canonical parameter and response mean are proposed and Theorems 2 and 4(b) show that the integrated likelihood quantities are asymptotically global likelihood quantities with the new estimators. The new estimator $\hat{\theta}_i^*$ or $\check{\theta}_i^*$ for the canonical parameter is formed by combining locally fitted $\hat{\theta}_i(x)$ or $\check{\theta}_i(x)$ through weighted integration and thus use all locally fitted parameters, which is different from the conventional approach of focusing on $\hat{\beta}_0$. The integrated likelihood approach of combining local likelihood appears to be new in the smoothing literature, though it was discussed by Severini (2007) and Lehmann (2006) in different settings. The numerical results of $n = 100$ and 200 show that the test statistics under the null hypothesis follow the asymptotic χ^2-distribution reasonably well and the performance under alternative hypotheses is sometimes more powerful than Wood (2013) in the R package `mgcv`. It has been suggested by a reviewer to investigate the asymptotic power of the proposed tests. Since there is no simple explicit expression for Fisher information for generalized linear models (1), we conjecture that the study of power may be focused on special cases of logistic and Poisson models, which will be explored for future research. For a smaller sample size such as $n = 50$, two reviewers have suggested to develop a bootstrap procedure for calculating the sample critical values for further investigation.

The local analysis of deviance in (9) and in Theorem 4(a) are derived assuming a fixed value of bandwidth. It is straightforward to obtain local analysis of deviance expressions with varying values of bandwidth at different x, but how to combine them to form global analysis of deviance will be an interesting problem. Like all smoothing-based tests, the p-values of the integrated likelihood ratio tests depend on the values of the smoothing parameter. We recommend plotting the "significant trace" (Bowman and Azzalini 1997) to assess the evidence across a wide range of values of h and looking for some overall trends. For fitting generalized partially linear models, a practical problem is how to choose the predictor to be modeled nonparametrically. One approach may be based on selecting the most significant predictor based on the smallest p-value of integrated likelihood ratio tests when using approximately the same degrees of freedom for smoothing. This idea and the related model selection problems with a diverging number of linear covariates (Wang et al. 2014) may be explored for future research. A topic for further investigation is the problem of bandwidth selection for models (4) and (6) based on cross-validating the deviance or minimizing the Akaike information criterion. Further extension on developing analysis of deviance for generalized partially linear models with noncanonical links, for multiplicative bias reduction methods (Kosmidis and Firth 2009), for hazard estimation (Nielsen and Tanggaard 2001) as the proportional hazards models and Poisson regression are connected, and for generalized additive models with multiple nonparametric functions remain to be investigated.

ACKNOWLEDGMENTS

We thank the editor, associate editor , and two anonymous referees for their constructive comments and suggestions. The work was conceived during the visit of the second author to the Humboldt-Universität zu Berlin supported by CRC 649 "Economic Risk." The support is greatly appreciated. The first author was partially supported by SKBI School of Business, Singapore Management University, the second author (corresponding author) was partially supported by the Ministry of Science and Technology NSC 101-2118-M-007-002-MY2 in Taiwan, and both authors were partially supported by CRC 649 "Economic Risk." The authors wish to thank Mr. Leslie Udvarhelyi who assisted in the proof-reading of the article.

ORCID

Li-Shan Huang http://orcid.org/0000-0001-8297-2804

[Received April 2016. Revised February 2017.]

REFERENCES

Bowman, A. W., and Azzalini, A. (1997), *Applied Smoothing Techniques for Data Analysis*, London: Oxford. [11]

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997), "Generalized Partially Linear Single-Index Models," *Journal of the American Statistical Association*, 92, 477–489. [1,3,6]

Chen, S., Härdle, W., and Moro, R. (2011), "Modeling Default Risk With Support Vector Machines," *Quantitative Finance*, 11, 135–154. [9,10,11]

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, London: Chapman and Hall. [2]

Fan, J., Heckman, N. E., and Wand, M. P. (1995), "Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions," *Journal of the American Statistical Association*, 90, 141–150. [2,3]

Fan, J., Zhang, C., and Zhang, J. (2001), "Generalized Likelihood Ratio Statistics and Wilks Phenomenon," *Annals of Statistics*, 29, 153–193. [1]

Green, P. J., and Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, London: Chapman and Hall. [1]

Härdle, W., Mammen, E., and Müller, M. (1998), "Testing Parametric Versus Semiparametric Modeling in Generalized Linear Models," *Journal of the American Statistical Association*, 93, 1461–1474. [1,6]

Härdle, W. K., Müller, M., Sperlich, S., and Werwatz, A. (2004), *Nonparametric and Semiparametric Models*, Berlin: Springer. [1]

Hastie, T. J., and Tibshirani, R. J. (1990), *Generalized Additive Models*, London: Chapman and Hall. [1,5,11]

——— (1987), "Local Likelihood Estimation," *Journal of the American Statistical Association*, 82, 559–567. [2]

Horowitz, J. L., and Spokoiny, V. G. (2001), "An Adaptive, Rate–Optimal Test of a Parametric Mean–Regression Model Against a Nonparametric Alternative," *Econometrica*, 69, 599–631. [6]

Huang, L.-S., and Chan, K.-S. (2014), "Local Polynomial and Penalized Trigonometric Series Regression," *Statistica Sinica*, 24, 1215–1238. [4]

Huang, L.-S., and Chen, J. (2008), "Analysis of Variance, Coefficient of Determination, and *F*-Test for Local Polynomial Regression," *Annals of Statistics*, 36, 2085–2109. [2,3,4]

Huang, L.-S., and Davidson, P. W. (2010), "Analysis of Variance and *F*-Tests for Partial Linear Models With Applications to Environmental Health Data," *Journal of the American Statistical Association*, 105, 991–1004. [2,5,6]

Hurvich, C. M., Simonoff, J. S., and Tsai, C.-L. (1998), "Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion," *Journal of the Royal Statistical Society*, Series B, 60, 271–293. [4,6]

Kosmidis, I., and Firth, D. (2009), "Bias Reduction in Exponential Family Nonlinear Models," *Biometrika*, 96, 793–804. [4,11]

Lehmann, E. L. (2006), "On Likelihood Ratio Tests," in *Optimality: The Second Erich L. Lehmann Symposium*, Institute of Mathematical Statistics Lecture Notes: Monograph Series vol. 49, ed. J. Rojo, Beachwood, OH: Institute of Mathematical Statistics, pp. 1–8. [2,4,11]

Li, R., and Liang, H. (2008), "Variable Selection in Semiparametric Regression Modeling," *Annals of Statistics*, 36, 261–286. [1,5]

Loader, C. (1999), *Local Regression and Likelihood*, New York: Springer. [2]

Lopez, J. A. (2004), "The Empirical Relationship Between Average Asset Correlation, Firm Probability of Default, and Asset Size," *Journal of Financial Intermediation*, 13, 265–283. [10]

Nielsen, J. P., and Tanggaard, C. (2001), "Boundary and Bias Correction in Kernel Hazard Estimation," *Scandinavian Journal of Statistics*, 28, 675–698. [11]

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman and Hall. [1,2,3,4]

Severini, T. A. (2007), "Integrated Likelihood Functions for Non-Bayesian Inference," *Biometrika*, 94, 529–542. [2,4,11]

Simon, G. (1973), "Additivity of Information in Exponential Family Probability Laws," *Journal of the American Statistical Association*, 68, 478–482. [2,3]

Wang, L., Xue, L., Qu, A., and Liang, H. (2014), "Estimation and Model Selection in Generalized Additive Partial Linear Models for Correlated Data With Diverging Number of Covariates," *Annals of Statistics*, 42, 592–624. [11]

Wood, S. N. (2006), *Generalized Additive Models: An Introduction With R*, Boca Raton, FL: Chapman and Hall. [1]

——— (2013), "On *p*-Values for Smooth Components of an Extended Generalized Additive Model," *Biometrika*, 100, 221–228. [1,2,6,7,11]

Journal of Forecasting, *J. Forecast.* **36**, 241–256 (2017)
Published online 20 July 2016 in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/for.2431

Adaptive Interest Rate Modelling

MENGMENG GUO[1]* AND WOLFGANG KARL HÄRDLE[2]

[1] *Research Institute of Economics and Management, Southwestern University of Finance and Economics, Chengdu, China*
[2] *Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Germany*

ABSTRACT

A good description of the dynamic process of interest rates is crucial to price derivatives and to hedge corresponding risk. An unstable macroeconomic context motivates the stochastic interest rate models with time-varying parameters. In this paper, the local parameter approach is introduced to adaptively estimate interest rate models. This method can be generally used in time-varying coefficient parametric models. It is used not only to detect jumps and structural breaks but also to choose the largest time homogeneous interval for each time point, such that in this interval the coefficients are statistically constant. We apply this adaptive approach in both simulations and real data analysis. Using the 3-month Treasury bill rate as a proxy of the short rate, we find that our method can detect the structural breaks as well as the stable intervals for homogeneously modelling of the interest rate process. The time homogeneous interval cannot persist in an unstable macroeconomy. Furthermore, our approach performs well in long horizon forecasting.

KEY WORDS CIR model; local parametric approach; time homogeneous interval; adaptive statistical techniques

INTRODUCTION

Interest rates are one of the most important factors in financial markets. For hedging purposes, they play an important role in pricing stocks and the corresponding derivatives. Hence it is crucial to be able to describe the dynamics of interest rates. Moreover, interest rates are a signal of the macroeconomy. If the macroeconomy is unstable, e.g. in the wake of a financial shock, we observe that interest rate volatility will be correspondingly larger, and vice versa. For instance, in 2002 bubbles existed in the stock market, and in 2003 the war in Iraq influenced the world macroeconomy. In 2007, the world economy fluctuated greatly due to the financial crisis. Along with these shocks, one finds the interest rate in these periods varying more significantly. In general, changes in business cycle conditions or macroeconomic shocks will affect the dynamics of interest rates in terms of mean and volatility. Moreover, in the empirical study described in this paper, we also find stronger fluctuations in the interest rate during the above mentioned periods. Alternately, changes in business cycle conditions may affect the dynamics of interest rates even from one period to another, which will also be tested in the paper.

On the other hand, these shocks or news items are dominated by announcements from central banks or government agencies, which release macroeconomic data at monthly or quarterly intervals. They may contain a large, unanticipated component. Moreover, interest rates respond immediately to these unanticipated announcements, which induces periodic fluctuations in the interest rate. Ultimately, the dynamics of interest rates do not follow a stable process. The corresponding findings are well documented in Jones *et al.* (1998) and Johannes (2004). All these factors lead us to believe that time-varying parameters would be more reasonable for describing interest rate dynamics.

Three main streams of literature exist that capture the instability of the dynamics of the short rate. In one branch of the literature, the described instability is modelled via structural breaks or jumps and captured by the general jump diffusion models. For instance, Das (2002) incorporated jumps into the Vasicek (1977) model and found strong evidence of jumps in the daily federal funds rate. Johannes (2004) used a nonparametric diffusion model to study secondary 3-month Treasury bills and concluded that jumps are generally generated by the arrival of macroeconomic news. A general conclusion would be that the dynamics of the short rate vary significantly due to shocks and jumps, which is also well described in Lettau and Ludvigson (2001), Goyal and Welch (2003) and Paye and Timmermann (2006). Another strand of literature uses regime switching models to capture the business cycle character of interest rates; see Ang and Bekaert (2002) and Bansal and Zhou (2002). They found that the interest rate is closely related

*Correspondence to: Mengmeng Guo, Research Institute of Economics and Management, Southwestern University of Finance and Economics, LiuTai Blvd. 555, Wenjiang District, 611130 Chengdu, Sichuan, China. E-mail: gmm@swufe.edu.cn

to the business cycle. The short rate has changed significantly, and its volatility performs differently in expansion regimes and recession regimes. More generally, many studies argue that the process parameters (drift or volatility) are assumed to be functions of time. This is well documented in numerous studies, such as Hull and White (1990), Black and Karasinski (1991), Aït-Sahalia (1996), Stanton (1997), Fan *et al.* (2003) and Arapis and Gao (2006). As an example, using semi- and nonparametric approaches, Aït-Sahalia (1996) found strong nonlinearity in the drift function of the interest rate model. Arapis and Gao (2006) applied nonparametric techniques to provide evidence that the specification of the drift has a considerable impact on the pricing of derivatives through its effect on the diffusion function. As a conclusion from these findings, one may say that the coefficients in the models, especially in the one-factor models, such as Vasicek (1977) model and Cox *et al.* (1985) model, are time-varying. Based on these studies, we may state that a short-rate model with constant parameters may not be valid for a long time period.

We introduce the time-varying Cox–Ingersoll–Ross (CIR) model and estimate it using a novel method: the local parametric approach (LPA). Our aim is to find the longest stable 'time homogeneous' interval for each time point t, where the parameters in the CIR model can be safely approximated as constants. Moreover, using this method, we can detect jumps and structural break points. Furthermore, this approach includes regime switching models as a special case, and it can also describe the time variation of the coefficients. Based on the parameters inside the selected interval, one may distinguish blooming and declining regimes of the economy. Moreover, the LPA has several attractive properties. First, it can capture the smooth time-varying property of parameters in interest rate models. The coefficients can be arbitrarily dependent on time: for instance, the smooth time trend. Second, this method allows for structural breaks and jumps in the parameter values; thus the length of the time homogeneous interval would depend on the time points of jumps or structure breaks. Third, there is no requirement concerning the number of observations before or after the break point.

The proposed approach has already been applied to different problems. Giacomini *et al.* (2009) considered time-varying copulae estimation, Cížek *et al*. (2009) applied it to compare the performance of global and time-varying autoregressive conditional heteroscedasticity (ARCH) and generalized ARCH (GARCH) specifications, and Härdle *et al.* (2010) applied this method to hierarchical Archimedean copulae, finding that the LPA can be used to detect both adaptive copulae parameters and local dependency structures.

To evaluate the performance of the LPA, we conduct both simulations and empirical studies. In the simulation exercise, we show that the proposed LPA is highly capable of detecting structural breaks, and all the true parameters are located in the pointwise confidence intervals of the estimators. In the empirical study, we use the 3-month Treasury bill rate as a proxy of the short rate and investigate the performance of the LPA compared to the time-varying CIR model by both in-sample fitting and out-of-sample forecasting via a comparison with moving window estimators.

The remainder of the paper is organized as follows. In the next section we provide a short review of standard interest rate models, and then we explain the LPA in detail in the third section. In the fourth section we present our simulation results. Empirical studies are presented in the fifth section. We conclude in the sixth section.

INTEREST RATE MODELS

In this section, we recall several standard one-factor short-rate models. One-factor short-rate models consider only one factor of uncertainty in the dynamics of the interest rate r_t. The general one-factor model can be written as the OU process:

$$\mathrm{d}r_t = \mu(r_t, \theta)\mathrm{d}t + \sigma(r_t, \theta)\mathrm{d}W_t$$

where $\mu(r_t, \theta)$ is the mean process, and $\sigma(r_t, \theta)$ stands for the volatility process, and W_t is a standard Brownian process. Specifically, we list several classical one-factor models:

Vasicek model (1977):

$$\mathrm{d}r_t = \kappa\{\mu - r_t\}\mathrm{d}t + \sigma\mathrm{d}W_t$$

where κ, μ and σ are constants. It is consistent with the mean reversion feature with a reversion speed κ to the long-run mean level μ. However, r_t can be negative in this model.

Cox *et al.* (CIR) model (1985):

$$\mathrm{d}r_t = \kappa\{\mu - r_t\}\mathrm{d}t + \sigma\sqrt{r_t}\mathrm{d}W_t \tag{1}$$

The drift function $a(r_t) = \kappa(\mu - r_t)$ is linear and possesses a mean reverting property, i.e. r_t moves in the direction of its long-run mean μ at speed κ. The diffusion function $\sigma^2(r_t) = r_t\sigma^2$ is proportional to the interest rate r_t and ensures that the process remains positive. Moreover, here r_t has a positive impact on the standard deviation through equation (1).

Hull–White model (1990):

$$\mathrm{d}r_t = \{\delta(t) - \kappa r_t\}\mathrm{d}t + \sigma \mathrm{d}W_t$$

This is an extended Vasicek model, where κ and σ are constant and $\delta(t)$ is a deterministic function of time. Moreover, this model uses the time-dependent reversion level $\delta(t)/\kappa$ for the long-run mean instead of the constant μ used in the Vasicek model.

Black–Karasinski model (1991):

$$\mathrm{d}\log r_t = \delta(t)\{\log \mu(t) - \log r_t\}\mathrm{d}t + \sigma(t)\mathrm{d}W_t$$

with $\delta(t)$, $\mu(t)$ and $\sigma(t)$ as a deterministic function of time, where $\mu(t)$ is the target interest rate. A drawback is that no closed-form formula for valuing bonds in terms of r_t can be derived by this model.

METHODOLOGY

In the Vasicek model, the interest rate r_t can be negative, whereas the CIR model guarantees the interest rate will be non-negative. In the Hull–White model, the volatility is a constant. The Black–Karasinski model assumes $\delta(t)$ and $\mu(t)$ are deterministic functions of time. Inherent to all these dynamics is that the coefficient functions cannot arbitrarily depend on time. This property may not be reasonable through a long economy developing process. Thus, in the paper, we introduce a time-varying one-factor model: specifically, the time-varying CIR model. The LPA allows the coefficients to be arbitrary functions of time, which further can be used to find the longest stable 'time homogeneous' interval for each time point, where the parameters in the CIR model can be safely assumed to be constant.

The time-varying CIR model is expressed as

$$\mathrm{d}r_t = \kappa_t\{\mu_t - r_t\}\mathrm{d}t + \sigma_t\sqrt{r_t}\mathrm{d}W_t \tag{2}$$

where W_t is the standard Wiener process. Denote the time-varying parameters as $\theta_t = (\kappa_t, \mu_t, \sigma_t)^\top$

This CIR model (2) includes all of the aforementioned parametric models, such as jump diffusion models, regime switching models, and even the nonparametric specified time-varying interest rate models.

For estimation, the discrete version of equation (2) is

$$r_{t+1} = a_t + (1 + b_t)r_t + \sigma_t\sqrt{r_t}z_t \tag{3}$$

where $a_t = \kappa_t\mu_t$, $b_t = -\kappa_t$, and $\{z_t\} \sim$ i.i.d. $N(0, 1)$. Hence the time-varying parameter set is redefined as $\theta_t = (a_t, b_t, \sigma_t)^\top$.

Likelihood function of the CIR process

Define $I_{t-1} = \{r_{t-1}, \ldots, r_1\}$ as the information set obtained at $t - 1$, and $\theta = (a, b, \sigma)$; then the conditional probability density function of r_t given I_{t-1} is

$$p(r_t|I_{t-1};\theta) = \frac{1}{\sqrt{2\pi\sigma^2 r_{t-1}}}\exp\left[-\frac{1}{2\sigma^2 r_{t-1}}\{r_t - (a + (1 + b)r_{t-1})\}^2\right] \tag{4}$$

Further, the log-likelihood function can be written as follows:

$$L(\theta) = \sum_{t=2}^{T}\log p(r_t|r_{t-1};\theta) \tag{5}$$

Now fix t; the MLE estimator $\tilde{\theta}_{I_k}$ in any interval $I_k = [t - m_k, t]$ is

$$\tilde{\theta}_{I_k} = \arg\max L_{I_k}(\theta) = \arg\max \sum_{i\in I_k}\log p(r_{t_{i+1}}|r_{t_i};\theta, \Delta t) \tag{6}$$

The accuracy of the estimation for a locally constant model with parameter θ_0 is measured via the log-likelihood ratio $L_{I_k}(\tilde{\theta}_{I_k}, \theta_0) = L_{I_k}(\tilde{\theta}_{I_k}) - L_{I_k}(\theta_0)$. In Cížek *et al.* (2009) it is proven that if Y_i follows a nonlinear process (2) then, given I_k for any $r > 0$, there exists a constant $\Re_r(\theta_0)$, such that

$$\mathsf{E}_{\theta_0} \left| L_{I_k} \left(\tilde{\theta}_{I_k}, \theta_0 \right) \right|^r \leq \Re_r(\theta_0) \tag{7}$$

Thus $\Re_r(\theta_0)$ can be treated as the parametric risk bound. This enables testing the parametric hypothesis on the basis of the fitted log-likelihood $L_{I_k}(\tilde{\theta}_{I_k}, \theta_0)$.

Test of homogeneous intervals

Mercurio and Spokoiny (2004), Cížek *et al.* (2009) and Spokoiny (2009) are informative references for the LPA. The general idea can be described as follows: suppose we have K (historical) candidate intervals with a starting interval I_0, i.e. $I_0 \subset I_1 \subset \ldots \subset I_K$, $I_k = [t - m_k, t]$ with $0 < m_k < t$. We increase the length from m_k to m_{k+1}, and test over the larger interval I_{k+1} whether $\tilde{\theta}_{k+1}$ is still consistent with $\tilde{\theta}_k$. To test an interval $I_k = [t - m_k, t]$, we fix the null hypothesis with a constant parameter $\theta_t \equiv \theta$. The alternative (a non-constant θ_t) is given by an unknown change point τ in I_k, i.e. $Y_{t'}$ follows one process when $t' \in J = [\tau + 1, t]$ with parameter θ_J, and it follows another process when $t' \in J^c = [t - m_{k+1}, \tau]$ with parameter θ_{J^c}, where $\theta_J \neq \theta_{J^c}$. With this alternative, the log-likelihood described in equation (5) can be expressed as $L_J \left(\tilde{\theta}_J \right) + L_{J^c} \left(\tilde{\theta}_{J^c} \right)$, giving the test statistics

$$T_{I_{k+1},\tau} = L_J \left(\tilde{\theta}_J \right) + L_{J^c} \left(\tilde{\theta}_{J^c} \right) - L_{I_{k+1}} \left(\tilde{\theta}_{I_{k+1}} \right) \tag{8}$$

where $\tau \in J_k = I_k \backslash I_{k-1}$ (see Figure 1). Because the change point $\tau \in I_k$ is unknown, we consider the maximum of the test statistics over J_k:

$$T_k = \max_{\tau \in J_k} T_{I_{k+1},\tau} \tag{9}$$

These statistics (equation (9)) are compared with critical values $\{\mathfrak{z}_k\}$; see below for more details.

The selected longest time homogeneous interval $I_{\hat{k}}$ satisfies

$$T_k \leq \mathfrak{z}_k, \quad \text{for } k \leq \hat{k} \tag{10}$$

and $T_{\hat{k}+1} > \mathfrak{z}_{\hat{k}+1}$. The interval $I_{\hat{k}}$ yields the adaptive estimator $\hat{\theta}_t = \hat{\theta}_{I_{\hat{k}}}$. The event $\{I_k \text{ is rejected}\}$ means that $T_\ell > \mathfrak{z}_\ell$ for some $\ell < k$, and hence a change point has been detected in the first k steps.

The local parametric approach

For any given t with intervals $I_0 \subset I_1 \subset \ldots \subset I_K$, the algorithm is described in four steps.

1. We estimate $\tilde{\theta}_{I_0}$ using the observations from the smallest interval $I_0 = [t - m_0, t]$, hence $\tilde{\theta}_{I_0}$ is always accepted.
2. We increase the interval to I_k, $(k \geq 1)$, find the estimator $\tilde{\theta}_{I_k}$ by MLE, and test homogeneity via equation (9); i.e. we test whether there is a change point in I_k. If equation (10) is fulfilled, we go on to step 3; otherwise we go to step 4.
3. Let $\hat{\theta}_{I_k} = \tilde{\theta}_{I_k}$, then further set $k = k + 1$, and continue to step 2.
4. Accept as the longest time homogeneous interval $I_{\hat{k}} = I_{k-1}$ and define the local adaptive estimator as $\hat{\theta}_{I_{\hat{k}}} = \tilde{\theta}_{I_{k-1}}$. Additionally, set $\hat{\theta}_{I_{\hat{k}}} = \hat{\theta}_{I_k} = \ldots = \hat{\theta}_{I_K}$ for all $k > \hat{k}$.

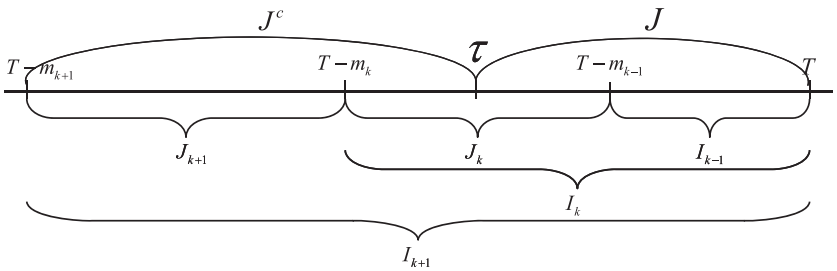


Figure 1. Construction of the test statistics for LPA: the involved interval I_k and J_k

For a change point τ in I_k, we obtain $\hat{k} = k - 1$; therefore $I_{\hat{k}} = I_{k-1}$ is the selected longest time homogeneous interval. We compare the test statistics with the critical value. If they are smaller than the critical value $\mathfrak{z}_k$ for interval I_k, we accept I_k as the time homogeneous interval, at which point we increase the interval to I_{k+1} and perform the test again. We repeat this procedure sequentially until we stop at some $k < K$ or we exhaust all the chosen intervals. For each time point t, we use the same algorithm, although we do not need to calculate the critical values a second time because they depend on only the parametric specification and the length of interval m_k.

To investigate the performance of the adaptive estimator, we introduce the small modelling bias (SMB). The SMB for interval I_k is

$$\Delta_{I_k}(\theta) = \sum_{t \in I_k} \mathcal{K}\{r_t, r_t(\theta)\} \tag{11}$$

with $\mathcal{K}$ the Kullback–Leibler (KL) divergence:

$$\mathcal{K}\{r_t, r_t(\theta)\} = \mathsf{E} \log \frac{p\{r_t\}}{p\{r_t(\theta)\}} \tag{12}$$

where $p(\cdot)$ and $p(\cdot;\theta)$ are the probability density functions of r_t and $r_t(\theta)$ respectively. The SMB measures in terms of KL divergence the closeness of a constant parametric model with $p(\cdot;\theta)$ to a time-varying nonparametric model with $p(\cdot)$. Suppose now for a fixed $\Delta > 0$

$$\mathsf{E}\, \Delta_{I_k}(\theta) \leq \Delta \tag{13}$$

Inequality (13) simply means that for some $\theta \in \Theta$, $\Delta_{I_k}(\theta)$ is bounded by a small constant, implying that the time-varying model can be well approximated (over I_k) by a model with a fixed parameter θ.

Under the SMB condition (equation (13)) for some interval I_k and $\theta \in \Theta$, one has with a risk bound $\mathfrak{R}_r(\theta)$:

$$\mathsf{E} \log \left\{ 1 + \frac{\left| L_{I_k}\left(\tilde{\theta}_{I_k}, \theta\right)\right|^r}{\mathfrak{R}_r(\theta)} \right\} \leq 1 + \Delta \tag{14}$$

If Δ is not large, equation (14) extends the parametric risk bound $\mathfrak{R}_r(\theta)$ to the nonparametric situation; for details see Cížek *et al*. (2009). An 'oracle' choice I_{k*} from the set $I_0, \ldots, I_K$ exists, which is defined as the largest interval satisfying equation (13). We denote the corresponding 'oracle' parameter as $\theta_{I_{k*}}$.

Two types of error occur in this algorithm, however: the first type is the rejection of the time homogeneous interval earlier than the 'oracle' step, which means $\hat{k} \leq k^*$. The other type is the selection of a homogeneous interval larger than the 'oracle', i.e. $\hat{k} > k^*$. The first type of error can be treated as a 'false alarm'; i.e. the algorithm stops earlier than the 'oracle' interval I_{k*}, which leads to selecting an estimate with a larger variation than $\theta_{I_{k*}}$. The second type of error arises if $\hat{k} > k^*$. Outside the oracle interval, we are exploiting data that do not support the SMB condition. Both errors will be specified in a propagation and stability condition in the next section.

Choice of critical values

The accuracy of the estimator can be measured by the log-likelihood ratio $L_{I_k}\left(\tilde{\theta}_{I_k}, \theta_0\right)$, which is stochastically bounded by the exponential moments (equation (14)). In general, $\tilde{\theta}_{I_k}$ differs from $\hat{\theta}_{I_k}$ only if a change point is detected at the first k steps. A small value of the likelihood ratio means that $\hat{\theta}_{I_k}$ belongs to the confidence set based on the estimate of $\tilde{\theta}_{I_k}$; i.e. statistically we 'accept' $\hat{\theta}_{I_k} = \tilde{\theta}_{I_k}$. If the procedure stops at some $k \leq K$ by a false alarm, i.e. a change point is detected in I_k with the adaptive estimator $\hat{\theta}_{I_k}$, then the accuracy of the estimator can be expressed via the propagation condition

$$\mathsf{E}_{\theta_0} \left| L_{I_k}\left(\tilde{\theta}_{I_k}, \hat{\theta}_{I_k}\right)\right|^r \leq \rho \mathfrak{R}_r(\theta_0) \tag{15}$$

In the parametric situation, we can calculate the left-hand side of equation (15) and choose the critical value $\mathfrak{z}_l$ based on this inequality. The situation in the first k steps can be distinguished as one of two cases: There is a change point detected at some step $l \leq k$, or there is no change point in the first k intervals. We denote by $\mathfrak{B}_l$ the event of rejection at step l, i.e.

$$\mathfrak{B}_l = \{T_1 \leq \mathfrak{z}_1, \ldots, T_{l-1} \leq \mathfrak{z}_{l-1}, T_l > \mathfrak{z}_l\} \tag{16}$$

and $\hat{\theta}_{I_k} = \tilde{\theta}_{I_{l-1}}$ on $\mathfrak{B}_l$, $l = 1, 2, \ldots, k$. Now choose $\mathfrak{z}_1$ by minimizing the following equation:

$$\max_{k=1,\ldots,K} \mathsf{E}_{\theta_0} \left| L\left(\tilde{\theta}_{I_k}, \tilde{\theta}_{I_0}\right)\right|^r \mathbf{1}(\mathfrak{B}_1) = \rho \mathfrak{R}_r(\theta_0)/K \tag{17}$$

For $\mathfrak{z}_l, l \geq 2$, we use the same algorithm to calculate them. The event $\mathfrak{B}_l$ depends on $\mathfrak{z}_1, \ldots, \mathfrak{z}_l$. Because $\mathfrak{z}_1, \ldots, \mathfrak{z}_{l-1}$ have been fixed by previous steps, the event $\mathfrak{B}_l$ is controlled only by $\mathfrak{z}_l$. Hence the minimal value of $\mathfrak{z}_l$ should ensure

$$\max_{k \geq l} \mathsf{E}_{\theta_0} \left| m_k \mathcal{K}\left(\tilde{\theta}_k, \tilde{\theta}_{l-1}\right)\right|^r \mathbf{1}(\mathfrak{B}_l) = \rho \mathfrak{R}_r(\theta_0)/K \tag{18}$$

or we can express the criterion via the log-likelihood ratio:

$$\max_{k \geq l} \mathsf{E}_{\theta_0} \left| L\left(\tilde{\theta}_{I_k}, \tilde{\theta}_{I_{l-1}}\right)\right|^r \mathbf{1}(\mathfrak{B}_l) = \rho \mathfrak{R}_r(\theta_0)/K \tag{19}$$

where ρ and r are two global parameters and m_k denotes the number of observations in I_k. The role of ρ is similar to the level of the test in hypothesis testing problems, while r describes the power of the loss function. We apply $r = 1/2$ in both the simulation and the real data analysis because it makes the procedure more stable and robust against outliers. We also choose $\rho = 0.2$; however, other values in the range [0.1, 1] lead to similar results, as referenced in Spokoiny (2009).

The critical value $\mathfrak{z}_l$ that satisfies equation (19) can be found numerically by Monte Carlo simulations from the parametric model. It is a decreasing function with respect to the log length of interval. When the interval is small, it is easier to accept it as the time homogeneous interval because there cannot be many jumps due to the short length. If we increase the length of interval, however, as more observations are included, it will contain more uncertain information, especially when large jumps or visible structural breaks can exist in the interval. In this case, it therefore tends to reject the test statistics, and the corresponding critical values will decrease as well.

The length of the tested interval is assumed to increase geometrically with $m_k = [m_0 a^k]$. m_0 is the length of initial interval I_0, which is time homogeneous as default. a can be chosen from 1.1 to 1.3. The experiments reveal, however, that the estimated results are not sensitive to the choice of a. In the time-varying CIR model, three parameters must be estimated. To guarantee a reasonable quality of the estimation, a large sample size is required. Therefore, we choose the length of the initial interval I_0 as $m_0 = 60$ and $a = 1.25$. As already discussed, interest rates are influenced by macroeconomic structures and may also be subject to regime shifts. Therefore, the longest interval we choose is supposed to cover one regime, and by initiation at least one change point should exist between the expansion and recession regimes. Referring to a business cycle of approximately 4 years, we choose the number of intervals $K = 13$ so that $m_K = 1100$ is the longest tested time homogeneous interval used in both the simulation and empirical exercises in this paper.

'Oracle' property of the estimators

In this section, we discuss the 'oracle' properties of the LPA estimators. Recall that for the 'oracle' choice k^* equation (13) holds. It also holds for every $k \leq k^*$ but does not hold for any $k \geq k^*$. However, the 'oracle' choice I_{k^*} and $\theta_{I_{k^*}}$ are, of course, unknown. The LPA algorithm works to mimic these oracle values. In Cížek *et al*. (2009) it is proven that under the SMB condition, i.e. when equation (13) holds, the 'oracle' property of the LPA estimator $\hat{\theta}_{I_{\hat{k}}}$ satisfies the following property:

For $\theta \in \Theta$ and letting $\max_{k \leq k^*} \mathsf{E} \left| L\left(\tilde{\theta}_{I_{k^*}}, \theta\right)\right|^r \mathbf{1}(\mathfrak{B}_1) \leq \mathfrak{R}_r(\theta)$, we have

$$\mathsf{E} \log \left\{ 1 + \frac{\left| L_{I_{k^*}}\left(\tilde{\theta}_{I_{k^*}}, \theta\right)\right|^r}{\mathfrak{R}_r(\theta)} \right\} \leq 1 + \Delta \tag{20}$$

Further, we obtain

$$\mathsf{E} \log \left\{ 1 + \frac{\left| L_{I_{k^*}}\left(\tilde{\theta}_{I_{k^*}}, \hat{\theta}_{I_{\hat{k}}}\right)\right|^r}{\mathfrak{R}_r(\theta)} \right\} \leq \rho + \Delta \tag{21}$$

This property tells us that although the false alarm occurs before the 'oracle' choice, i.e. $\hat{k} \leq k^*$, under the SMB condition, the adaptive estimator $\hat{\theta}_{I_{\hat{k}}}$ does not stray far from the oracle value, which implies the LPA estimator does not introduce large errors into the estimation.

However, the SMB condition does not hold if $\hat{k} > k^*$, which means the detected interval is larger than the 'oracle' interval; then, the LPA estimator $\hat{\theta}_{I_{\hat{k}}}$ satisfies Theorem 4.3 in Cížek *et al.* (2009):

Let $\mathsf{E}\,\Delta_{I_{k^*}(\theta)} \leq \Delta$ for $k^* \leq K$, then $L_{I_{k^*}}\left(\tilde{\theta}_{I_{k^*}}, \hat{\theta}_{I_{\hat{k}}}\right)\mathbf{1}\left(\hat{k} \geq k^*\right) \leq \mathfrak{z}_{k^*}$:

$$\mathsf{E}\log\left\{1 + \frac{\left|L_{I_{k^*}}\left(\tilde{\theta}_{I_{k^*}}, \hat{\theta}_{I_{\hat{k}}}\right)\right|^r}{\mathfrak{R}_r(\theta)}\right\} \leq \rho + \Delta + \log\left\{1 + \frac{\mathfrak{z}_{k^*}^r}{\mathfrak{R}_r(\theta)}\right\} \tag{22}$$

This indicates that $\hat{\theta}_{I_{\hat{k}}}$ belongs with a high probability to the confidence interval of the oracle estimate $\tilde{\theta}_{I_{k^*}}$; i.e. it is still a reliable approximation of the oracle value $\theta_{I_{k^*}}$.

SIMULATION STUDY

In this section, we first calculate the critical values via simulation. As described in Spokoiny (2009), the critical values are not sensitive to the parameter setting, although they are crucial to the model setting. Therefore, in the paper, the parameters used for the calculation of critical values are as follows: $r_0 = 5.2$, $a = 0.0024$, $b = -0.0015$ and $\sigma = 0.059$; which are estimated from the real data. The performance of the critical values is described in Figure 2. One can note that the critical value decreases as the length of the intervals increases, which is consistent with the theory mentioned above. Moreover, we also change the parameter settings for the simulation, while under the null hypothesis there are no especially significant differences between the critical values in different scenarios.

We also evaluate the performance of the LPA using Monto Carlo simulation. We designed several scenarios with the structural breaks at different times for the three parameters and find satisfactory results. For brevity, we concentrate here on one scenario in which we simultaneously change all three parameters $(a_t, b_t, \sigma_t)^\top$ and assume there are two structural break points for each parameter in the process. We simulate the CIR process with two structural breaks in θ and a total of 2250 observations over 200 simulations, where the parameters are estimated by the real data samples. Table I summarizes the parameter settings for simulations of the CIR model, the chosen values located in the range of estimators from the global CIR model with the 3-month Treasury bill rate that is used in the empirical analysis.

The estimators $\hat{a}$, $\hat{b}$, and $\hat{\sigma}$ are described in Figures 3–6. In each figure, the blue line depicts the mean of the corresponding estimators from the 200 simulations, the two dotted red lines are the 5%–95% pointwise confidence intervals for the estimators, and the black line describes the respective real parameters. We use the first 250 data points as the training set referring to the moving window estimator, after which we estimate the CIR model using the LPA from time point 251 to 2250. One can observe that for the parameter a the LPA estimator $\hat{a}$ covers the true parameter a

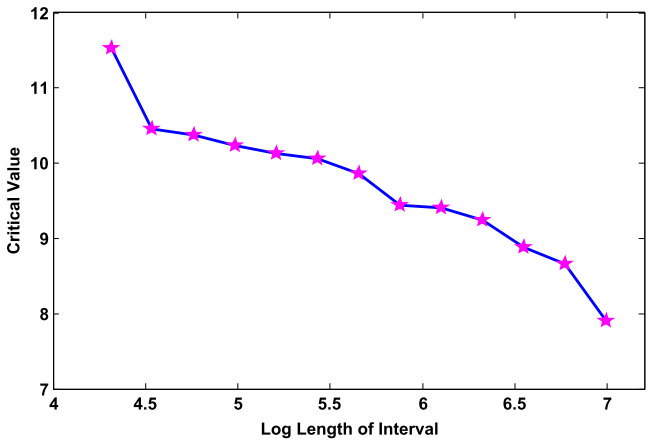


Figure 2. Critical values with $m_0 = 60$, $K = 13$ from 500 simulations. [Colour figure can be viewed at wileyonlinelibrary.com]

Table I. Parameter settings for simulations of the CIR process

t	a	b	σ
$t \in [1, 750]$	0.533	−0.103	0.022
$t \in [751, 1500]$	0.115	−0.073	0.050
$t \in [1501, 2250]$	0.373	−0.132	0.084

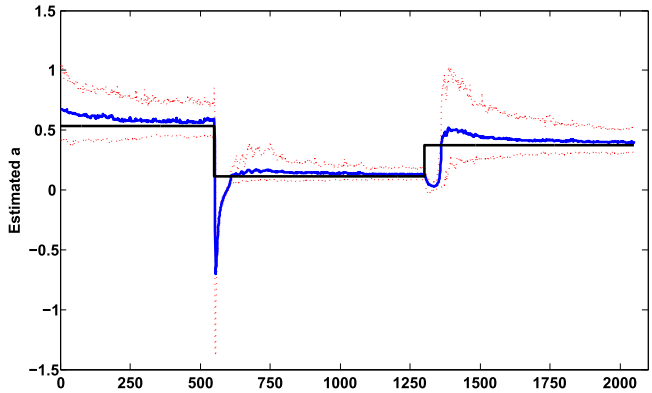


Figure 3. LPA estimator $\hat{a}$ with simulated CIR paths. The dotted red lines are the 5%–95% pointwise confidence intervals of $\hat{a}$, the blue line is the mean of $\hat{a}$, and the black line indicates the true process of a set in Table I. [Colour figure can be viewed at wileyonlinelibrary.com]

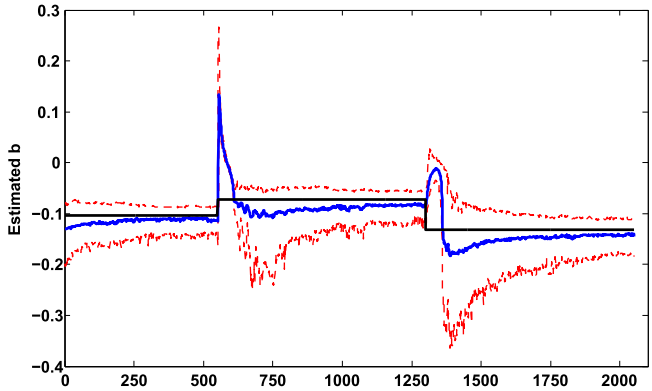


Figure 4. LPA estimator $\hat{b}$ with simulated CIR paths. The dotted red lines are the 5%–95% confidence interval of $\hat{b}$, the blue line is the mean of $\hat{b}$, and the black line indicates the true process of b set in Table I. [Colour figure can be viewed at wileyonlinelibrary.com]

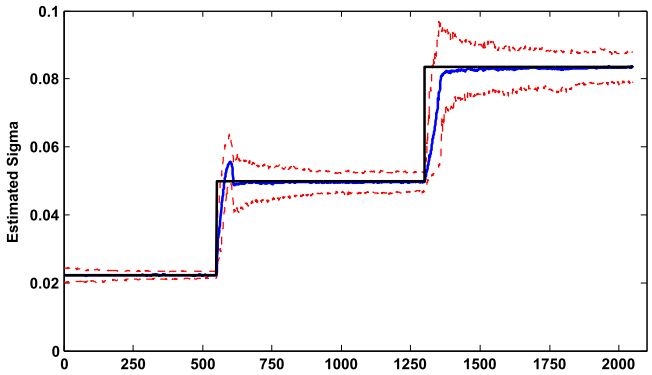


Figure 5. LPA estimator $\hat{\sigma}$ with simulated CIR paths. The dotted red lines are the 5%–95% confidence interval of $\hat{\sigma}$, the blue line is the mean of $\hat{\sigma}$, and the black line indicates the true process σ set in Table I. [Colour figure can be viewed at wileyonlinelibrary.com]

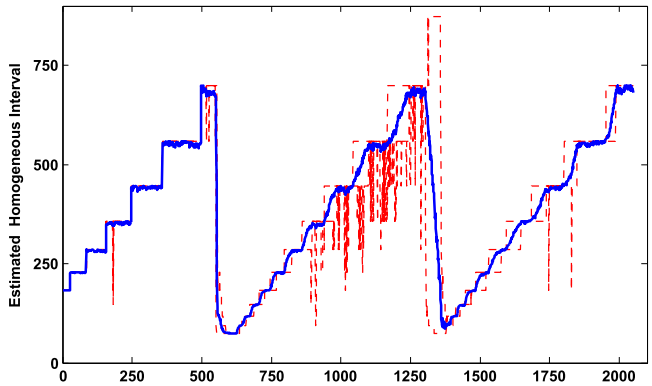


Figure 6. The length of time homogeneous intervals for simulated CIR paths. The dotted red lines are the 5–95% confidence interval and the blue line is the mean of estimated length of time homogeneous intervals.[Colour figure can be viewed at wileyonlinelibrary.com]

quite well, which is described in Figure 3. There are clearly two jump points, located around time point 500 and 1250, respectively, which are the structural break points designed in the simulation. Figure 4 presents the performance of the LPA estimator $\hat{b}$. Its performance is quite reasonable. Taking a little delayed time into consideration, the performance of $\hat{b}$ coincides with the true process, as does the LPA estimator $\hat{\sigma}$, as shown in Figure 5. Briefly, it is worth noting that the performance here is preferable to that of both $\hat{a}$ and $\hat{b}$. The structural change points are evident in Figure 5. Both the mean process and the volatility process of the estimator have the same pattern as the true parameter path, which indicates the LPA can capture precise information for the stochastic process when structural breaks occur.

Figure 6 depicts the selected longest time homogeneous interval for each time point. The dotted red lines are the 5–95% confidence interval and the blue line is the mean of the estimated length of time homogeneous intervals. One can compare the selected homogeneous intervals with the LPA estimators in other figures, all of which provide consistent evidence for its performance. In the initial setting, we have two jumps respectively at 750 and 1500. One can easily detect in Figure 6 that the two jump points are located close to 500 and 1250, due to some delayed time. Further, both the 5–95% pointwise confidence intervals and the mean of the length of the selected intervals coincide with the parameter settings, which coincide with the estimators.

EMPIRICAL STUDY

Data description

We use the 3-month Treasury bill rate from the Federal Reserve Bank of St Louis as a proxy for the short rate. This rate has been used frequently in the term structure literature. The data range from 2 January 1998 to 31 December 2013, and the summary statistics are shown in Table II.

Table II. Statistical summary of 3-month Treasury bill rate: 2 January 1998 to 31 December 2013

	Mean	SD	Skewness	Kurtosis
r_t	2.4419	2.0317	0.2441	1.5225
$\mathrm{d}r_t$	−0.0013	0.0574	−0.6699	82.7109

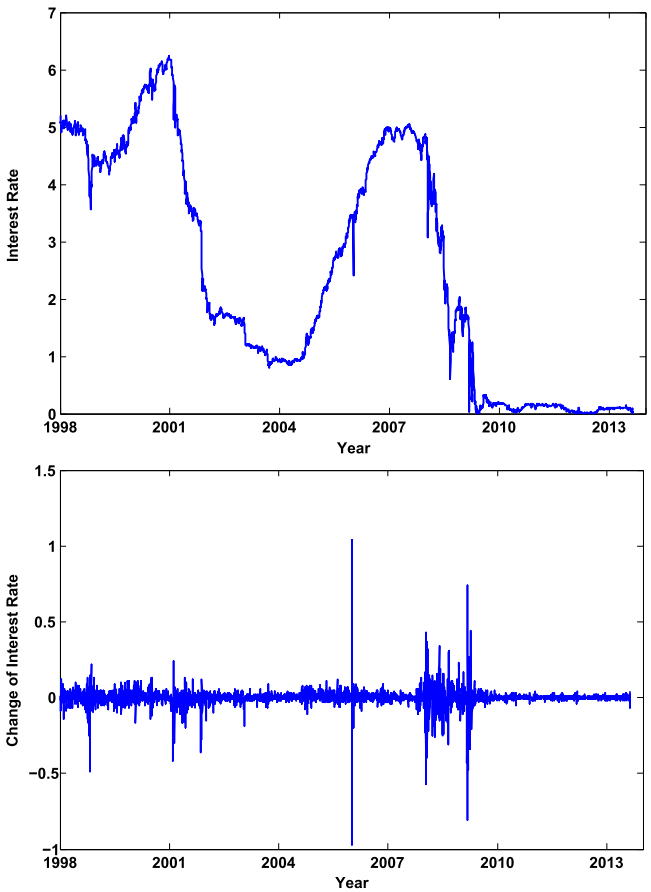


Figure 7. Three-month Treasury bill rate: 2 January 1998 to 31 December 2013. Top panel: daily yields; bottom panel: changes of daily yields. [Colour figure can be viewed at wileyonlinelibrary.com]

Table III. Estimated parameters of CIR model by MLE in different time periods

Sample size	$\hat{a}$	$\hat{b}$	$\hat{\sigma}$
2 Jan 1998–31 Dec 2013	0.0024	−0.0015	0.0590
2 Jan 1998–31 Jul 2007	0.0012	−0.0004	0.0285
1 Aug 2007–31 Dec 2009	0.0116	−0.0097	0.1273
2 Jan 2010–31 Dec 2013	0.0026	−0.0292	0.0428

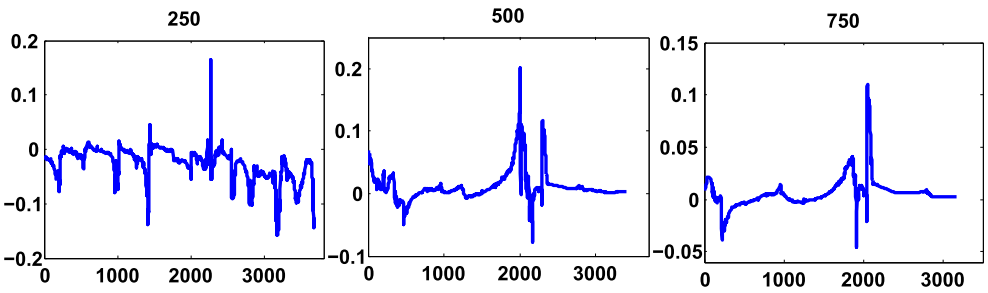


Figure 8. Moving window estimator $\hat{a}$ with window sizes 250, 500 and 750 (from left to right). [Colour figure can be viewed at wileyonlinelibrary.com]

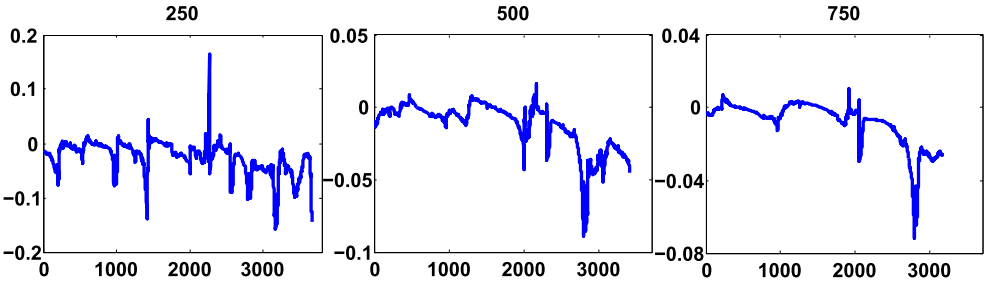


Figure 9. Moving window estimator $\hat{b}$ with window sizes 250, 500 and 750 (from left to right).[Colour figure can be viewed at wileyonlinelibrary.com]

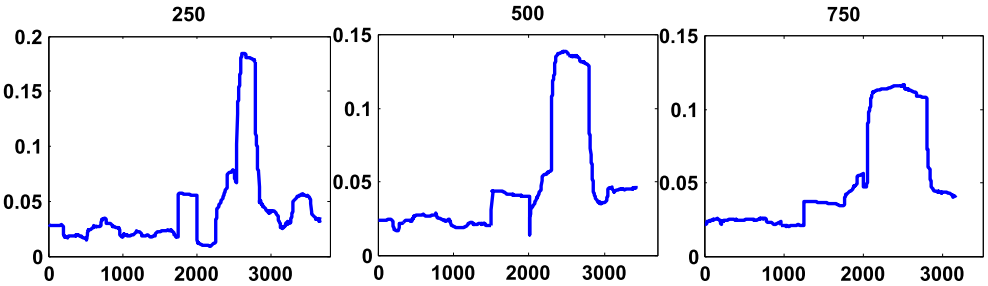


Figure 10. Moving window estimator $\hat{\sigma}$ with window sizes 250, 500 and 750 (from left to right). [Colour figure can be viewed at wileyonlinelibrary.com]

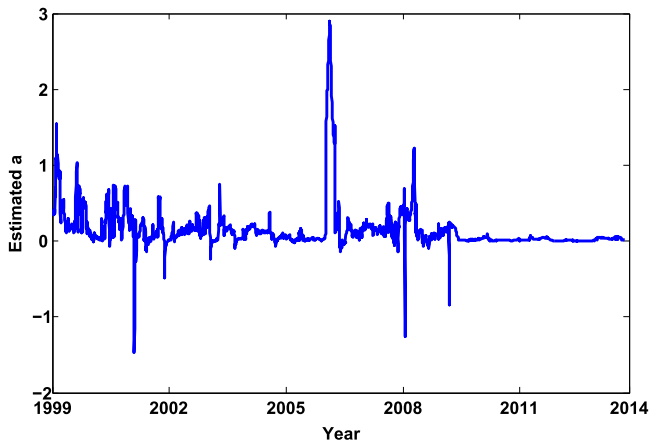


Figure 11. Estimated $\hat{a}$ for CIR model using 3-month Treasury bill rate by the LPA. [Colour figure can be viewed at wileyonlinelibrary.com]

The short rate and its daily change are displayed in Figure 7. The volatility of the short rate clearly changes over time, as mentioned above. Moreover, there are several jumps and structural breaks across the entire time period. Specifically, the short rate from 1999 to 2001 is relatively less volatile; however, from mid 2007 to 2009, the volatility is much higher than in other periods. One more noticeable phenomenon is that after the financial crisis both the level and the difference value of the short rate decrease greatly and stay low until the end of the study period and, therefore, with quite a small volatility. The variation of the short rate is time-varying; after fitting the CIR model separately with three different time periods, the estimation results are presented in Table III. The first row in the table uses the entire sample, the second row comprises observations from the beginning of 1998 to the end of July 2007, the third estimate period is during the financial crisis, and the final row shows the estimation results after the financial crisis. All three

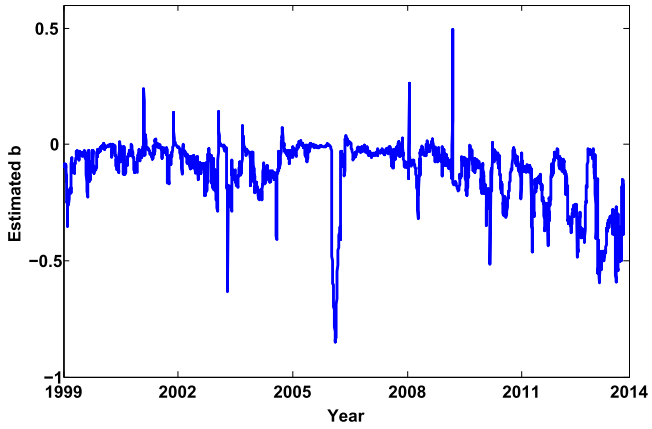


Figure 12. Estimated $\hat{b}$ for CIR model using 3-month Treasury bill rate by the LPA. [Colour figure can be viewed at wileyonlinelibrary.com]

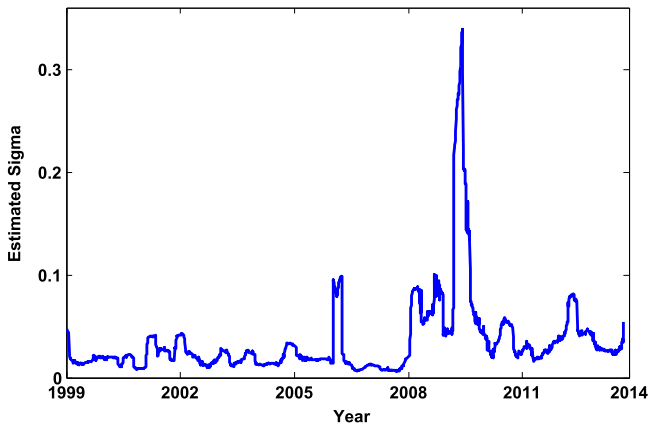


Figure 13. Estimated $\hat{\sigma}$ for CIR model using 3-month Treasury bill rate by the LPA.[Colour figure can be viewed at wileyonlinelibrary.com]

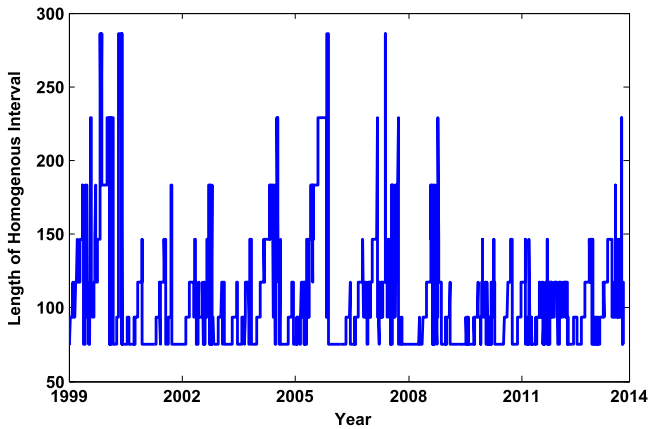


Figure 14. The selected longest time homogeneous intervals using 3-month Treasury bill rate with $\rho = 0.2$, and $r = 0.5$. The first reported time is 1999. [Colour figure can be viewed at wileyonlinelibrary.com]

parameters are significantly different during the three different periods. For instance, $\hat{a}$ grew approximately 10 times during the financial crisis period, compared with other periods or even the entire period. $\hat{b}$ also performs differently during each period, as does the estimated volatility $\hat{\sigma}$. It is relatively low from 1998 to 2007, then increases to 0.1273 during the financial crisis. After the crisis, the volatility is also higher than expected, which can also be verified by Figure 7.

First, we use the moving window estimation to investigate the stability of all three parameters in the CIR model. We specify three different window sizes as $l = 250$, $l = 500$ and $l = 750$, corresponding to 1-year, 2-year and

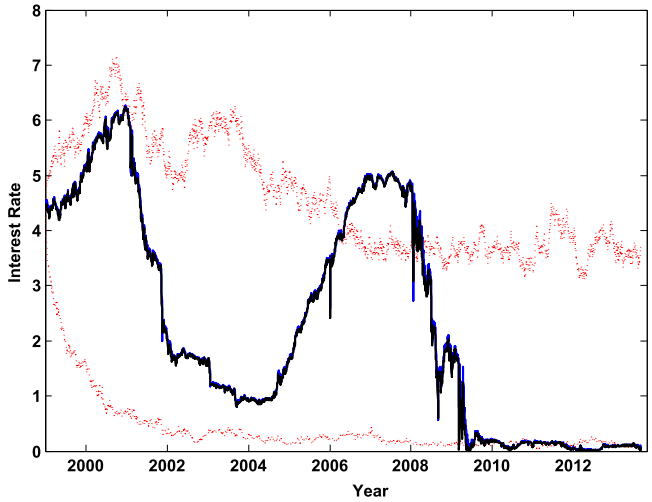


Figure 15. In-sample fitting for CIR model using 3-month Treasury bill rate. The black line is the real data; the blue line is the fitted CIR path with the estimators by LPA; the two red lines are 5%–95% confidence intervals simulated with the global estimators. [Colour figure can be viewed at wileyonlinelibrary.com]

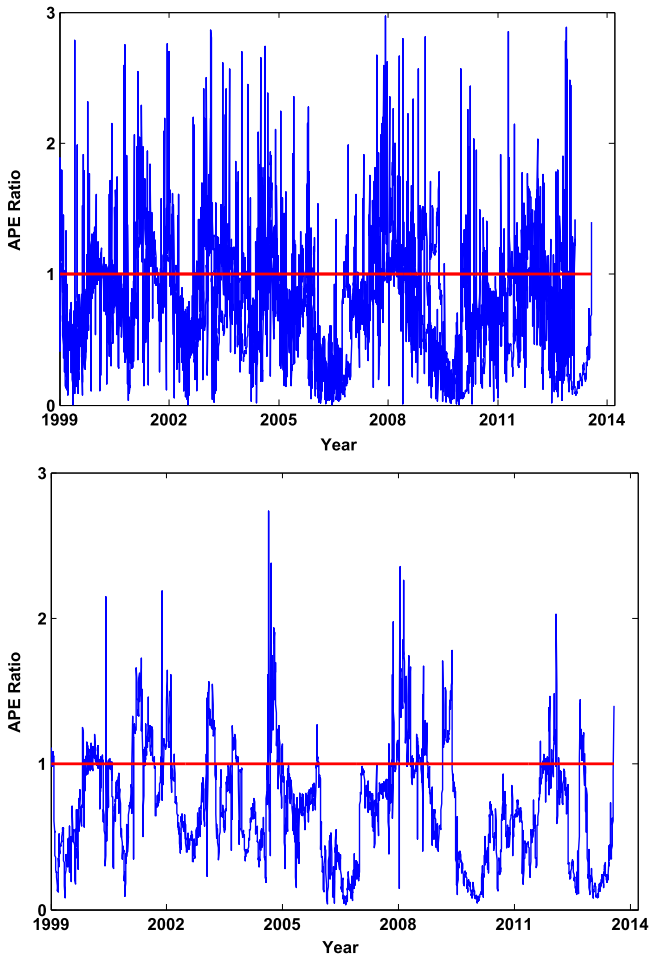


Figure 16. Ratio of the absolute prediction errors between the estimators by LPA (numerator) and moving window estimator (denominator) with window size 250. Top: 1-day horizon; bottom: 10-day horizon. [Colour figure can be viewed at wileyonlinelibrary.com]

3-year window size. Figures 8, 9 and 10 separately present the moving window estimators $\hat{a}$, $\hat{b}$ and $\hat{\sigma}$. Fairly similar performances are illustrated in both $\hat{a}$ and $\hat{b}$. One can find that large variations exist in the moving window estimation process. $\hat{a}$ shows high variation, especially around observation 2000, approximately the financial crisis time period, while before and after the period $\hat{a}$ performance is relatively stable. Similarly, for $\hat{b}$, there is clear pattern during different periods, while across the periods, the variation is still quite high, even though it begins to decrease after the crisis. Volatility $\hat{\sigma}$, however, performs in a much more stable way. The pattern is quite clear: before the financial crisis the volatility is relatively low, whereas during the crisis period it jumps to a high level, and then decreases once again after the crisis.

We then apply the LPA to estimate the time-varying CIR model. The estimation results are displayed in Figures 11–14. The performance of $\hat{a}$ from the LPA is very similar to that of the moving window estimator $\hat{a}$. It retains a relatively stable pattern, with some exceptions; during the financial crisis the variation of $\hat{a}$ increases significantly, while after the crisis it performs quite stably. The performance of $\hat{b}$ varies differently in different periods. It is relatively stable from 1999 to 2008, after which its variation becomes larger from 2008 to the end of the study period, during which time it also shows a decreasing tendency. $\hat{\sigma}$ shows a relatively stable performance compared with the other two estimators in the CIR model during the entire time series, and it shows a clearer pattern that is consistent with the behaviour of the length of the selected time homogeneous interval described in Figure 14. Moreover, one can easily detect the largest structural break points: from 2001 to 2008, the fluctuation of $\hat{\sigma}$ is relatively small, while after 2008 the variation becomes quite large, especially during the financial crisis period.

Figure 14 describes the selected time homogeneous interval for each time point t. Here we evaluate the estimation starting from year 1999, and treat the first year as a time homogeneous interval. One can note that the interval $I_{\hat{k}}$ can drop rapidly when the LPA diagnoses a change point. After a drop, the intervals increase slowly as the LPA gains more confidence in the stability of its parameters. Moreover, the length of the selected time homogeneous interval is closely correlated with the regimes of the macroeconomy. The recession regime induces shorter homogeneous intervals, while the length is extended in the expansion periods where the macroeconomy is in a stable state. Let us

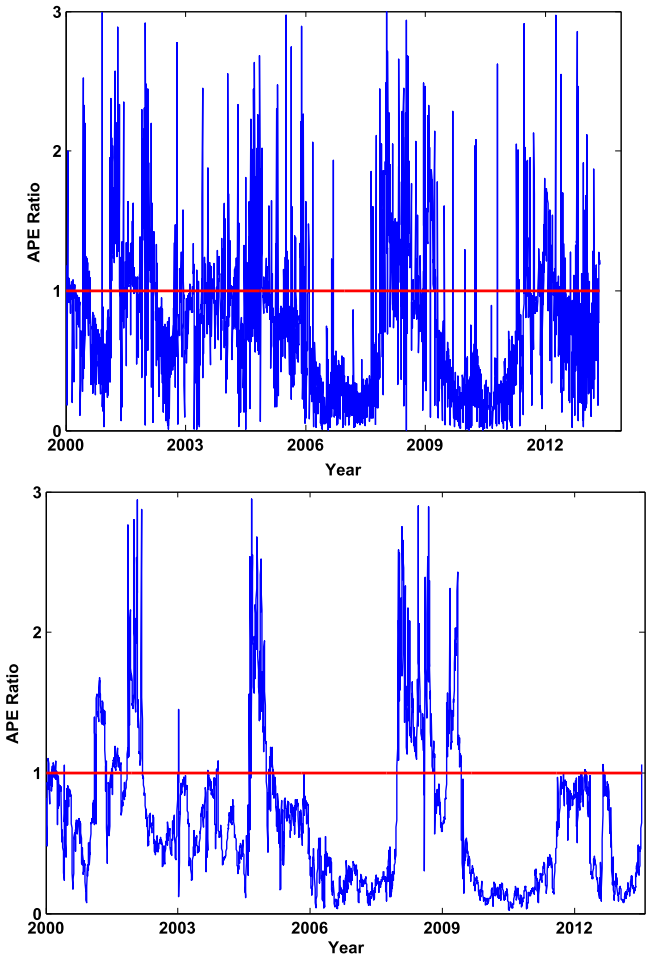


Figure 17. Ratio of the absolute prediction errors between the estimators by LPA (numerator) and moving window estimator (denominator) with window size 500. Top: 1-day horizon; bottom: 10-day horizon. [Colour figure can be viewed at wileyonlinelibrary.com]

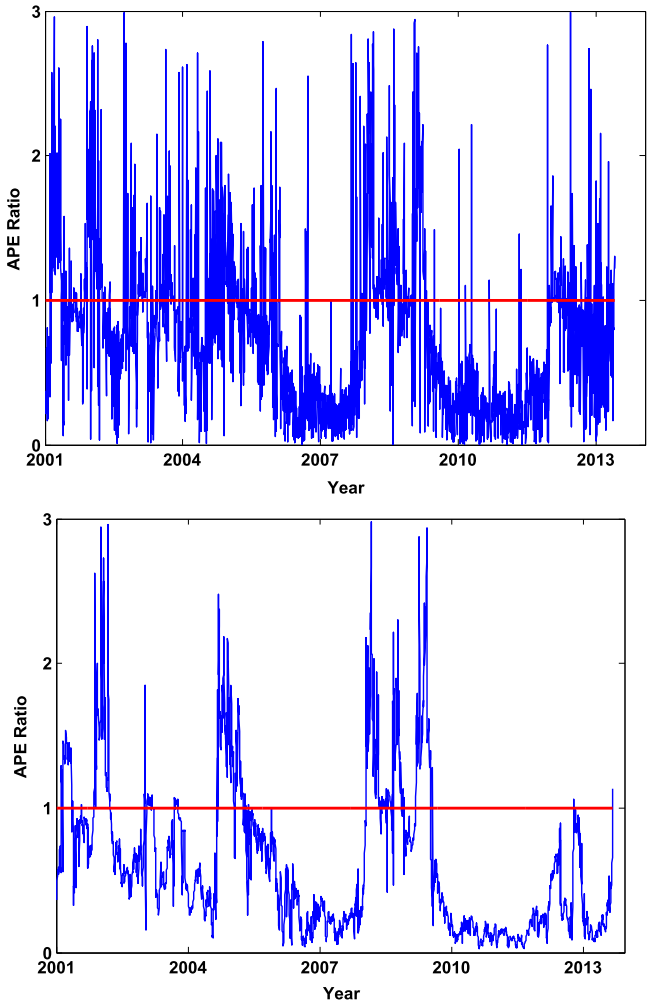


Figure 18. Ratio of the absolute prediction errors between the estimators by LPA (numerator) and moving window estimator (denominator) with window size 750. top: 1-day horizon; bottom: 10-day horizon. [Colour figure can be viewed at wileyonlinelibrary.com]

first analyse the short rate before 2001. In that period, economic activity continued to expand briskly, and the variation of the short rate was relatively small. Then, in the period from 2001 to 2003, the US economy went into recession, influenced by the terrorist attack on 11 September 2001, the 2002 stock market crash and the war in Iraq in 2003. These events induced a fairly unstable macroeconomy marked by increased oil prices, overstretched investment and excessively high productivity. Further, these factors led to short selected homogeneous intervals. From 2004 until 2007, the economy headed towards a stable state again. The selected intervals lasted longer than before. In 2008, due to the financial crisis, the situation reversed itself and another global recession began. Again, it can be confirmed by the shorter length of the selected intervals and the fact that the interest rate remains quite volatile. After the financial crisis, however, the economy continued to develop at a very low speed while also remaining in a recession regime, which indicates that the length of time homogeneous interval still does not hold for very long.

Figure 15 depicts the in-sample fit. The real data are described by the black line, and the two red dashed lines indicate the 5–95% pointwise confidence intervals from the simulated data, which is the same as that used in calculating the critical values. The blue line is the in-sample fit path with the values estimated by the LPA. It is clear that the fitted sample path by the LPA estimator matches the real data quite well; i.e. the LPA has an acceptable performance for in-sample fit. The structural break points from the fitted LPA path occur very close to the real data path. Moreover, there are two periods when both the real data and the fitted value are located outside the confidence interval, which also indicates that the CIR model with constant parameters cannot capture the dynamics of interest rate particularly well.

We further evaluate the out-of-sample forecasting performance. To compare the forecasting result of the LPA with the moving window estimation results, we take the absolute prediction error (APE) as the criterion. It is defined over a prediction period horizon $\mathcal{H}$, $\mathrm{APE}(t) = \sum_{h\in\mathcal{H}} |r_{t+h} - \hat{r}_{t+h|t}|/|\mathcal{H}|$, where $\hat{r}_{t+h|t}$ represents the interest rate prediction by a particular model. Both 1-day- and 10-day-ahead forecasting are considered. Figures 16–18 show the comparison results. In each figure, the top panel shows the forecast ratio for the 1-day horizon, while the bottom panel shows the 10-day horizon. There is no doubt that the LPA performs well, especially for the long horizon forecasting.

Table IV. Forecast evaluation criteria for 1-day and 10-day horizons of the short rate based on the LPA and moving window (MW) estimation

Horizon		MAPE		
		$l = 250$	$l = 500$	$l = 750$
1 day	LPA	**0.0448**	**0.0450**	**0.0450**
	MW	0.0516	0.0549	0.0553
10 days	LPA	**0.1971**	**0.2006**	**0.2020**
	MW	0.2640	0.2918	0.2962

Note: The left-hand columns refer to the forecasting horizon; the right-hand columns represent the mean absolute prediction error (MAPE) according to different moving window sizes.

First, one can observe that the LPA is generally preferable compared to the moving window estimation in one-step-ahead prediction. The results are better for the 10-day forecast horizon, when the LPA performs better than the MW by a large percentage.

Table IV summarizes the prediction performance for the LPA and the moving window (MW) estimations with forecast horizons of 1 day and 10 days. We compare the mean of absolute prediction errors (MAPE) for each method. Note that for 1-day-ahead forecasting there is no significant difference between the results from both methods, and both of their MAPEs are quite small; still, the LPA does perform slightly better than the MW. Over the 10-day horizon, however, the difference in quality increases, and the accuracy of the MW decreases greatly compared with the LPA. The larger the size of the window, the larger is the MAPE from the MW estimation method.

CONCLUSION

There is considerable statistical evidence, in addition to economic reasons, indicating that the short-rate process does not follow stable stochastics. We apply a modern statistical method to describe the changing dynamics of the short rate. With the simple CIR model and the LPA, we detect structural breaks for the short-rate process, which is consistent with the conclusion from the existing literature. Our study proves that interest rate dynamics are not stable. Moreover, We obtain time homogeneous intervals for each time point, which is useful to explain the structural breaks. We further compare our results with moving window estimators, and the results show that the LPA performs better in both in-sample fit and out-of-sample forecasting, independent of whether data come from a stable period.

ACKNOWLEDGEMENTS

We thank our Editor, Associate Editor and referees for their very helpful comments. Financial support from the Deutsche Forschungsgemeinschaft via SFB 649 'Ökonomisches Risiko', Humboldt-Universität zu Berlin and from the NSFC 71401141, China, is gratefully acknowledged.

REFERENCES

Aït-Sahalia Y. 1996. Testing continuous-time models of the spot interest rate. *Review of Financial Studies* **9**: 385–426.
Ang A, Bekaert B. 2002. Regime switches in interest rates. *Journal of Business and Economic Statistics* **20**: 163–182.
Arapis M, Gao J. 2006. Empirical comparisons in short-term interest rate models using nonparametric methods. *Journal of Financial Econometrics* **4**: 310–345.
Bansal V, Zhou H. 2002. Term structure of interest rates with regime shifts. *Journal of Finance* **57**(5): 1997–2043.
Black F, Karasinski P. 1991. Bond and option pricing when short rates are lognormal. *Financial Analysts Journal* **47**: 52–59.
Cížek P, Härdle W, Spokoiny V. 2009. Adaptive pointwise estimation in time-inhomogeneous conditional heteroscedasticity models. *Econometric Journal* **12**: 1–25.
Cox C, Ingersoll E, Ross A. 1985. A theory of the term structure of interest rates. *Econometrica* **53**: 385–407.
Das SR. 2002. The surprise element: jumps in interest rates. *Journal of Econometrics* **106**: 27–65.
Fan J, Jiang J, Zhang C, Zhou Z. 2003. Time-dependent diffusion models for term structure dynamics. *Statistica Sinica* **13**: 965–992.
Giacomini E, Härdle W, Spokoiny V. 2009. Inhomogeneous dependence modeling with time-varying copulae. *Journal of Business and Economic Statistics* **27**(2): 224–234.
Goyal A, Welch I. 2003. Predicting the equity premium with dividend ratios. *Management Science* **49**(5): 639–654.
Härdle W, Okhrin O, Okhrin Y. 2010. Time varying hierarchical Archimedean copulae. *SFB Discussion Paper*.
Hull J, White A. 1990. Pricing interest rate derivative securities. *Review of Financial Studies* **3**(4): 573–592.
Johannes M. 2004. The statistical and economic role of jumps in continuous-time interest rate models. *Journal of Finance* **51**(1): 227–260.

Jones M, Lamont O, Lumsdaine R. 1998. Macroeconomic news and bond market volatility. *Journal of Financial Economics* **47**: 315–337.
Lettau M, Ludvigson S. 2001. Consumption, aggregate wealth, and expected stock returns. *Journal of Finance* **56**(3): 815–849.
Mercurio D, Spokoiny V. 2004. Statistical inference for time-inhomogeneous volatility models. *Annals of Statistics* **32**(2): 577–602.
Paye B, Timmermann A. 2006. Instability of return prediction models. *Journal oF Empirical Finance* **13**: 274–315.
Spokoiny V. 2009. Multiscale local change point detection with applications to value-at-risk. *Annals of Statistics* **37**(3): 1405–1436.
Stanton R. 1997. A nonparametric model of term structure dynamics and the market price of interest rate risk. *Journal of Finance* **52**: 1973–2002.
Vasicek O. 1977. An equilibrium characterization of the term structure. *Journal of Financial Economics* **5**: 177–188.

Authors' biographies:

Mengmeng Guo is associate professor at Southwestern University of Finance and Economics, Chengdu, China. She got the doctor degree in economics from Humboldt-Universität zu Berlin in 2012. Her main research fields include financial econometrics, applied econometrics.

Wolfgang Karl Härdle is Professor in Institute of Statistics and Econometrics and director of Center for Applied Statistics and Economics at Humboldt-Universität zu Berlin. His main research fields include nonparametric estimation, financial statistics, risk management.

Authors' addresses:

Mengmeng Guo, Research Institute of Economics and Management, Southwestern University of Finance and Economics, LiuTai Blvd. 555, Wenjiang District, 611130 Chengdu, Sichuan, China.

Wolfgang Karl Härdle, Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Spandauer Straße 1, 10178 Berlin, Germany.

Härdle, Wolfgang Karl; Lee, David Kuo Chuen; Nasekin, Sergey; Ni, Xinwen; Petukhina, Alla

Working Paper

Tail event driven ASset allocation: Evidence from equity and mutual funds' markets

SFB 649 Discussion Paper, No. 2015-045

Provided in Cooperation with:
Collaborative Research Center 649: Economic Risk, Humboldt University Berlin

SFB 649 Discussion Paper 2015-045

Tail Event Driven ASset allocation: evidence from equity and mutual funds' markets

Wolfgang Karl Härdle*
David Lee Kuo Chuen*[2]
Sergey Nasekin*
Xinwen Ni*[3]
Alla Petukhina*

* Humboldt-Universität zu Berlin, Berlin, Germany
*[2] Singapore Management University, Singapore
*[3] Nanyang Technology University, Singapore

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

http://sfb649.wiwi.hu-berlin.de
ISSN 1860-5664

SFB 649, Humboldt-Universität zu Berlin
Spandauer Straße 1, D-10178 Berlin

Tail Event Driven ASset allocation: evidence from equity and mutual funds' markets*

Wolfgang Karl Härdle, †‡ David Lee Kuo Chuen ‡
Sergey Nasekin† Xinwen Ni § Alla Petukhina†

September 11, 2015

Abstract

Classical asset allocation methods have assumed that the distribution of asset returns is smooth, well behaved with stable statistical moments over time. The distribution is assumed to have constant moments with e.g., Gaussian distribution that can be conveniently parameterised by the first two moments. However, with market volatility increasing over time and after recent crises, asset allocators have cast doubts on the usefulness of such static methods that registered large drawdown of the portfolio. Others have suggested dynamic or synthetic strategies as alternatives, which have proven to be costly to implement. The authors propose and apply a method that focuses on the left tail of the distribution and does not require the knowledge of the entire distribution, and may be less costly to implement. The recently introduced TEDAS -Tail Event Driven ASset allocation approach determines the dependence between assets at tail measures. TEDAS uses adaptive Lasso based quantile regression in order to determine an active set of portfolio elements with negative non-zero coefficients. Based on these active risk factors, an adjustment for intertemporal dependency is made. The authors extend TEDAS methodology to three gestalts differing in allocation weights' determination: a Cornish-Fisher Value-at-Risk minimization,

*The financial support from the Deutsche Forschungsgemeinschaft via CRC 649 "Ökonomisches Risiko", Humboldt-Universität zu Berlin, IRTG 1792 and Sim Kee Boon Institute for Financial Economics, Singapore Management University.

†C.A.S.E.- Center for Applied Statistics & Economics, Humboldt-Universität zu Berlin, Berlin, Germany

‡Sim Kee Boon Institute for Financial Economics, Singapore Management University, Singapore

§School of Humanities and Social Sciences, Nanyang Technology University, Singapore

Markowitz diversification rule and naive equal weighting. TEDAS strategies significantly outperform other widely used allocation approaches on two asset markets: German equity and Global mutual funds.

Key words: adaptive lasso, portfolio optimisation, quantile regression, Value-at-Risk, tail events

JEL Classification: C00, C14, C50, C58

Portfolio allocation and selection go hand in hand with risk management, and are not only important concepts in quantitative finance and applied statistics, but are important determinants for long term portfolio returns for large funds. Over the past 60 years, several long-term asset allocation methods have been implemented. With each crisis occurring, more advanced methods were proposed after previous techniques failed to deliver. Notable approaches are the traditional 60/40-portfolio investment adopted by pension funds, Transparent Beta Base Model adopted by the Norwegian Sovereign Wealth Fund (NSWF), the Endowment Model popularised by University Endowments, the Core-Satellite Strategy introduced in the early 2000's , Risk Parity Model originated from the fund management firm Bridgewater, Factor Models/ Insurance and Option Overlay studied by academics and adopted by practitioners and insurers, Value and Focus Investing Model by Warren Buffett and other value investors, and ad-hoc Family Office/Real Estate Model that, however, has a notable bias of real estate in the portfolio although favoured by Asian tycoons, see Swensen [2009].

Absence of significant correlation among various asset classes is the essential motivation for traditional portfolio allocation. In reality, some strategies contradicted this principle, such as the traditional 60 equity/40 bond portfolio approach: the correlation between the bond market and the stock market was 0.98 in the last 15 years (Geczy [2014]). During the Global Financial Crisis the Endowment Model underperformed due to increased correlation across assets, Swensen [2009]. The Risk Parity strategy recommended a significant allocation to bonds amidst the implementation of quantitative easing and performed poorly because of interest rate volatility (Kazemi [2012] and Nathan [2013]). The Norwegian SWF model, strongly relied on the CAPM beta, which itself was unclear (Klarman [1991]). Performance of other models varied among investors, e.g., Factor Models that employed single or multiple factors, for instance, macroeconomic, risk or market factors, which were difficult to interpret; the Value Investing Model/ Warren Buffett that underperformed in recent years, and the Family Office Model that performed well during real asset bubble, Hamilton [2002].

A pillar in portfolio theory, mean-variance (MV) portfolio optimisation by Markowitz [1952] proposed to study semi-variance even though the optimisation was not straightforward given the low computation power at that time. As the computing capacities increased, later models incorporated optimisation involving higher and time varying moments. The mean-variance and subsequent refined models did not perform well during volatile periods and there were technical problems that were not addressed adequately. When the number of assets (p) is larger than the number of observations (n), there is a statistical problem, Bai et al. [2009] proved that the asset return estimate given by the Markowitz MV model was always larger than the theoretical return and the rate of the difference was related to p/n, the ratio of the dimension to the sample size. Jobson et al. [1979] and Jobson and Korkie [1980] showed that the Markowitz

mean-variance efficient portfolios were highly sensitive to p/n. They suggested to shrink the number of estimators or assets. From this point of view, the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani [1996]) may provide a feasible solution.

According to Lee et al. [2006], the inclusion of assets with returns that are skewed and leptokurtic in the portfolio will enhance returns. These assets provide the opportunity of downside protection especially during periods of high volatility. Härdle et al. [2014] introduce a new asset allocation strategy Tail Event Driven Asset Allocation (TEDAS), which exploits negative co-movement of alternative assets in the tail to hedge for downside risk. The subset of alternative or satellite assets performs the role of downside protection. Successful protection of the portfolio by limiting the downside risk during volatile periods allows the portfolio returns to recover sooner. It is not surprising that TEDAS, with smaller drawdowns, outperforms more traditional methods that suffer larger drawdowns during extreme events. Given that a subset of alternative or satellite assets is chosen from a larger universe of assets, TEDAS can also be viewed as an application of the Core-Satellite model. For big data, where the number of possible alternatives is larger than the number of observations, the Adaptive LASSO quantile regression (ALQR) is introduced to address this issue and is used to simultaneously pursue variable selection and measure relations between variables at tail quantiles. In order to deal with changing volatility and correlation structure problem and to better match the higher moments of the portfolio distribution, one applies Cornish-Fisher VaR (Value-at Risk) model with Dynamic Conditional Correlation (DCC) and higher moments, such as skewness and kurtosis, can be used to obtain optimal asset weights among chosen products. Here, we extend TEDAS by introducing three modifications, which we call *"TEDAS gestalts"*: *TEDAS basic*, *TEDAS naive*, which places equal weights on every satellite asset, and *TEDAS hybrid*, which uses the most common Markowitz variance-covariance rule to select the weights.

In addition, we apply the TEDAS methods to global mutual fund and German stock market data. First, TEDAS yields robust and consistent results, with various assets,time periods, parameter frequencies, and in big and small data. Secondly, we show results that compare different TEDAS methods. Finally, the results are presented with transaction costs incorporated into our portfolio rebalancing.

The rest of the paper is organised as follows. In section 2, we introduce the framework of TEDAS. In section 3 and 4, we apply the methods to different markets and compare the performance between different models. Section 5 will present the conclusion and discussion. All codes and datasets are available as Quantlets on Quantnet (Borke and Härdle [2015]) .

TEDAS - Tail Event Driven Portfolio Allocation

The basic elements of TEDAS are presented in Härdle et al. [2014]. The proposed tool set has important implications for portfolio risk management and asset allocation decisions. Along with the basic setup we propose two modifications: *TEDAS naive* and *TEDAS hybrid*.

The TEDAS strategy is based on a simple idea widely used in core-satellite approach. The core asset is chosen to be e.g., the DAX index or S&P 500. The strategy is to select satellite assets to complement the core portfolio. The core portfolio is chosen by the fund manager and the satellite assets are chosen by TEDAS to limit the downside of the core portfolio during extreme events.

The second step is a selection of satellite portfolio constituents. In TEDAS the Adaptive Lasso Quantile Regression (ALQR) is applied to pick assets for a new portfolio Zheng et al. [2013]. This technique allows to simultaneously solve two challenges for portfolio managers. It shrinks the high dimensional universe of satellite assets to potential candidates for portfolio's constituents. ALQR also provides the information concerning the dependence between core portfolio and satellites at different quantiles (for various tail events). TEDAS employs 5%, 10%, 25%, 35% and 50% tail events. Assets with negative ALQR coefficients, i.e. assets adversely moving with the core for chosen level of a tail event, are constituents of a new rebalanced portfolio. For the case with only positive ALQR coefficients received, it is supposed, the value of the portfolio does not change in comparison with the previous period (a portfolio manager keeps a so-called "stay-in-cash" position). Technical details for the ALQR are provided in appendix.

The third step is a determination of portfolio weights for assets selected on the second step. TEDAS proposes three alternative ways to solve this task, we refer to them as TEDAS gestalts, which is originally a german word to indicate an organised whole that is perceived as more than the sum of its parts and literally can be translated as "form, shape" (Oxford Dictionary of English [2010]). Depending on a volatility-modeling method and the portfolio weights' optimisation rule three TEDAS gestalts can be applied. The *TEDAS basic* gestalt employs the dynamic conditional correlation model (DCC) is used (Engle [2002], Franke et al. [2015]) to account for time-varying covariance structure and correlation shifts in returns' covariance. The weights of satellites are defined based on the Cornish-Fischer Value-at-Risk (VaR) minimization rule, Favre and Galeano [2002] (Technical details are included in appendix).

The *TEDAS naive* gestalt assigns to every satellite asset the same portfolio weight.

The *TEDAS hybrid* after LASSO selection employs the simplest approach to estimate the covariance structure of assets' returns, the historical covariance matrix; portfolio weights are calculated according to classical mean-variance optimisation procedure (Markowitz diversification rule), Markowitz [1952].

The choice of satellite assets and data description

Small and mid-cap stocks

Banz [1981] found smaller firms (small caps) have had higher risk-adjusted returns, on average, than larger firms. Reinganum [1981] observed portfolios based on firm size or earnings/price ratios experienced average returns systematically different from those predicted by the CAPM. Since these pioneer papers the effect of relation between size and expected return attracted a significant attention of academics and practitioners. Research in this area is often referred to as "small cap premium", "size premium", or "size anomaly" literature. The size premium effect was preserved even after controlling for market factor and the value effect Fama and French [1993] , the momentum effect Jegadeesh and Titman [1993] and Carhart [1997], liquidity effects Pastor and Stambaugh [2003] and Ibbotson et al. [2013], industry factors as well as high leverage, low liquidity, Menchero et al. [2008]. Moreover, studies of stock returns across many separate countries and regions also confirmed the size phenomenon, Rizova [2006] summarised the academic evidence on the international existence of the size effect.

What is the source of the size premium? The traditional theory claimed that firm size was a proxy for systematic risk, small cap stocks were riskier than large cap stocks, and, therefore, market forced exert downward pressure on the prices of small cap stocks to provide investors with higher returns, Fama and French [1993]. Subsequent researchers explored the underlying sources of such risk, but the results were controversial. For example, Amihud and Mendelson [1986] proposed to link the size effect with liquidity risk, measured as bid-ask spread, and their results demonstrated the size premium effect was mostly a liquidity driven. Amihud [2002] found that smaller firms' returns were more sensitive to market illiquidity and that small cap stocks had more liquidity risk than large caps stocks, Liu [2006] also argued that small caps required higher returns for accepting liquidity risk. Zhang [2006] proposed another source of risk, namely 'information uncertainty', which linked small caps to law quality of the information disclosure and information about a firms' volatile fundamentals. Chan and Chen [1991] and Dichev [1998] suggested that size served as a proxy for financial distress, Vassalou and Xing [2004] stated the size effect was a default effect and together with value (the book-to-market) effect existed only in market segments with high default risk. Overall, this group of literature explored reasons, why

higher risks were linked to small and mid caps. Lakonishok et al. [1994] proposed an alternative explanation and proved that small caps were mispriced by investors due to behavioural biases and not because these types of assets were fundamentally riskier.

After first discovering and documentation of size premium in Banz [1981], Fama and French [1993] also observed a premium of 0.27% per month in the US over the period 1963 to 1991. However, more recent studies documented the size anomaly disappeared (see, e.g., Amihud [2002] , Dichev [1998]) since 1980 in the US. Furthermore, Fama and French [2012] observed no size effect across 23 countries from November 1990 to September 2010. At the same time Hou and Dijk [2010] argued that U.S. stocks of smaller firms had not had higher returns since the early 1980s because of firm profitability "shocks": smaller firms had negative earnings surprises and larger firms had positive earnings surprises during this time. Based on this argument, they claimed that the size effect still existed even it was not so obvious (see also Crain [2011]). Three studies on the size anomaly in Germany provided inconsistent results. Namely, Stehle [1992] found some evidence of a size effect in Germany, especially in January, whereas Schlag and Wohlschieß [1992] obtained very low t-statistics for size as an explanatory variable for mean returns. Sauer [1994] too did not detect a size related anomaly for stock returns in Germany. For an extensive literature review concerning a size effect we refer to e.g., Crain [2011]. It can be summarised, that the size effect has been challenged along many fronts. Over the last decade, however, global small caps and mid caps have been relatively strong again and outperformed large caps (Figure 1). The existence of size effect as well as the benefits of diversification (see, e.g., Bender et al. [2012]) strongly motivates inclusion of small and mid cap stocks into allocation strategies. In our research we utilise small and mid cap stocks as satellite assets for the TEDAS strategy.

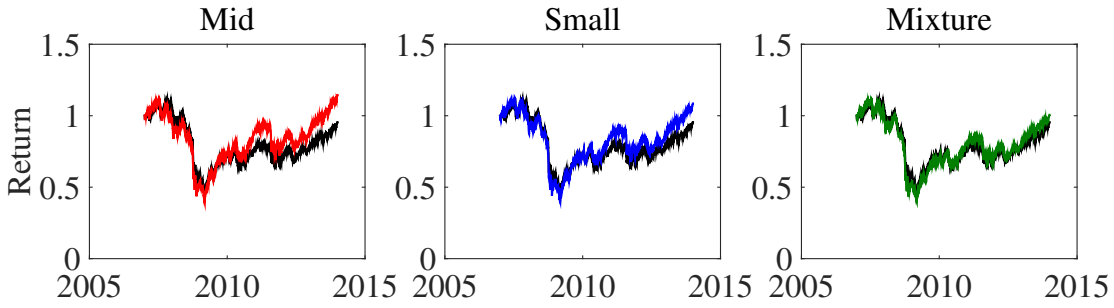


Figure 1: Daily cumulative returns of MSCI World Large Cap index (black) from 1 Jan 2007 to 31 Dec 2014 against MSCI World Mid Cap index (red), MSCI World Small Cap index (blue) and MSCI World Small and Mid Cap / mixture index (green)

The empirical analysis of TEDAS application to equity market focuses on the German stock market. As the core-asset DAX index is employed and 125 constituents of indices SDAX, MDAX and TecDAX construct the universe of hedging assets – small and mid-cap stocks. The collected data cover the time period from 21 Dec 2012 to

27 Nov 2014 (Source: *Datastream*). The performance of TEDAS strategy for German equity market was analysed on 41 sixty-weeks moving windows.

Mutual funds

The role of Mutual Funds in world economy has increased in the 20 years or so due to their fast growth (from 52 746 in 1999 they of Mutual Finds has reached 76 200 by 2013) (Figure 2). The US economy is the market that accounts for about half of the global mutual fund market of $30 trillion which underlines its importance in the US economy. In addition mutual fund investment companies account for 88 percent of investment companies in total. The popularity of mutual funds is due to their perceived safety compared to alternatives, notably stocks. This perception has resulted in a situation where almost half specifically, 46.3 percent of US households have participated in such funds. All this underlines the sheer size and the importance of the US mutual find market which, therefore, should provide us with an important test case for the evaluation of the performance of TEDAS strategy and would show whether TEDAS can handle cases of big data.

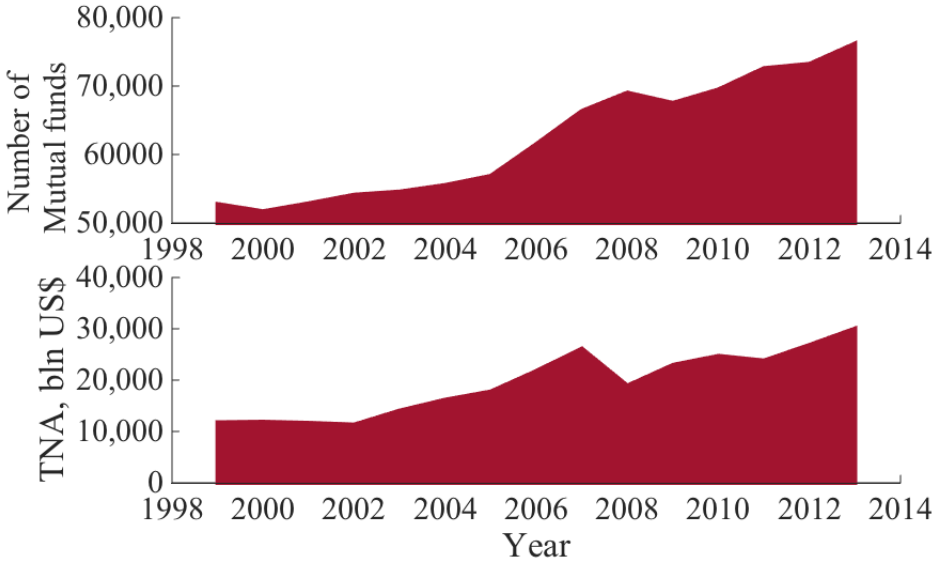


Figure 2: Number (upper graph) and Total net Assets (lower graph) of Worldwide Mutual Funds from 1999 to 2013

The potential of diversification, a major determinant for asset allocation, is a major and very attractive characteristic of mutual funds. In the 2013 US market, 38 percent of all industry assets were held by domestic equity funds and an additional 14 percent by world equity funds. Moreover, it is pointed out that the percentage of mutual fund assets that were in the form of bond funds is at 22 percent, whereas money market

funds covered 18 percent and 8 percent, the remaining, was accounted by hybrid funds. Finally, it has been observed that in the US there has been a tendency towards equity mutual funds regarding portfolio diversification, which means increased investment rates in foreign (non-US) markets.

The data for this study come from Datastream and represent the period from January 1998 to December 2013, i.e. a period of 192 months. The classification of the data was performed on the basis of three locations in which they originated: United States, Singapore and the World. At first hand, cross-sectional data from 2616 funds were retrieved, but only that from funds that had had a life of at least 10 years. Not surprisingly, the US market had the largest representation in the data set with 2347 cases of mutual funds, whereas Singapore had only 13 and the other markets 256. To simplify the processing of the data some further reduction of the data set was applied: inactive cases – the ones which showed no price change for 3 months – were excluded resulting to a total of 583 remaining cases which provided the dataset for our calculations. S&P 500 provided the core asset, whereas Bloomberg was the source of the data from the same time space.

Empirical results

Results for German equity market

The comparison of the three TEDAS gestalts with the core DAX30 index is given on Figure 3. As is seen, all three TEDAS strategies demonstrate almost equal results in terms of cumulative return. At the end of the analysed timespan these strategies yield 41-42 % of cumulative return taking into account 1% of transaction fees (The cumulative returns reach even 60 % - 70 % without the transaction costs). The asset allocation decision is twofold: one has to define which assets to buy and which proportions to use to construct the portfolio (solution of weights' optimisation problem). One observes though the main driving factor of the overperformance for TEDAS strategy comes from the portfolio assets' selection and not really from weights' optimisation. A conducted sign test confirms the absence of difference in medians of returns for the three TEDAS gestalts (on 5% significance level).

TEDAS needs to be benchmarked with three alternative widely used strategies: Risk-Parity portfolio (Equal risk contribution portfolio), OGARCH mean-variance strategy, 60/40 portfolio. The mean-variance (MV) portfolio selection has been widely used by the financial community and is the common benchmark for every newly introduced asset allocation strategy. The traditional Markowitz portfolio optimisation approach as has been shown in previous literature has some drawbacks especially for the case when $p>n$. The portfolio formed by using the classical mean-variance approach

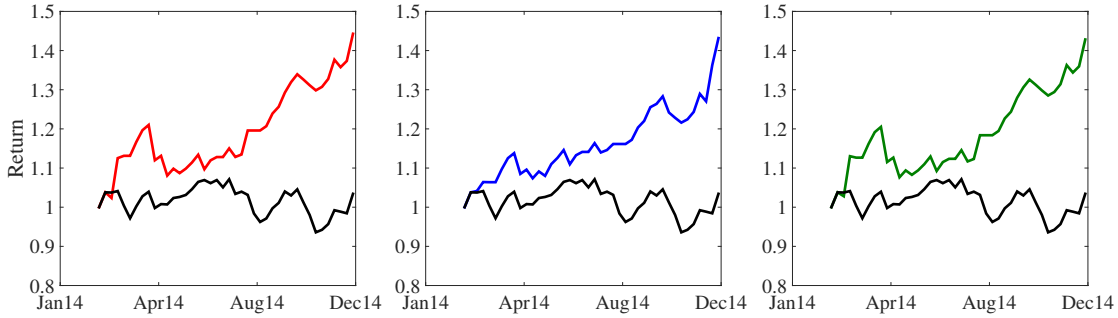


Figure 3: Weekly cumulative returns of DAX30 index (black) from 21 Dec 2012 to 27 Nov 2014 against TEDAS basic (red), TEDAS naive (blue) and TEDAS hybrid (green) strategies applied to German stocks

TEDAS_gestalts

always results in extreme portfolio weights Jorion [1985], that fluctuate substantially over time and perform poorly in the sample estimation (for example, Frankfurter et al. [1971], Simaan [1997], Kan and Zhou [2007]) as well as in the out-of-sample forecasting.

Different studies provide different observations and suggestions to investigate the reasons, why the MV optimisation estimate is so far away from its theoretic counterpart. So far, all believe that the reason behind this outcome is that the "optimal" return is formed by a combination of returns from an extremely large number of assets (see McNamara [1998]). Use of Markowitz optimisation procedure efficiently depends on whether the expected return and the covariance matrix can be estimated accurately. Many studies have improved the estimate of the classical Markowitz MV approach by using different approaches. For our comparative study, the conditional variance-covariance matrix was estimated with Orthogonal GARCH factors. In our study we use dynamic Markowitz risk-return optimisation with portfolio covariance matrix modelled by the basic orthogonal GARCH method. The Orthogonal GARCH model was first proposed in Alexander [2001], and is based on Principal Components Analysis (PCA).

60/40 portfolio allocation strategy implies the investing of 60% of the portfolio value in stocks (often via a broad index such as S& P500) and 40% in government or other high-quality bonds, with regular rebalancing to keep proportions steady. German market's 60/40 portfolio is constructed with DAX and RDAX indices.

Risk-parity portfolio-strategy is based on allocation by risk, not by capital. In this case, the portfolio manager defines a set of risk budgets and then computes the weights of the portfolio such that the risk contributions match the risk budgets (for details see Maillard et al. [2010]).

The comparison of cumulative returns achieved with *TEDAS hybrid* and alternative

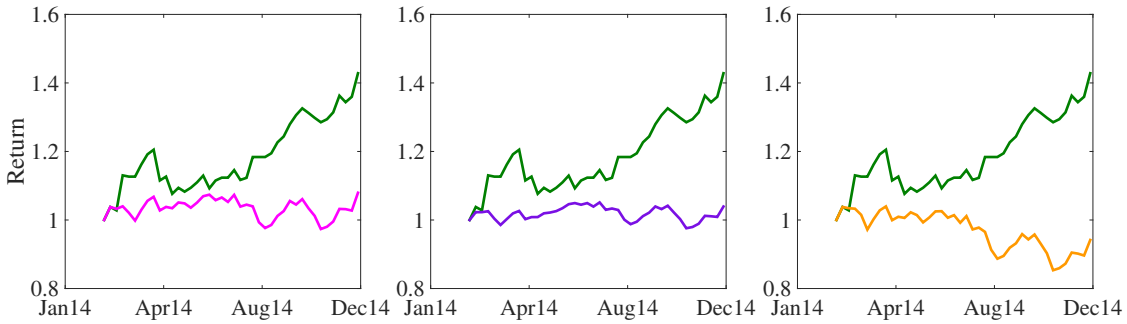


Figure 4: Weekly cumulative returns of TEDAS Hybrid (green) from 21 Dec 2012 to 27 Nov 2014 against MV OGARCH (magenta), 60/40-portfolio (purple) and Risk Parity (orange) strategies applied to German stocks

TEDAS_gestalts

strategies, demonstrated in Figure 4 , shows that TEDAS performs significantly better than other considered approaches.

Strategy	Cumulative return	Sharpe ratio	Maximum drawdown
TEDAS basic	143%	0.3184	0.1069
TEDAS naive	144%	0.3792	0.0564
TEDAS hybrid	143%	0.3079	0.1068
MV OGARCH	108%	0.0687	0.0934
Risk-Parity	95%	-0.0693	0.1792
60/40 portfolio	121%	0.0306	0.0718
DAX30	103%	0.0210	0.1264

Table 1: Strategies' performance overview: German stocks' sample

TEDAS_perform

The rebalancing of portfolio to hedge the core asset occurred 21 times out of 41 moving-window estimation periods. Table 1 summarises the performance of portfolio strategies in terms of cumulative returns as well as in terms of risk. We used two traditional measures to evaluate portfolios' risk-adjusted returns: Sharpe ratios and maximum drawdown. As it can be seen from the results, the most attractive strategy is *TEDAS basic*, which gives the highest excess return for the extra volatility. At the same time, *TEDAS naive* demonstrates the lowest financial risk, measured with maximum drawdown. In general we can conclude that TEDAS strategies show better risk-adjusted returns than all other analysed benchmarks and have comparatively the same level of risk.

Figure 5 shows the frequency of the number of selected variables for different quantiles. As can be noticed, the number of selected satellites in most of cases is

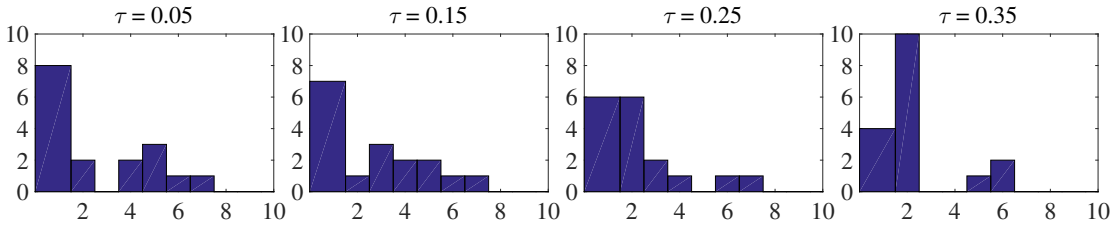


Figure 5: Frequency of the number of selected stocks for 4 different quantiles (German stocks' sample)

less than five, which is also indicative of this strategy and the simplicity of portfolio rebalancing. Furthermore, we analyze how frequently certain stocks were selected as satellites (i.e. how often they have significant ALQR non-positive coefficients) the results of which are given in figure 6. More frequently small stocks (first 50 stocks on the graph) and stocks of high-tech companies (last 30 stocks) hedge the core. This conclusion is also confirmed by table 2, which lists the most frequently used German stocks for 5 % quantile and most part of them operate in the high technology innovative industries.

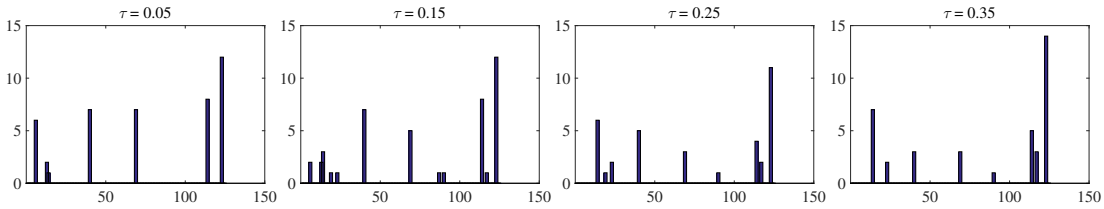


Figure 6: Frequency of selected stocks for 4 different quantiles (German stocks' sample)

Top 5 influential stocks	Frequency	Index	Industry
Sartorius Aktiengesellschaft	12	TecDAX	Provision of laboratory and process technologies and equipment
XING AG	8	TecDAX	Online business communication services
Surteco SE	7	SDAX	Household Goods & Home Construction
Kabel Deutschland Holding AG	7	MDAX	Cable-based telecommunication services
Biotest AG	6	SDAX	Producing biological medications

Table 2: The selected German Stocks for 5% quantile

All TEDAS gestalts applied to the universe of German stocks outperform both traditional benchmark strategies such as Markowitz rule or 60/40 and more sophisticated ones such as the risk-parity model. Our analysis leads us to believe

that using the ALQR technique delivers good results in reducing the dimensionality of the asset universe for more effective portfolio allocation.

Results for global mutual funds

Since the number of satellites after filtering (p=583) is very large, the moving window for Mutual funds' sample is adjusted to 120. We assume in December 2007 one starts to allocate 1 unit of money using each strategy and calculated the 73 monthly cumulative returns until Dec 2013.

Similar to the previous analysis, the outcomes of the three TEDAS strategies are compared. From 2007 to the end of 2013, the *TEDAS Naive* yields the highest return, 454%. *TEDAS Hybrid* and *TEDAS Basic* setups show similar returns of 433% and 421% respectively (Figure7).

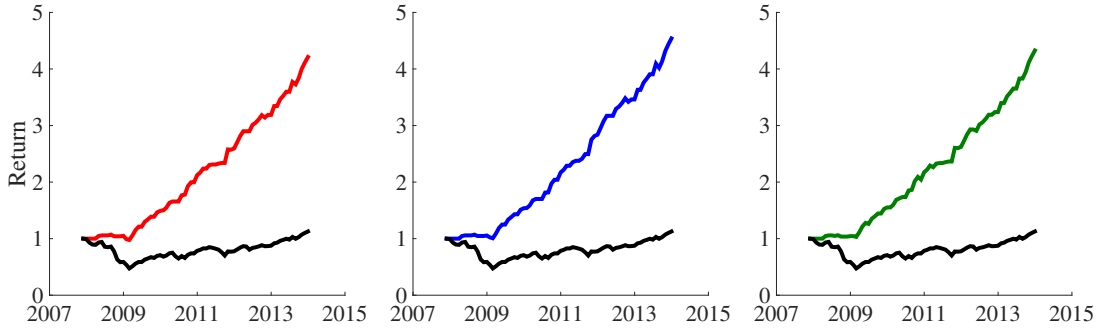


Figure 7: Monthly cumulative returns of S&P 500 from Dec 2007 to Dec 2013 against TEDAS basic (red), TEDAS naive (blue) and TEDAS hybrid (green) strategies applied to Mutual funds

TEDAS_gestalts

In order to check whether TEDAS is significantly better than popular methods that have been applied in the past years, we employed the same four benchmarks in the case of German stocks. We constructed a 60/40 portfolio using NASDAQ composite and the Barclays US treasury index. For the base case, we buy and hold the core asset, S&P 500, during the whole period. By comparing the TEDAS hybrid and the benchmarks, we can tell that TEDAS is out-performed. 60/40 and Risk-Parity portfolios have high correlation with the S&P 500 and these three gave similar returns of around 125% (Figure 8). By Sign Test between *TEDAS Hybrid* and other four benchmarks, we could get the p-values, which are all smaller than 5% and therefore, we could conclude that the return of our strategy is statistically and significantly different from others.

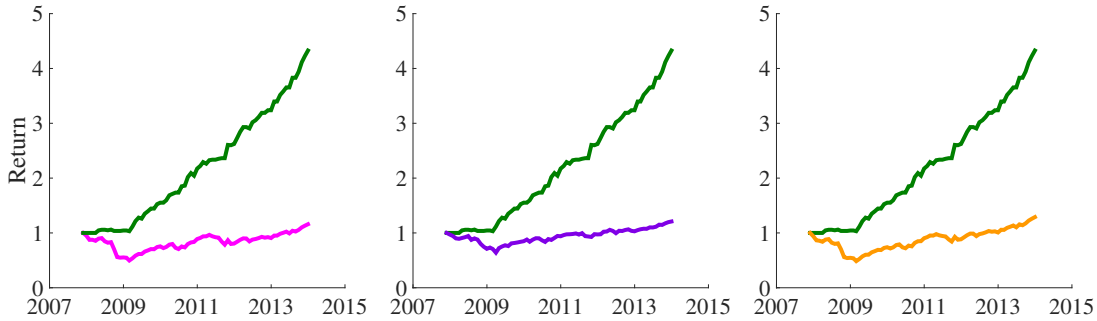


Figure 8: Monthly cumulative returns of TEDAS Hybrid (green) from Dec 2007 to Dec 2013 against MV OGARCH (magenta), 60/40-portfolio (purple) and Risk Parity (orange) strategies applied to Mutual funds

TEDAS_gestalts

Strategy	Cumulative return	Sharpe ratio	Maximum drawdown
TEDAS basic	421%	0.6393	0.0855
TEDAS naive	454%	0.6974	0.0583
TEDAS hybrid	433%	0.6740	0.0276
MV OGARCH	116%	0.0214	0.4772
Risk-Parity	129%	0.0487	0.4899
60/40 portfolio	121%	0.0252	0.3473
S&P500	113%	0.0132	0.5037

Table 3: Strategies' performance overview: Mutual funds' sample

TEDAS_perform

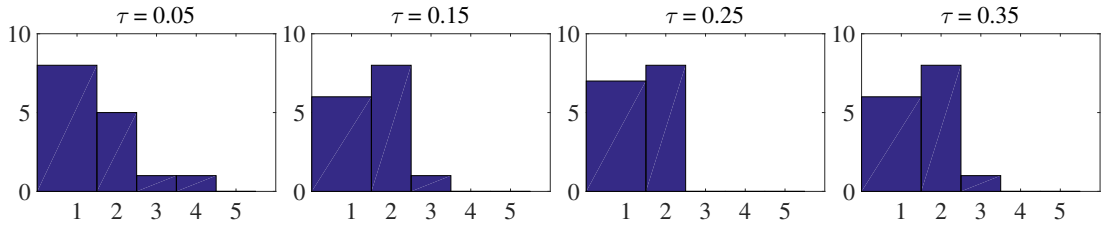


Figure 9: Frequency of the number of selected variables for 4 different quantiles (Mutual funds' sample)

Figure 9 shows the different frequencies of the number of selected variables from 4 quantiles (0.05, 0.15, 0.25 and 0.35). Unexpectedly, the number of selected satellites is all less than four in all cases, which is similar with the German Stock data. Compared with the number of selection pool (583 Mutual Funds), 4 is really small.

One explanation might be that even though Mutual Funds consist of combinations of many products (different kinds of bonds, domestic and international equities), and have many different investment ways, there is a huge part of the investment pool has been allocated into the U.S. stock markets or into related products. As a result of globalization, the U.S. market strongly affects other markets.

Influential Mutual Funds	Frequency	Market
Blackrock Eurofund Class I	12	U.S.
Pimco Funds Long Term United States Government Institutional Shares	8	U.S.
Prudential International Value Fund Class Z	4	U.S.
Artisan International Fund Investor Shares	3	U.S.
American Century 2OTH Century International Growth Investor Class	1	U.S.
First Eagle Overseas Fund Class A	1	U.S.

Table 4: The selected Mutual Funds for for 5% quantile

TEDAS does not select many different Mutual Funds, only 6 Mutual funds hedged the core in extreme events throughout the analysed period. From Table 4 we can see that all selected Mutual Funds exchanged in U.S. market, but most of them are related to the products outside the U.S. markets.

Conclusion and Discussion

Asset allocators have difficulties in constructing a portfolio that can sufficiently protect the downside with acceptable level of drawdown. Each crisis, previously adopted methods failed to limit the downside as suggested by empirical stress testing based on historical data. Here, we have proposed a method that focuses on the co-movement of the core and the universe of satellite assets during extreme events. The degree of extremeness is defined as the percentage of historical observations in the tail, also known as quantiles. By selecting and reducing the universe of satellite assets to a manageable subset and at the same time having the properties of negative or zero correlation with the core during extreme event is the innovation of this paper.

The main contribution of this paper is to demonstrate the practical significance of the TEDAS tool set for a wide range of both institutional and private investors in various settings. We conducted an empirical study on the performance of TEDAS strategy applied to a broad spectrum of core and satellite configurations. The testing of TEDAS strategy for Global Mutual funds and German equity data leads to conclusion TEDAS is meaningful for geographically different markets (global and Germany), using weekly and monthly returns as well as for different levels of dimensionality of the universe of potential portfolio constituents. This paper demonstrated the power of the TEDAS strategy for different asset markets, such as equity, Mutual funds and Hedge funds. Furthermore, compared with four conventional benchmark allocation

approaches, TEDAS cumulative returns are significantly higher. Investigation of TEDAS outperformance in terms of risk measures, such as Sharpe ratio and maximal drawdown, also demonstrates better results than other benchmark strategies. Finally, when we relaxed the assumption of zero transaction fees TEDAS still demonstrates superior performance, significantly different from other traditional approaches.

There are many ways in which we envision the research reported here can be extended. The results of three modifications of TEDAS adopted in this study are robust. Theoretically speaking, *TEDAS basic*, which takes the third and fourth moments into account, should perform better than the other two. However, we do not observe it in our empirical study. There are some possible explanations and directions for further analysis. One is to solve the utility maximization problem with higher moments or to include time-varying modelling of higher portfolio moments as in Ghalanos et al. [2015].

Analysing the superior returns of TEDAS strategies, it is necessary to keep in mind all results were received based on realized returns and not on expected returns. Therefore, the possible direction for a further development of TEDAS strategy might be an incorporation of returns' forecasting and examining of out-of-sample performance. In conclusion, the results suggest that these TEDAS methods, while still relying on historical methods, are producing promising results. The caveat remains that history may not necessarily repeats itself and further studies are needed.

Appendix

Adaptive LASSO Quantile regression (ALQR)

Adaptive Lasso Procedure

Introduced in Bassett and Koenker [1978] quantile regression (QR) estimates conditional quantile functions–models in which quantiles of the conditional distribution of the response variable are expressed as functions of observed covariates (see Koenker and Hallock [2001]).

L_1 - penalty is considered to nullify "excessive" coefficients (Belloni and Chernozhukov [2011]). Simple lasso-penalized QR optimisation problem is:

$$\hat{\beta}_{\tau,\lambda} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(Y_i - X_i^\top \beta) + \lambda \|\beta\|_1 \tag{1}$$

The adaptive Lasso, Zou [2006], yields a sparser solution and is less biased. L_1 -

penalty is replaced by a re-weighted version:

$$\hat{\beta}^{\text{adapt}}_{\tau,\lambda_n} = \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(Y_i - X_i^\top \beta) + \lambda_n \|\hat{\omega}^\top \beta\|_1 \tag{2}$$

here $\tau \in (0,1)$ is a quantile level, $\rho_\tau(u) = u\{\tau - \mathbf{I}(u<0)\}$ piecewise loss function, λ_n regularization parameter. Weights $\hat{\omega} = 1/|\hat{\beta}^{\text{init}}|$, $\hat{\beta}^{\text{init}}$ is obtained from (1). In TEDAS setup $Y \in \mathbb{R}^n$ represents core log-returns (DAX or S&P500 indices)and $X \in \mathbb{R}^{n\times p}$ – satellites' log-returns (German stocks or Mutual funds), $p > n$.

Algorithm for Adaptive Lasso Penalized QR

The optimisation for the adaptive Lasso can be re-formulated as a Lasso problem:

- the covariates are rescaled: $\tilde{X} = (X_1 \circ \hat{\beta}^{\text{init}}_1, \ldots, X_p \circ \hat{\beta}^{\text{init}}_p)$;
- the lasso problem (1) is solved:

$$\hat{\tilde{\beta}}_{\tau,\hat{\lambda}} = \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(Y_i - \tilde{X}_i^\top \beta) + \lambda \|\beta\|_1 \tag{3}$$

- the coefficients are re-weighted as $\hat{\beta}^{\text{adapt}} = \hat{\tilde{\beta}}_{\tau,\hat{\lambda}} \circ \hat{\beta}^{\text{init}}$

Cornish-Fisher VaR optimisation

A modification of VaR via the Cornish-Fisher (CF) expansion improves its precision adjusting estimated quantiles for non-normality. To obtain asset allocation weights the following VaR-minimization problem is solved (for details see Favre and Galeano [2002], Härdle et al. [2014]):

$$\begin{aligned} &\underset{w\in\mathbb{R}^d}{\text{minimize}} \quad W_t\{-q_\alpha(w_t)\cdot\sigma_p(w_t)\} \\ &\text{subject to} \quad w_t^\top \mu = \mu_p,\ \ w_t^\top 1 = 1,\ \ w_{t,i} \geq 0 \end{aligned} \tag{4}$$

here $W_t \overset{\text{def}}{=} W_0 \cdot \prod_{j=1}^{t-1} w_{t-j}^\top (1 + r_{t-j})$, $\tilde{w}$, W_0 initial wealth, $\sigma_p^2(w) \overset{\text{def}}{=} w_t^\top \Sigma_t w_t$,

$$q_\alpha(w_t) \overset{\text{def}}{=} z_\alpha + (z_\alpha^2 - 1)\frac{S_p(w_t)}{6} + (z_\alpha^3 - 3z_\alpha)\frac{K_p(w_t)}{24} - (2z_\alpha^3 - 5z_\alpha)\frac{S_p(w_t)^2}{36}, \tag{5}$$

here $S_p(w_t)$ skewness, $K_p(w_t)$ excess kurtosis, z_α is $\mathrm{N}(0,1)$ α-quantile. If $S_p(w_t)$, $K_p(w_t)$ are zero, then the problem reduces to the Markowitz case.

Mean-variance optimisation procedure (Markowitz diversification rule)

Mean-variance optimisation procedure is based on four inputs: the weights of total funds invested in each security w_i, $i = 1, \ldots, d$, the expected returns μ approximated as averages $\overline{r}$, volatilities (standard deviations) σ_i associated with each security and covariances σ_{ij}, $j = 1, \ldots, d; i \neq j$ between returns. Portfolio weights w_i are obtained from the quadratic optimisation problem, see Brandimarte [2006], p. 74

$$\begin{aligned} \underset{w \in \mathbb{R}^d}{\text{minimize}} \quad & \sigma_p^2(w_t) \overset{\text{def}}{=} w_t^\top \Sigma w_t \\ \text{subject to} \quad & w_t^\top \mu = r_T, \\ & \sum_{i=1}^{d} w_{i,t} = 1, \\ & w_{i,t} \geq 0 \end{aligned} \tag{6}$$

where $\Sigma \in \mathbb{R}^{d \times d}$ is the covariance matrix for d portfolio asset returns, r_T is the "target" return for the portfolio assigned by the investor. Markowitz optimisation procedure gives the same result as CF-VaR optimisation in case of skewness and excess kurtosis are zero (in excess of 3, which corresponds to a Gaussian distribution).

Reference

Alexander, C. "A Primer on the Orthogonal GARCH Model." Unpublished manuscript, ISMA Centre, University of Reading, UK, 2001.

Amihud, Y. and H. Mendelson. "Asset pricing and the bid-ask spread." Journal of Financial Economics, Vol. 17, No. 2 (1986), pp. 223–249.

Amihud, Y. "Illiquidity and stock returns: cross-section and time-series effects." Journal of Financial Markets, 5(2002), pp. 31–56.

Bai, Z. , H. Liu, and W.K. Wong. "Enhancement of the applicability of Markowitz's portfolio optimisation by utilizing random matrix theory." Mathematical Finance, Vol. 19, No. 4 (2009), pp. 639-667.

Banz, R. W. "The relationship between return and market value of common stocks." Journal of Financial Economics, Vol. 9, No. 1 (1981), pp. 3–18.

Bassett, G. and R. Koenker. "Regression Quantiles." Econometrica, 46 (1978), pp. 33–50.

Belloni, A. and V. Chernozhukov. "L_1-penalized Quantile Regression in High-Dimensional Sparse Models." Annals of Statistics, Vol. 39, No. 1 (2011), pp. 82–130.

Bender, J., R. Briand, G. Fachinotti, and S. Ramachandran. "Small Caps? No Small Oversight: Institutional Investors and Global Small Cap Equities." MSCI Research Insight, March 2012, $www.msci.com/www/research-paper/small-caps-no-small-oversight/014391548$

Borke, L. and W. K. Härdle. "Q3-D3-LSA." SFB 649 Discussion paper (forthcoming), Humboldt Universität zu Berlin, 2015.

Brandimarte, P. Numerical Methods in Finance and Economics. A Matlab-Based Introduction, Wiley, 2006.

Carhart, M. "On Persistence in Mutual Fund Performance." The Journal of Finance, Vol. 52, No. 1 (1997), pp. 57–82.

Chan, L. K. C., and N.F. Chen. "Structural and Return Characteristics of Small and Large Firms." Journal of Finance, 46 (1991), pp. 1467–1484.

Chunhachinda P., K. Dandapani, S. Hamid, and A.J.Prakash. "Portfolio selection and skewness: Evidence from International Stock Markets." Journal of Banking and Finance, Vol. 21, No. 2 (1997), pp. 143-167.

Cornish E.A. and R.A. Fisher. "The percentile points of distributions having known cumulants." Technometrics Vol. 2, No. 2 (1960), pp. 209–225.

Crain, M. "A literature review of the size effect." Working paper, October 2011, $www.ssrn.com/abstract\_id = 1710076$

Dichev, I. "Is the Risk of Bankruptcy a Systematic Risk?" The Journal of Finance, 53 (1998), pp. 1131–1148.

Efron, B. and Tibshirani, R.J. An Introduction to the Bootstrap, Chapman & Hall/CRC, 1993.

Engle, R. "Dynamic Conditional Correlation: a Simple Class of Multivariate GARCH Models." Journal of Business and Economic Statistics, Vol. 20, No. 3(2002), pp. 339–350.

Favre, L. and J.-A. Galeano. "Mean-Modified Value-at-Risk optimisation with Hedge Funds." The Journal of Alternative Investments, Vol. 5, No. 2(2002), pp. 21–25

Fama, E., and K. French. "Common risk factors in the returns on stocks and bonds." Journal of Financial Economics, Vol. 33, No. 1 (1993), pp. 3–56.

Fama, E., and K. French. "Size, value, and momentum in international stock returns." Journal of Financial Economics, 105 (2012), pp.457–472.

Franke, J. and Härdle, W.K., Hafner, C.M. Statistics of Financial Markets. An Introduction, 4th Edition. Springer, 2015.

Frankfurter G.M., H.E. Phillips, and J.P. Seagle. "Portfolio Selection: The Effects of Uncertain Means, Variances and Covariances." Journal of Financial and Quantitative Analysis, 6 (1971), pp. 1251–1262.

Ghalanos A., E. Rossi, and G. Urga. "Independent Factor Autoregressive Conditional Density Model." Econometric Reviews,Vol. 34, No. 5 (2015), pp. 594–616.

Geczy C. "The new diversification: open your eyes to alternatives." Blackrock, 2014.

Härdle, W.K., S. Nasekin , D.K.C. Lee, and K.F. Phoon. "TEDAS - Tail Event Driven Asset Allocation." SFB 649 Discussion Paper 2014-032, Humboldt University zu Berlin, 2014.

Hou, K., and M.A.v. Dijk. "Profitability shocks and the size effect in the cross-section of expected stock returns." Working paper, January 2010, $www.ssrn.com/abstract\_id = 1536804$

Huang X. "Mean-semivariance models for fuzzy portfolio selection." Journal of Computational and Applied Mathematics, 217 (2008), pp. 1–8.

Hamilton S. The Multi-Family Office Mania, Trusts& Estates, 2002.

Ibbotson, R.G., Z. Chen, D. Y.-J. Kim, and W.Y. Hu. "Liquidity as an Investment Style." Financial Analysts Journal, Vol. 69, No. 3 (2013), pp. 30–44.

Jegadeesh, N. and S. Titman. "Returns to Buying Winners and Selling Losers: Implications for Market Efficiency." Journal of Finance, Vol. 48, No. 1 (1993), pp.65–91.

Jobson, J.D., B. Korkie, and V.Ratti. " Improved Estimation for Markowitz Portfolios using James-Stein Type Estimators, Proceedings of the American Statistical Association. " Business and Economics Statistics, 41 (1979), pp. 279–284.

Jobson, J.D. and B. Korkie. " Estimation for Markowitz efficient portfolios." Journal of the American Statistical Association, 75 (1980), pp. 544–554.

Jorion, P. "International Portfolio Diversification with Estimation Risk." Journal of Business, Vol. 58, No. 3 (1985), pp. 259–278.

Kan, R., and G. Zhou. "Optimal Portfolio Choice With Parameter Uncertainty." Journal of Financial and Quantitative Analysis, Vol. 42, No. 3 (2007), pp. 621–656.

Kazemi, H. Alternative Investment Analyst Review, Chartered Alternative Investment Analyst Association, 2012.

Klarman, S. Margin of Safety: Risk-Averse Value Investing Strategies for the Thoughtful Investor, Harper Business, 1991.

Koenker, R. and K.F. Hallock. "Quantile Regression." The Journal of Economic Perspectives, Vol. 15, No. 4 (Fall 2001), pp. 143–156

Konno, H. and K. Suzuki. " A mean-variance-skewness optimisation model." Journal of Operations Research Society of Japan, Vol. 38, No. 2 (1995), pp. 137-187.

Lakonishok, J., A. Shleifer, and R. Vishny. "Contrarian Investment, Extrapolation, and Risk." The Journal of Finance, Vol. 49, No. 5 (1994), pp. 1541–1578.

Lee, D.K.C., F. P. Kok, and Y.W. Choon. "Moments analysis in risk and performance measurement." The Journal of Wealth Management, 9.1 (2006), pp. 54–65.

Liu S.C., S.Y. Wang , and W.H. Qiu. " A mean-variance-skewness model for portfolio selection with transaction costs." International Journal of Systems Science, 34 (2003), pp. 255–262.

Liu, W. "A liquidity-augmented capital asset pricing model." Journal of Financial Economics, Vol. 82, No. 3 (2006), pp. 61–671.

Maillard S., T. Roncalli, J. Teiletche. "The Properties of Equally Weighted Risk Contribution Portfolios." Journal of Portfolio Management, Vol. 36, No. 4 (2010), pp. 60-70.

Markowitz, H. "Portfolio selection." The journal of finance, Vol. 7, No. 1 (Mar. 1952), pp. 77–91.

Markowitz, H., P. Todd, G. Xu, and Y. Yamane. " Computation of mean-semivariance effcient sets by the critical line algorithm." Annals of Operational Research, 45 (1993), pp. 307–317.

McNamara, J. R. "Portfolio Selection Using Stochastic Dominance Criteria." Decision Sciences , Vol. 29, No.4 (1998), pp. 785–801.

Menchero, J., A. Morozov, and P. Shepard. "Global equity model (GEM2) Research notes." MSCI Research Insight, September 2008, $www.msci.com/www/research-paper/global-equity-model-gem2/015905139$

Nathan A. Bond Bubble Breakdown, Commodities and Strategy Research, 2013.

Stevenson, A., ed. Oxford Dictionary of English, 3rded. Oxford University Press, 2010, $www.oxforddictionaries.com$

Pastor, L. and R.F. Stambaugh. "Liquidity risk and expected stock returns." Journal of Political Economy, 111 (2003), pp. 642–685.

Reinganum, M. R. "A New Empirical Perspective on the CAPM." The Journal of Financial and Quantitative Analysis, Vol. 16, No. 4 (1981), pp. 439–462.

Rizova, S. International Evidence on the Size Effect, Dimensional Fund Advisors, 2006.

Sauer, A. "Faktormodelle und Bewertung am deutschen Aktienmarkt." Fritz Knapp Verlag, Frankfurt, 1994.

Schlag, C. and V. Wohlschieß "Is β dead? - Results for the German Stock Market." Discussion Paper Universität Karlsruhe (TH), 1992.

Simaan Y. "Estimation Risk in Portfolio Selection: The Mean Variance Model versus the Mean Absolute Deviation Model. " Management Science, Vol. 43, No. 10 (1997), pp. 1437-1446.

Stehle, R. "The Size Effect in the German Stock Market." Working Paper Universität Augsburg, 1992.

Swensen D. F. Pioneering Portfolio Management: An Unconventional Approach to Institutional Investment, Fully Revised and Updated, Free Press, 2009.

Vassalou, M. and Y. Xing. "Default Risk in Equity Returns." Journal of Finance, 59 (2004), pp. 831–861.

Tibshirani, R. "Regression Shrinkage and Selection via the Lasso." Journal of Royal Statistical Society, Vol. 58 No. 1 (1996), pp. 267–288.

Zhang, X.F. "Information uncertainty and stock returns." Journal of Finance, Vol. 61, No. 1 (2006), pp. 105–137.

Zheng, Qi, C. Gallagher, and K.B. Kulasekera. "Adaptive Penalized Quantile Regression for High-Dimensional Data." Journal of Statistical Planning and Inference, Vol. 143, No. 6 (June 2013), pp. 1029–1038.

Zou, H. "The Adaptive Lasso and its Oracle Properties." Journal of Statistical Planning and Inference, Vol. 101, No. 476 (2006), pp. 1418–1429.

SFB 649 Discussion Paper Series 2015

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

001 "Pricing Kernel Modeling" by Denis Belomestny, Shujie Ma and Wolfgang Karl Härdle, January 2015.
002 "Estimating the Value of Urban Green Space: A hedonic Pricing Analysis of the Housing Market in Cologne, Germany" by Jens Kolbe and Henry Wüstemann, January 2015.
003 "Identifying Berlin's land value map using Adaptive Weights Smoothing" by Jens Kolbe, Rainer Schulz, Martin Wersing and Axel Werwatz, January 2015.
004 "Efficiency of Wind Power Production and its Determinants" by Simone Pieralli, Matthias Ritter and Martin Odening, January 2015.
005 "Distillation of News Flow into Analysis of Stock Reactions" by Junni L. Zhang, Wolfgang K. Härdle, Cathy Y. Chen and Elisabeth Bommes, January 2015.
006 "Cognitive Bubbles" by Ciril Bosch-Rosay, Thomas Meissnerz and Antoni Bosch-Domènech, February 2015.
007 "Stochastic Population Analysis: A Functional Data Approach" by Lei Fang and Wolfgang K. Härdle, February 2015.
008 "Nonparametric change-point analysis of volatility" by Markus Bibinger, Moritz Jirak and Mathias Vetter, February 2015.
009 "From Galloping Inflation to Price Stability in Steps: Israel 1985–2013" by Rafi Melnick and till Strohsal, February 2015.
010 "Estimation of NAIRU with Inflation Expectation Data" by Wei Cui, Wolfgang K. Härdle and Weining Wang, February 2015.
011 "Competitors In Merger Control: Shall They Be Merely Heard Or Also Listened To?" by Thomas Giebe and Miyu Lee, February 2015.
012 "The Impact of Credit Default Swap Trading on Loan Syndication" by Daniel Streitz, March 2015.
013 "Pitfalls and Perils of Financial Innovation: The Use of CDS by Corporate Bond Funds" by Tim Adam and Andre Guettler, March 2015.
014 "Generalized Exogenous Processes in DSGE: A Bayesian Approach" by Alexander Meyer-Gohde and Daniel Neuhoff, March 2015.
015 "Structural Vector Autoregressions with Heteroskedasticy" by Helmut Lütkepohl and Aleksei Netšunajev, March 2015.
016 "Testing Missing at Random using Instrumental Variables" by Christoph Breunig, March 2015.
017 "Loss Potential and Disclosures Related to Credit Derivatives – A Cross-Country Comparison of Corporate Bond Funds under U.S. and German Regulation" by Dominika Paula Gałkiewicz, March 2015.
018 "Manager Characteristics and Credit Derivative Use by U.S. Corporate Bond Funds" by Dominika Paula Gałkiewicz, March 2015.
019 "Measuring Connectedness of Euro Area Sovereign Risk" by Rebekka Gätjen Melanie Schienle, April 2015.
020 "Is There an Asymmetric Impact of Housing on Output?" by Tsung-Hsien Michael Lee and Wenjuan Chen, April 2015.
021 "Characterizing the Financial Cycle: Evidence from a Frequency Domain Analysis" by Till Strohsal, Christian R. Proaño and Jürgen Wolters, April 2015.

SFB 649, Spandauer Straße 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

SFB 649 Discussion Paper Series 2015

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

022 "Risk Related Brain Regions Detected with 3D Image FPCA" by Ying Chen, Wolfgang K. Härdle, He Qiang and Piotr Majer, April 2015.
023 "An Adaptive Approach to Forecasting Three Key Macroeconomic Variables for Transitional China" by Linlin Niu, Xiu Xu and Ying Chen, April 2015.
024 "How Do Financial Cycles Interact? Evidence from the US and the UK" by Till Strohsal, Christian R. Proaño, Jürgen Wolters, April 2015.
025 "Employment Polarization and Immigrant Employment Opportunities" by Hanna Wielandt, April 2015.
026 "Forecasting volatility of wind power production" by Zhiwei Shen and Matthias Ritter, May 2015.
027 "The Information Content of Monetary Statistics for the Great Recession: Evidence from Germany" by Wenjuan Chen and Dieter Nautz, May 2015.
028 "The Time-Varying Degree of Inflation Expectations Anchoring" by Till Strohsal, Rafi Melnick and Dieter Nautz, May 2015.
029 "Change point and trend analyses of annual expectile curves of tropical storms" by P.Burdejova, W.K.Härdle, P.Kokoszka and Q.Xiong, May 2015.
030 "Testing for Identification in SVAR-GARCH Models" by Helmut Luetkepohl and George Milunovich, June 2015.
031 "Simultaneous likelihood-based bootstrap confidence sets for a large number of models" by Mayya Zhilova, June 2015.
032 "Government Bond Liquidity and Sovereign-Bank Interlinkages" by Sören Radde, Cristina Checherita-Westphal and Wei Cui, July 2015.
033 "Not Working at Work: Loafing, Unemployment and Labor Productivity" by Michael C. Burda, Katie Genadek and Daniel S. Hamermesh, July 2015.
034 "Factorisable Sparse Tail Event Curves" by Shih-Kang Chao, Wolfgang K. Härdle and Ming Yuan, July 2015.
035 "Price discovery in the markets for credit risk: A Markov switching approach" by Thomas Dimpfl and Franziska J. Peter, July 2015.
036 "Crowdfunding, demand uncertainty, and moral hazard - a mechanism design approach" by Roland Strausz, July 2015.
037 ""Buy-It-Now" or "Sell-It-Now" auctions : Effects of changing bargaining power in sequential trading mechanism" by Tim Grebe, Radosveta Ivanova-Stenzel and Sabine Kröger, August 2015.
038 "Conditional Systemic Risk with Penalized Copula" by Ostap Okhrin, Alexander Ristig, Jeffrey Sheen and Stefan Trück, August 2015.
039 "Dynamics of Real Per Capita GDP" by Daniel Neuhoff, August 2015.
040 "The Role of Shadow Banking in the Monetary Transmission Mechanism and the Business Cycle" by Falk Mazelis, August 2015.
041 "Forecasting the oil price using house prices" by Rainer Schulz and Martin Wersing, August 2015.
042 "Copula-Based Factor Model for Credit Risk Analysis" by Meng-Jou Lu, Cathy Yi-Hsuan Chen and Karl Wolfgang Härdle, August 2015.
043 "On the Long-run Neutrality of Demand Shocks" by Wenjuan Chen and Aleksei Netsunajev, August 2015.

SFB 649, Spandauer Straße 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

SFB 649 Discussion Paper Series 2015

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

044 "The (De-)Anchoring of Inflation Expectations: New Evidence from the Euro Area" by Laura Pagenhardt, Dieter Nautz and Till Strohsal, September 2015.

045 "Tail Event Driven ASset allocation: evidence from equity and mutual funds' markets" by Wolfgang Karl Härdle, David Lee Kuo Chuen, Sergey Nasekin, Xinwen Ni and Alla, September 2015.

SFB 649, Spandauer Straße 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

Journal of Business & Economic Statistics

ISSN: 0735-0015 (Print) 1537-2707 (Online) Journal homepage: http://amstat.tandfonline.com/loi/ubes20

Single-Index-Based CoVaR With Very High-Dimensional Covariates

Yan Fan, Wolfgang Karl Härdle, Weining Wang & Lixing Zhu

To cite this article: Yan Fan, Wolfgang Karl Härdle, Weining Wang & Lixing Zhu (2017): Single-Index-Based CoVaR With Very High-Dimensional Covariates, Journal of Business & Economic Statistics, DOI: 10.1080/07350015.2016.1180990

To link to this article: https://doi.org/10.1080/07350015.2016.1180990

Supplementary materials for this article are available online. Please go to http://tandfonline.com/r/JBES

Single-Index-Based CoVaR With Very High-Dimensional Covariates

Yan FAN
School of Statistics and Information, Shanghai University of International Business and Economics, Shanghai 201620, China (*fanyan212@162.com*)

Wolfgang Karl HÄRDLE
C.A.S.E.—Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, 10178 Berlin, Germany, and Singapore Management University, Singapore 178899 (*haerdle@wiwi.hu-berlin.de*)

Weining WANG
King's College London, London, United Kingdom, and Ladislaus von Bortkiewicz Chair of Statistics, Humboldt-Universität zu Berlin, 10178 Berlin, Germany (*wangwein@cms.hu-berlin.de*)

Lixing ZHU
Department of Mathematics, Hong Kong Baptist University, Hong Kong, China, and School of Statistics, Beijing Normal Univeristy, Beijing, China (*lzhu@hkbu.edu.hk*)

Systemic risk analysis reveals the interdependencies of risk factors especially in tail event situations. In applications the focus of interest is on capturing joint tail behavior rather than a variation around the mean. Quantile and expectile regression are used here as tools of data analysis. When it comes to characterizing tail event curves one faces a dimensionality problem, which is important for CoVaR (Conditional Value at Risk) determination. A projection-based single-index model specification may come to the rescue but for ultrahigh-dimensional regressors one faces yet another dimensionality problem and needs to balance precision versus dimension. Such a balance is achieved by combining semiparametric ideas with variable selection techniques. In particular, we propose a projection-based single-index model specification for very high-dimensional regressors. This model is used for practical CoVaR estimates with a systemically chosen indicator. In simulations we demonstrate the practical side of the semiparametric CoVaR method. The application to the U.S. financial sector shows good backtesting results and indicate market coagulation before the crisis period. Supplementary materials for this article are available online.

KEY WORDS: Composite quasi-maximum likelihood estimation; CoVaR; Lasso; Minimum average contrast estimation; Model selection; Quantile single-index regression.

1. INTRODUCTION

It is known to be a challenging task to manage financial risk due to joint extreme events, reflecting the fact that in times of crisis losses tend to spread across a portfolio. The key interest is to understand and forecast the risk exposure of, for example, a financial institution in the market for firm leaders or to identify and select systemic risk relevant factors for government regulators. There is a large amount of literature on measuring systemic risk. We focus on the line of research adopting quantile methods to quantify the tail dependence among financial institutions. In particular, Adrian and Brunnermeier (2011) proposed a systemic risk measure, called CoVaR, with balance sheet characteristics driven individual risk exposure. Furthermore, Hautsch, Schaumburg, and Schienle (2014) introduced an applicable measure of a firm's systemic relevance, explicitly accounting for the company's interconnectedness within the financial sector.

The underlying statistical setting involved is a two-stage linear quantile regression. Several elements of the existing CoVaR methodology are, however, based on questionable assumptions: First, a significant degree of nonlinearity occurs when modeling conditional tail curves. Second, the number of potential risk factors is large in comparison with the amount of available observations. Third, the selected factors are difficult to be interpreted, and need to be summarized to an index. Therefore, one calls for a data driven technique that combines dimension reduction, variable selection, and generalized tail events, for example, expectiles. In this article we address these points and provide a practical CoVaR estimate together with a systemically chosen indicator. The systemic indicator is chosen by the single-index approach, which has a unique feature: the index that yields interpretability and low dimension simultaneously. However, in the case of ultrahigh-dimensional regressors X the single-index approach suffers from singularity problems. Efficient variable selection is the strategy to employ here. Specifically we consider composite regression with general weighted loss and possible ultrahigh-dimensional covariates. Our setup is general, and includes quantile, expectile (and therefore mean as a special case) regression. We offer theoretical properties and demonstrate our

Journal of Business & Economic Statistics
xxxx 2017, Vol. XX, No. XX
DOI: 10.1080/07350015.2016.1180990
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/jbes.

method with applications to firm risk analysis in a CoVaR estimation context.

The basic element of our CoVaR estimation is quantile regression(QR). In many fields of applications such as quantitative finance, econometrics, marketing, and also medical and biological sciences, QR is a fundamental element for data analysis, modeling, and inference. An application in finance is the analysis of time varying value-at-risk (VaR) using the conditional autoregressive value at risk (CaViaR) model; see Engle and Manganelli (2004). The QR estimation may be seen as an estimation problem by assuming an asymmetric ALD (asymmetric Laplace distribution) pseudo likelihood, which not necessarily return an efficient estimator. Therefore, different flexible loss functions are considered in the literature to improve the estimation efficiency, such as composite quantile regression (Zou et al. 2008; Kai, Li, and Zou 2010, 2011). Moreover, Bradic, Fan, and Wang (2011) proposed a general loss function framework for linear models, with a weighted sum of different kinds of loss functions, and the weights are selected to be data driven. Another type of loss considered is in Newey and Powell (1987) corresponding to expectile regression (ER). This is similar in spirit to QR but contains mean regression as a special case. Nonparametric expectile smoothing work with applications to demography can be found in Schnabel and Eilers (2009). The ER curves are alternatives to the QR curves and give us an alternative regression picture.

The difficulty of characterizing an entire distribution partly arises from the high dimensionality of covariates. This asks to strike a balance between model flexibility and statistical precision. To crack this tough nut, dimension reduction techniques of semiparametric type, such as the single-index model, came into the focus of statistical modeling. Wu, Yu, and Yu (2010) and Kong and Xia (2012) considered quantile regression via a single-index model. However, to our knowledge there is no further literature on generalized QR for the single-index model.

In addition to the dimension reduction, there is also the problem (incurred in our CoVaR estimation procedure) of choosing the right variables for projection. This motivates our second goal of this research: variable selection. Kong and Xia (2007), Wang and Yin (2008), and Zeng, He, and Zhu (2012) focused on variable selection in mean regression for the single-index model. The set of ideas presented there, however, have never been applied to a quantile, composite quantile framework, or to an even more general (composite) quasi-likelihood framework. The semiparametric single-index approach that we consider herein will be a good tool for practitioners, as it combines flexibility in modeling with applicability for even very high-dimensional data.

This article is organized as follows: In Section 2, we introduce the basic setup and the estimation algorithm. In Section 3, we build up asymptotic theorems for our model. In Section 4, simulations are carried out. In Section 5, we illustrate our methodology by estimating CoVaR. All the technical details can be found in the Appendix.

2. MACE FOR SINGLE-INDEX MODEL

Let X and Y be p dimensional, continuous random variables, respectively; p can be very large, namely of the rate $\exp(n^{\delta})$, where δ is a constant whose range will be defined in Condition 4 in Section 3. The single-index model (SIM) is defined to be:

$$Y = g\left(X^{\top}\beta^{*}\right) + \varepsilon, \tag{2.1}$$

where $g(\cdot) : \mathbb{R}^1 \longmapsto \mathbb{R}^1$ is an *unknown* smooth link function, β^* is the vector of index parameters, and ε is a continuous variable with mean zero. The interest here is to simultaneously estimate β^* and $g(\cdot)$. The assumptions on error structure can be seen in Condition 3.

2.1 Quasi-Likelihood for the Single-Index Model

Several estimation techniques exist for (2.1), among which the average derivative estimator (ADE) method is one of the oldest ones; see Härdle and Stoker (1989). The semiparametric SIM (2.1) also permits a one-step projection pursuit interpretation, therefore estimation tools from this stream of literature might also be employed; see Huber (1985). The minimum average variance estimation (MAVE) technique aimed at simultaneous estimation of $(\beta^*, g(\cdot))$ was proposed by Xia et al. (2002). Here we will apply a minimum average contrast estimation approach, called MACE. Similar to MAVE, the MACE technique uses double integration but allows more general loss functions. Our estimation framework is new in three aspects. First, we consider a general class of contrast functions that allow us to identify and estimate conditional quantiles, expectiles, and other tail-specific objects. Second, we consider the situation where p might be very large and we add penalty terms that lead to an automatic model selection framework of, for example, the least absolute shrinkage and selection operator (Lasso) or smoothly clipped absolute deviation (SCAD) type. Third, we implement a composite estimation technique for efficiency improvement.

In our theoretical setup, we identify the parameter via a minimum contrast with ρ_{w} as the contrast function. It corresponds, as mentioned above, to a quasi maximum likelihood framework: the direction β^* (for known $g(\cdot)$) is the solution of

$$\min_{\beta} \mathrm{E}\, \rho_{\mathrm{w}}\left\{Y - g\left(X^{\top}\beta\right)\right\}, \tag{2.2}$$

with the general quasi-likelihood loss function $\rho_{\mathrm{w}}(\cdot) = \sum_{k=1}^{K} \mathrm{w}_k \rho_k(\cdot)$, where $\rho_1(\cdot), \ldots, \rho_K(\cdot)$ are convex loss functions and $\mathrm{w}_1, \ldots, \mathrm{w}_K$ are positive weights.

Equivalently, β is the solution to

$$\mathrm{E}\left(\psi_{\mathrm{w}}\left\{Y - g\left(X^{\top}\beta\right)\right\} | X\right) = 0 \quad \text{a.s.}$$

(where $\psi_{\mathrm{w}}(\cdot)$ is the derivative (a subgradient) of $\rho_{\mathrm{w}}(\cdot)$). This weighted loss function includes many situations such as ordinary least square, quantile regression (QR), expectile regression(ER), composite quantile regression (CQR), and so on. For model identification, we assume that the L_2-norm of β^*, $\|\beta^*\|_2 = 1$ and the first component of β^* is positive.

The standard situation of QR is with $K = 1$ and the conditional quantile function $F^{-1}_{\varepsilon|X}(\tau) = 0$. This means to take the loss function as

$$\rho_{\mathrm{w}}(u) = \tau u \mathbf{1}(u \geq 0) - (1-\tau) u \mathbf{1}(u < 0), \tag{2.3}$$

where $\mathbf{1}(A)$ is equal to 1 if A is true and 0 otherwise. Moreover, for ER with $K = 1$, we have:

$$\rho_{\mathrm{w}}(u) = \tau u^2 \mathbf{1}(u \geq 0) + (1-\tau) u^2 \mathbf{1}(u < 0). \tag{2.4}$$

The general form of $\rho_{\mathrm{w}}(\cdot)$ boils down to CQR when one employs K different quantiles $\tau_1, \tau_2, \ldots, \tau_K$, with $w_k = 1/K$,

$k = 1, \ldots, K$ and

$$\rho_k(u) = \tau(u - b_k)\mathbf{1}(u - b_k \geq 0) + (1 - \tau)(u - b_k)\mathbf{1}(u - b_k < 0), \tag{2.5}$$

where b_k is the τ_k quantile of the error distribution; see Bradic, Fan, and Wang (2011).

Let us now launch the MACE. First, we approximate $g(X_i^\top \beta)$ for $x^\top \beta$ near $X_i^\top \beta$:

$$g\left(X_i^\top \beta\right) \approx g\left(x^\top \beta\right) + g'\left(x^\top \beta\right)(X_i - x)^\top \beta. \tag{2.6}$$

In the context of local linear smoothing, a first-order proxy of β (given x) can therefore be constructed by minimizing

$$L_x(\beta, g(\cdot)) \stackrel{\text{def}}{=} \mathrm{E}\, \rho_{\mathrm{w}}\{Y - g\left(x^\top \beta\right) - g'\left(x^\top \beta\right)(X_i - x)^\top \beta\}. \tag{2.7}$$

The empirical version of (2.7) requires minimizing, with respect to β and function $g(\cdot)$:

$$L_{n,x}(\beta, g(\cdot)) \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^{n} \rho_{\mathrm{w}}\left\{Y_i - g\left(x^\top \beta\right) - g'\left(x^\top \beta\right)(X_i - x)^\top \beta\right\} \times K_h\{(X_i - x)^\top \beta\}, \tag{2.8}$$

where $K_h(\cdot)$ is a kernel function with $K_h(u) = h^{-1}K(u/h)$ and h is a bandwidth parameter. We adopt now the double integration idea of MAVE, that is, we integrate with respect to the empirical distribution function of the covariates leading to the following loss function:

$$L_n(\beta, g(\cdot)) \stackrel{\text{def}}{=} n^{-2} \sum_{j=1}^{n}\sum_{i=1}^{n} \rho_{\mathrm{w}}\left\{Y_i - g(X_j^\top \beta) - g'(X_j^\top \beta)(X_i - X_j)^\top \beta\right\} K_h\{(X_i - X_j)^\top \beta\}. \tag{2.9}$$

Minimizing (2.9) with respect to β and $g(\cdot)$ is the basic idea.

For simplicity, from now on we write $g(X_j^\top \beta)$ and $g'(X_j^\top \beta)$ as $a(X_j)$ and $b(X_j)$ or a_j and b_j, respectively. The calculation of the above minimization problem can be decomposed into two subproblems, motivated by, for example, Leng, Xia, and Xu (2008):

a. Given β, the estimation of $a(\cdot)$ and $b(\cdot)$ are obtained through local linear minimization.
b. Given $a(\cdot)$ and $b(\cdot)$, the minimization with respect to β is carried out by the interior point method.

2.2 Variable Selection for Single-Index Model

The dimension of covariates (p) is large, even one can allow $p = \mathcal{O}\{\exp(n^\delta)\}$, so selecting important covariates is a necessary step. Without loss of generality assume that the first q components of β^* minimizing (2.2) are nonzero. To point this out write $\beta^* = (\beta_{(1)}^{*\top}, \beta_{(0)}^{*\top})^\top$ with $\beta_{(1)}^* \stackrel{\text{def}}{=} (\beta_1, \ldots, \beta_q)^\top \neq 0$ and $\beta_{(0)}^* \stackrel{\text{def}}{=} (\beta_{q+1}, \ldots, \beta_p)^\top = 0$ element-wise. Accordingly we denote $\mathbf{X}_{(1)}$ and $\mathbf{X}_{(0)}$ as the first q and the last $p - q$ column of design matrix $\mathbf{X}$, corresponding to $\beta_{(1)}^{*\top}$ and $\beta_{(0)}^{*\top}$, respectively.

Suppose $\{(X_i, Y_i)\}_{i=1}^n$ are n independent and identically distributed (iid) copies of (X, Y). Consider first estimating the SIM coefficient β^* by solving the optimization problem

$$\min_{(a_j, b_j)'\mathrm{s}, \beta} n^{-1} \sum_{j=1}^{n}\sum_{i=1}^{n} \rho_{\mathrm{w}}\left(Y_i - a_j - b_j X_{ij}^\top \beta\right)\omega_{ij}(\beta) + \sum_{l=1}^{p} \gamma_\lambda(|\hat{\beta}_l^{(0)}|)|\beta_l|, \tag{2.10}$$

where $X_{ij} \stackrel{\text{def}}{=} X_i - X_j$, $\omega_{ij}(\beta) \stackrel{\text{def}}{=} K_h(X_{ij}^\top \beta)/\sum_{i=1}^n K_h(X_{ij}^\top \beta)$. Here $\gamma_\lambda(t)$ is some nonnegative function, and $\hat{\beta}^{(0)}$ is an initial estimator of β^* (e.g., linear QR with variable selection). The penalty term in (2.10) is quite general and it covers the most popular variable selection criteria as special cases: the Lasso (Tibshirani 1996) with $\gamma_\lambda(x) = \lambda$ and the SCAD (Fan and Li 2001) with

$$\gamma_\lambda(x) = \lambda\left\{\mathbf{1}(|x| \leq \lambda) - \frac{(|x|^2 - 2c_1\lambda|x| + \lambda^2)_+}{|x|(c_1 - 1)2\lambda} \times \mathbf{1}(\lambda < |x| \leq c_1\lambda) + \frac{(c_1 + 1)\lambda}{2|x|}\mathbf{1}(|x| > c_1\lambda)\right\},$$

with ($c_1 > 2$) and $\gamma_\lambda(x) = \lambda|x|^{-c_2}$ for some $c_2 > 0$ corresponding to the adaptive Lasso (Zou 2006).

We propose to estimate β^* in (2.10) with the MACE iterative procedure described below. Denote $\hat{\beta}_{\mathrm{w}}$ the final estimate of β^*. Specifically, for $t = 1, 2, \ldots$, iterate the following two steps. Denote $\hat{\beta}^{(t)}$ as the estimate at step t.

a. Given $\hat{\beta}^{(t)}$, standardize $\hat{\beta}^{(t)}$ so that $\hat{\beta}^{(t)}$ has length one and positive first component. Then compute

$$(\hat{a}_j^{(t)}, \hat{b}_j^{(t)}) \stackrel{\text{def}}{=} \arg\min_{(a_j, b_j)'\mathrm{s}} \sum_{i=1}^{n} \rho_{\mathrm{w}}\left(Y_i - a_j - b_j X_{ij}^\top \hat{\beta}^{(t)}\right)\omega_{ij}(\hat{\beta}^{(t)}). \tag{2.11}$$

b. Given $(\hat{a}_j^{(t)}, \hat{b}_j^{(t)})$, solve

$$\hat{\beta}^{(t+1)} = \arg\min_{\beta} \sum_{j=1}^{n}\sum_{i=1}^{n} \rho_{\mathrm{w}}\left(Y_i - \hat{a}_j^{(t)} - \hat{b}_j^{(t)} X_{ij}^\top \beta\right) \times \omega_{ij}(\hat{\beta}^{(t)}) + n\sum_{l=1}^{p} \hat{d}_l^{(t)}|\beta_l|, \tag{2.12}$$

where $\hat{d}_l^{(t)} \stackrel{\text{def}}{=} \gamma_\lambda(|\hat{\beta}_l^{(t)}|)$. Please note here that the kernel weights $\omega_{ij}(\cdot)$ use the $\hat{\beta}^{(t)}$ from the step before.

When choosing the penalty parameter λ, we adopt a C_p-type criterion as in Yuan and Lin (2006) instead of the computationally involved cross-validation method. We choose the optimal weights of the convex loss functions ρ_{w} by minimizing the asymptotic variance of the resulting estimator of β^*, and the bandwidth h by criteria proposed in Yu and Jones (1998) for $g(\cdot)$.

3. MAIN THEOREMS

Define $\hat{\beta}_{\mathrm{w}} \stackrel{\text{def}}{=} (\hat{\beta}_{\mathrm{w}(1)}^\top, \hat{\beta}_{\mathrm{w}(2)}^\top)^\top$ as the estimator for $\beta^* \stackrel{\text{def}}{=} (\beta_{(1)}^{*\top}, \beta_{(2)}^{*\top})^\top$ attained by the procedure in (2.11) and (2.12). Let $\hat{\beta}_{\mathrm{w}(1)}$ and $\hat{\beta}_{\mathrm{w}(2)}$ be the first q components and the remaining $p - q$ components of $\hat{\beta}_{\mathrm{w}}$, respectively. If in the iterations, we have the initial estimator $\hat{\beta}_{(1)}^{(0)}$ as a $\sqrt{n/q}$ consistent one for $\beta_{(1)}^*$

(2.12), we will obtain with a very high probability, an oracle estimator of the following type, say $\hat{\beta}_{\mathrm{w}} = (\hat{\beta}^{\top}_{\mathrm{w}(1)}, \mathbf{0}^{\top})^{\top}$, since the oracle knows the true active set $\mathcal{M}_* \stackrel{\mathrm{def}}{=} \{l : \beta^*_l \neq 0\}$. The following theorem shows that the penalized estimator enjoys the oracle property. Define $\hat{\beta}^0$ (note that it is different from the initial estimator $\hat{\beta}^{(0)}_{(1)}$) as the minimizer with the same loss in (2.10) but within subspace $\{\beta \in \mathbb{R}^p : \beta_{\mathcal{M}^c_*} = \mathbf{0}\}$.

We make the following assumptions for the proofs of the theorems in this article. Let $Z_i \stackrel{\mathrm{def}}{=} X_i^{\top}\beta^*$ and $Z_{ij} \stackrel{\mathrm{def}}{=} Z_i - Z_j$.

Condition 1. The kernel $K(\cdot)$ is a continuous symmetric function. The link function $g(\cdot) \in C^2$, where C^2 is the function space consisting of functions with second-order continuous derivatives.

Condition 2. Assume that for all $k = 1, \ldots, K$, $\rho_k(x)$ is convex and not continuous on finite number of points. Suppose $\psi_k(x)$, the derivative (or a subgradient of) of $\rho_k(x)$, satisfies $\mathrm{E}\{\psi_k(\varepsilon)|Z_i\}$ would only be a function related to k such that $\mathrm{E}\{\psi_{\mathrm{w}}(\varepsilon)|Z_i\} = 0$, a.s., $\mathrm{E}\{\psi^2_k(\varepsilon)|Z_i\} < \infty$. Let $H_i(c) \stackrel{\mathrm{def}}{=} \inf_{|v|\le c} \partial\, \mathrm{E}\, \psi_{\mathrm{w}}(\varepsilon_i - v) = C_1$, where $\partial\, \mathrm{E}\, \psi_k(\varepsilon - v)$ is the partial derivative with respect to v, and c and C_1 are positive constants.

Condition 3. $\{(X_i, Y_i)\}_{i=1}^n$ be n iid copies of (X, Y). The density of $\beta^{*\top}X$ is bounded with bounded absolute continuous first-order derivatives on its support. Let $X_{i(1)}$ denote the sub-vector of X_i consisting of its first q elements.

Define

$$C_{1(1)} \stackrel{\mathrm{def}}{=} \mathrm{E}\Big[\, \mathrm{E}_{\varepsilon_i|Z_i} \psi^2_{\mathrm{w}}(\varepsilon_i) \big\{[g'(Z_i)]^2 \big(\mathrm{E}(X_{i(1)}) - X_{i(1)}\big) \big(\mathrm{E}(X_{i(1)}) - X_{i(1)}\big)^{\top}\big\}\Big] \tag{3.1}$$

$$C_{0(1)} \stackrel{\mathrm{def}}{=} \mathrm{E}\Big[\partial\, \mathrm{E}_{\varepsilon_i|Z_i} \psi_{\mathrm{w}}(\varepsilon_i) \big\{[g'(Z_i)]^2 \big(\mathrm{E}(X_{i(1)}|Z_i) - X_{i(1)}\big) \big(\mathrm{E}(X_{i(1)}|Z_i) - X_{i(1)}\big)^{\top}\big\}\Big] \tag{3.2}$$

and the matrix $C_{0(1)}$ satisfies $0 < L_1 \le \lambda_{\min}(C_{0(1)}) \le \lambda_{\max}(C_{0(1)}) \le L_2 < \infty$ for positive constants L_1 and L_2. There exists a constant C_3 such that for all $\beta \in \{\|\beta - \beta^*\| \le C_3\}$,

$$\| \mathrm{E}\big[\partial\, \mathrm{E}_{\varepsilon|Z_i}\{\psi_{\mathrm{w}}(\varepsilon)\} g'(Z_i) \big\{(X_{(0)}|Z_i) - X_{i(0)}\big\} \times \big\{(X_{(1)}|Z_i) - X_{i(1)}\big\}^{\top}\big] \|_{2,\infty} = \mathcal{O}(1),$$

where for a matrix B, $\|B\|_{2,\infty} = \max_{\|u\|=1} \|Bu\|_{\infty}$.

Condition 4. Let $d_l \stackrel{\mathrm{def}}{=} \gamma_{\lambda}(|\beta^*_l|)$ with the penalty parameter $\liminf_{n\to\infty} \lambda \ge n^{-1/2+\alpha_2/2}$ and $D_n \stackrel{\mathrm{def}}{=} \max\{d_l : l \in \mathcal{M}_*\} = \mathcal{O}(n^{\alpha_1-\alpha_2/2}\lambda)$, where $\mathcal{M}_* = \{l : \beta^*_l \neq 0\}$ be the true model. Assume that $\liminf_{n\to\infty} \min_j\{d_j/\lambda : j \in \mathcal{M}^c_*\} > 0$. Furthermore, assume $qh \to 0$ and $h^{-1}\sqrt{q/n} = \mathcal{O}(1)$ as n goes to infinity, $q = \mathcal{O}(n^{\alpha_2})$, $p = \mathcal{O}\{\exp(n^{\delta})\}$, $nh^3 \to \infty$, and $h \to 0$. Also, $0 < \delta < \alpha < \alpha_2/2 < 1/2, \alpha_2/2 < \alpha_1 < 1$.

Condition 5. The error term ε_i satisfies $\mathrm{var}(\varepsilon_i) < \infty$. Assume that for any integer $m \ge 1$

$$\mathrm{E}\big|\psi^m_{\mathrm{w}}(\varepsilon_i)/m!\big| \le s_0 M^m \tag{3.3}$$

where s_0 and M are constants, and $\psi_{\mathrm{w}}(\cdot)$ is the derivative (a subgradient) of $\rho_{\mathrm{w}}(\cdot)$.

Condition 6. The conditional density function $f(\varepsilon|Z_i = u)$ is bounded and absolutely continuous differentiable.

Condition 1 is commonly used and the standard normal probability density function is a kernel satisfying this condition. Condition 2 is made on the weighted loss function so that it admits a quadratic approximation. Condition 2 assumes the dependence structure between errors and the covariates. For the CQR estimation in case of $K > 1$, it means that $F^{-1}_{Y|X}(\tau_k) = g(\beta^{*\top}X) + c(\tau_k)$ for all $\tau_1 \le \tau_k \le \tau_K$, where $c(\tau_k)$ is only a constant depending on τ_k; this is a similar condition as Wang, Li, and He (2012). For $K = 1$ the assumption $\mathrm{E}\{\psi_{\mathrm{w}}(\varepsilon)|X\} = 0$ a.s. to $F^{-1}_{\varepsilon|X}(\tau) = 0$. Under Condition 3, the matrix in the quadratic approximation is nonsingular, so that the resulting estimate of β has a nondegenerate limiting distribution. Condition 4 guarantees that the proposed variable selection and estimation procedure for β is model-consistent. Condition 5 implies a common tail behavior that we employ. Condition 6 is essential for the uniform Bahadur representation, which we adopt in the proof.

Theorem 1. Under Conditions 1–6, the estimators $\hat{\beta}^0$ and $\hat{\beta}_{\mathrm{w}}$ exist and coincide on a set with probability tending to 1. Moreover,

$$\mathrm{P}(\hat{\beta}^0 = \hat{\beta}_{\mathrm{w}}) \ge 1 - (p - q)\exp(-C'n^{\alpha}) \tag{3.4}$$

for a positive constant C'.

It is worth noting that the above results imply the usual sign consistency; see, for example, Fan and Lv (2010). In addition, the theorem requires a relationship between the order of p, q, and the parameter α; see Condition 4.

Theorem 2. Under Conditions 1–6, we have

$$\|\hat{\beta}_{\mathrm{w}(1)} - \beta^*_{(1)}\| = \mathcal{O}_p\{(D_n + n^{-1/2})\sqrt{q}\}. \tag{3.5}$$

For any unit vector $\mathbf{b}$ in $\mathbb{R}^q$, we have

$$\mathbf{b}^{\top} C^{1/2}_{0(1)} C^{-1/2}_{1(1)} C^{1/2}_{0(1)} \sqrt{n}(\hat{\beta}_{\mathrm{w}(1)} - \beta^*_{(1)}) \xrightarrow{\mathcal{L}} \mathrm{N}(0, 1), \tag{3.6}$$

where recall that $C_{1(1)} \stackrel{\mathrm{def}}{=} \mathrm{E}\{\mathrm{E}\{\psi^2_{\mathrm{w}}(\varepsilon_i)|Z_i\}[g'(Z_i)]^2[\mathrm{E}(X_{i(1)}|Z_i) - X_{i(1)}][\mathrm{E}(X_{i(1)}|Z_i) - X_{i(1)}]^{\top}\}$, and $C_{0(1)} \stackrel{\mathrm{def}}{=} \mathrm{E}\{\partial\, \mathrm{E}\, \psi_{\mathrm{w}}(\varepsilon_i)|Z_i\}\{[g'(Z_i)]^2(\mathrm{E}(X_{i(1)}|Z_i) - X_{i(1)})(\mathrm{E}(X_{i(1)}|Z_i) - X_{i(1)})\}^{\top}$. Note that $\mathrm{E}(X_{i(1)}|Z_i)$ denotes a $q \times 1$ dimension vector, and $Z_i \stackrel{\mathrm{def}}{=} X_i^{\top}\beta^*$, $\psi_{\mathbf{w}}(\varepsilon)$ is a choice of the subgradient of $\rho_{\mathbf{w}}(\varepsilon)$ and $\sigma^2_{\mathbf{w}} \stackrel{\mathrm{def}}{=} \mathrm{E}\{[\psi_{\mathrm{w}}(\varepsilon_i)]^2\}/[\partial\, \mathrm{E}\, \psi_{\mathrm{w}}(\varepsilon_i)]^2$, where

$$\partial\, \mathrm{E}\{\psi_{\mathrm{w}}(\cdot)|Z_i\} = \frac{\partial\, \mathrm{E}\{\psi_{\mathrm{w}}(\varepsilon_i - v)^2|Z_i\}}{\partial v}\Big|_{v=0}. \tag{3.7}$$

It is worth noting that in the case of quantile regression, $\sigma^2_{\mathbf{w}} = \tau(1-\tau)/f_{\varepsilon|Z}(0)^2$.

Let us now look at the distribution of the estimated link function $\hat{g}(x^{\top}\hat{\beta}_{\mathrm{w}})$ with the consistent estimate for β^* and the estimate $\hat{g}'(x^{\top}\hat{\beta}_{\mathrm{w}})$ with the consistent estimate of β^* plugged in.

Theorem 3. Under conditions 1–6, let $\mu_j \stackrel{\mathrm{def}}{=} \int u^j K(u)du$ and $\nu_j \stackrel{\mathrm{def}}{=} \int u^j K^2(u)du$, $j = 0, 1, 2$. For any interior point $z = x^{\top}\beta^*$, $f_Z(z)$ is the density of $Z_i, i = 1, \ldots, n$, if $nh^3 \to \infty$

and $h \to 0$, we have

$$\sqrt{nh}\sqrt{f_Z(z)/(\nu_0\sigma_{\mathrm{w}}^2)}\left\{\hat{g}(x^\top\hat{\beta}_{\mathrm{w}}) - g(x^\top\beta^*) - \frac{1}{2}h^2 g''(x^\top\beta^*)\mu_2\partial\,\mathrm{E}\,\psi_{\mathrm{w}}(\varepsilon)\right\} \xrightarrow{\mathcal{L}} \mathrm{N}(0,\ 1).$$

Also, we have

$$\sqrt{nh^3}\sqrt{\left\{f_Z(z)\mu_2^2\right\}/(\nu_2\sigma_{\mathrm{w}}^2)}\left\{\hat{g}'(x^\top\hat{\beta}_{\mathrm{w}}) - g'(x^\top\beta^*)\right\} \xrightarrow{\mathcal{L}} \mathrm{N}(0,\ 1),$$

not that $\sqrt{f_Z(z)/(\nu_0\sigma_{\mathrm{w}}^2)}$ and $\sqrt{f_Z(z)\mu_2^2/(\nu_2\sigma_{\mathrm{w}}^2)}$ are the scaling according to the standard deviations of the estimates, and recall $\sigma_{\mathbf{w}}^2 \stackrel{\mathrm{def}}{=} \mathrm{E}\{[\psi_{\mathrm{w}}(\varepsilon_i)]^2\}/[\partial\,\mathrm{E}\,\psi_{\mathrm{w}}(\varepsilon_i)]^2$.

All the proofs of the theorems can be found in the online supplementary materials.

4. SIMULATION

In this section, we evaluate our technique in several settings, involving different combinations of link functions $g(\cdot)$, distributions of ε, and different choices of (n, p, q, τ)'s, where n is the sample size, p is the dimension of the true parameter β^*, q is the number of nonzero components in β^*, and τ represents the quantile level. The evaluation is first done with a simple quantile loss function, and then with the composite $L_1 - L_2$ and the composite quantile cases. The weights $\mathrm{w}_1, \ldots, \mathrm{w}_K$ are preestimated by minimizing the object $\sum_l^K \sum_k^K \mathrm{w}_l\mathrm{w}_k \sum_{i=1}^n \psi_l(\hat{\varepsilon}_i^{(0)})\psi_k(\hat{\varepsilon}_i^{(0)})$, where $\hat{\varepsilon}_i^{(0)}$'s are residuals for the initial estimator.

4.1 Link Functions

Consider the following nonlinear link functions $g(\cdot)$s. Model 1:

$$Y_i = 5\cos(D_1 \cdot Z_i) + \exp\left(-D_1 \cdot Z_i^2\right) + \varepsilon_i, \tag{4.1}$$

where $Z_i = X_i^\top\beta^*$, $D_1 = 0.01$ is a scaling constant, and ε_i is an error term. Model 2:

$$Y_i = 10\sin\{\pi(A \cdot Z_i - B)\} + \varepsilon_i, \tag{4.2}$$

with the parameters $A = 0.3$, $B = 3$. Finally, Model 3 is with $D_2 = 0.1$:

$$Y_i = 10\sin(D_2 \cdot Z_i) + \sqrt{|\sin(0.5 \cdot Z_i) + \varepsilon_i|}. \tag{4.3}$$

4.2 Criteria

For estimation accuracy for β and $g(\cdot)$, we use the following four criteria to measure:

1. Standardized L_2 norm:

$$\mathrm{Dev} \stackrel{\mathrm{def}}{=} \frac{\|\beta^* - \widehat{\beta}\|}{\|\beta^*\|},$$

2. Sign consistency:

$$\mathrm{Acc} \stackrel{\mathrm{def}}{=} \sum_{l=1}^{p} |\mathbf{1}\{\beta_l^* \neq 0\} - \mathbf{1}\{\widehat{\beta}_l \neq 0\}|,$$

3. Least angle:

$$\mathrm{Angle} \stackrel{\mathrm{def}}{=} \frac{\beta^{*\top}\widehat{\beta}}{\|\beta^*\| \cdot \|\widehat{\beta}\|},$$

4. Average squared error:

$$\mathrm{ASE} \stackrel{\mathrm{def}}{=} \frac{1}{n}\sum_{i=1}^{n}\left\{g(Z_i) - \widehat{g}(\widehat{Z}_i)\right\}^2.$$

4.3 L_1-Norm Quantile Regression

We adopt the algorithm for the L_1-norm quantile regression developed by Li and Zhu (2008). The initial estimate of β^* can be calculated by the L_1-norm quantile regression, and then we perform the two-step iterations mentioned in Section 2. Recall that $\mathbf{X}$ is a $p \times n$ matrix, and q is the number of nonzero components in β^*. The jth column of $\mathbf{X}$ is an iid sample from $\mathrm{N}(j/2, 1)$. Two error distributions are considered: $\varepsilon_i \sim \mathrm{N}(0, 0.1)$ and $t(5)$. Note that $\beta_{(1)}^*$ is the vector of the nonzero components in β^*. In the simulation, we consider different $\beta_{(1)}^*$: $\beta_{(1)}^{*\top} = (5, 5, 5, 5, 5)$, $\beta_{(1)}^{*\top} = (5, 4, 3, 2, 1)$, and $\beta_{(1)}^{*\top} = (5, 2, 1, 0.8, 0.2)$. Here the indices Z_i's are rescaled to $[0, 1]$ for nonparametric estimation. The bandwidth is selected as in Yu and Jones (1998):

$$h_\tau = h_{\mathrm{mean}}\left[\tau(1-\tau)\varphi\{\Phi^{-1}(\tau)\}^{-2}\right]^{0.2},$$

where h_{mean} can be calculated by using the direct plug-in methodology of a local linear regression described by Ruppert, Sheather, and Wand (1995). To see the performance of the bandwidth selection, we compare the estimated link functions with different bandwidths. Figure 1 is an example showing the true link function (gray) and the estimated link function (black). The left plot in Figure 1 is with the bandwidth ($h = 0.68$) selected

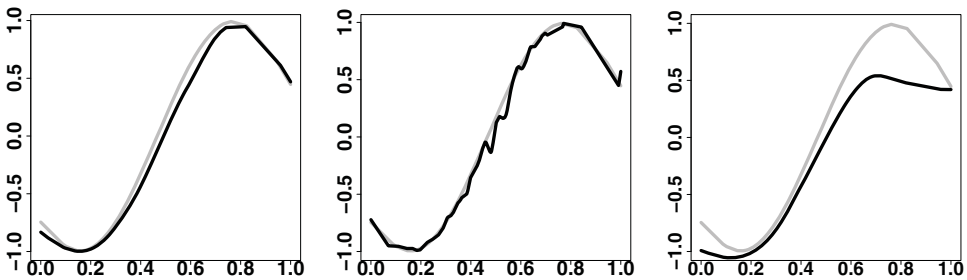


Figure 1. The true link functions (gray) and the estimated link functions (black) in Model 2 with $\beta_{(1)}^{*\top} = (5, 5, 5, 5, 5)$, and $\varepsilon \sim \mathrm{N}(0, 0.1)$, $n = 100$, $p = 10$, $q = 5$, $\tau = 0.05$, where $h = 0.68$ (left), $h = 0.068$ (middle), and $h = 0.8$ (right).

Table 1. Criteria evaluated with different models and quantiles

$g(\cdot)$	ε	τ	Dev	Acc	Angle	ASE
		0.95	1.213 (0.332)	0.949 (0.327)	9.656 (0.086)	0.357 (0.085)
	N	0.50	1.132 (0.137)	0.993 (0.244)	9.736 (0.022)	0.247 (0.044)
		0.05	1.346 (0.532)	1.335 (0.443)	9.626 (0.116)	0.260 (0.076)
Model 1		0.95	1.736 (0.744)	0.926 (0.478)	9.548 (0.135)	0.809 (0.097)
	t	0.50	1.236 (0.246)	1.157 (0.357)	9.667 (0.040)	0.448 (0.093)
		0.05	1.536 (0.737)	2.447 (0.446)	9.570 (0.126)	0.923 (0.097)
		0.95	4.679 (0.854)	6.579 (0.643)	9.581 (0.658)	1.768 (0.247)
	N	0.50	1.489 (0.458)	5.015 (0.436)	9.455 (0.274)	1.156 (0.464)
		0.05	1.501 (0.825)	6.858 (0.747)	9.388 (0.658)	2.015 (0.274)
Model 2		0.95	5.325 (0.960)	9.226 (0.758)	9.360 (0.567)	2.467 (0.351)
	t	0.50	1.689 (0.557)	7.004 (0.879)	9.409 (0.379)	1.279 (0.473)
		0.05	2.065 (0.847)	8.546 (0.951)	9.475 (0.531)	2.639 (0.368)
		0.95	0.757 (0.269)	1.702 (0.248)	9.966 (0.013)	0.569 (0.162)
	N	0.50	0.618 (0.175)	1.434 (0.186)	9.867 (0.021)	0.695 (0.104)
		0.05	0.558 (0.315)	1.845 (0.173)	9.979 (0.024)	0.758 (0.173)
Model 3		0.95	0.625 (0.287)	1.849 (0.284)	9.836 (0.038)	0.736 (0.174)
	t	0.50	0.647 (0.135)	1.655 (0.303)	9.758 (0.029)	0.789 (0.115)
		0.05	0.918 (0.260)	1.879 (0.334)	9.879 (0.036)	0.847 (0.283)

NOTE: $\beta_{(1)}^{*\top} = (5, 5, 5, 5, 5)$, N means the error ε follows a $\mathrm{N}(0, 0.1)$ distribution, t means the error ε follows a $t(5)$ distribution. In 10,000 simulations we set $n = 100$, $p = 10$, $q = 5$. Standard deviations are given in brackets. Dev, Acc, Angle, Error, and their standard deviations are reported in 10^{-1}. ASE and its standard deviations are reported in 10^{-2}.

by applying the aforementioned bandwidth selection. We can see that the estimated link function curve is relatively smooth. The middle plot shows the estimated link function with a smaller bandwidth ($h = 0.068$). It can be seen that the estimated curve is wiggly shaped. The right plot shows the estimated link function with a larger bandwidth ($h = 0.8$); the deviation between the estimated link function curve and the true curve is very large.

Table 1 shows the criteria evaluated with different models and quantile levels. Here $\beta_{(1)}^{*\top} = (5, 5, 5, 5, 5)$, the error ε follows a $\mathrm{N}(0, 0.1)$ distribution or follows a $t(5)$ distribution. In 10,000 simulations we set $p = 10$, $q = 5$. Standard deviations are given in brackets. We find that for quantile levels 0.95 and 0.05, the errors are usually slightly larger than the median. Although the estimations for Model 2 are not as good as for Models 1 and 3, the errors are still moderate. Figures 2 and 3 present the plots of the true link functions against the estimated ones for different quantile levels.

Table 2 reports on the criteria evaluated under different $\beta_{(1)}^{*}$ cases. In this table two different $\beta_{(1)}^{*}$ are considered: (a) $\beta_{(1)}^{*\top} = (5, 4, 3, 2, 1)$, (b) $\beta_{(1)}^{*\top} = (5, 2, 1, 0.8, 0.2)$, the error ε

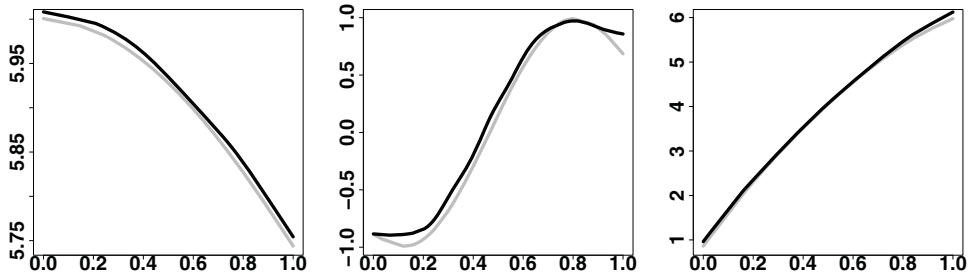


Figure 2. The true link functions (gray) and the estimated link functions (black) with $\beta_{(1)}^{*\top} = (5, 5, 5, 5, 5)$, and $\varepsilon \sim \mathrm{N}(0, 0.1)$, $n = 100$, $p = 10$, $q = 5$, $\tau = 0.95$, Model 1 (left) with $h = 1.02$, Model 2 (middle) with $h = 0.15$, and Model 3 (right) with $h = 0.76$.

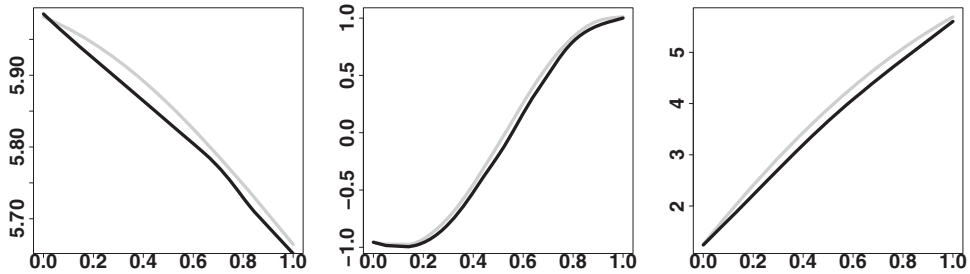


Figure 3. The true link functions (gray) and the estimated link functions (black) with $\beta_{(1)}^{*\top} = (5, 5, 5, 5, 5)$, and $\varepsilon \sim \mathrm{N}(0, 0.1)$, $n = 100$, $p = 10$, $q = 5$, $\tau = 0.05$, Model 1 (left) with $h = 0.78$, Model 2 (middle) with $h = 0.12$, and Model 3 (right) with $h = 0.78$.

Table 2. Criteria evaluated with different models

n	$g(\cdot)$	$\beta^*_{(1)}$	Dev	Acc	Angle	ASE
	Model 1	(a)	1.402 (0.351)	1.009 (0.361)	9.735 (0.071)	0.232 (0.094)
		(b)	1.718 (0.393)	1.313 (0.391)	9.391 (0.084)	0.353 (0.119)
100	Model 2	(a)	1.849 (0.867)	7.367 (0.944)	9.446 (0.423)	1.451 (0.852)
		(b)	2.304 (0.913)	9.505 (0.958)	9.341 (0.556)	1.845 (0.914)
	Model 3	(a)	0.406 (0.256)	1.519 (0.243)	9.643 (0.066)	0.857 (0.125)
		(b)	0.835 (0.294)	1.781 (0.289)	9.426 (0.073)	0.906 (0.136)
	Model 1	(a)	1.318 (0.368)	0.825 (0.221)	9.756 (0.062)	0.179 (0.088)
		(b)	1.409 (0.312)	0.956 (0.252)	9.682 (0.079)	0.302 (0.073)
200	Model 2	(a)	1.833 (0.751)	5.126 (0.936)	9.476 (0.392)	1.338 (0.701)
		(b)	2.257 (0.887)	7.366 (0.910)	9.385 (0.460)	1.754 (0.843)
	Model 3	(a)	0.389 (0.231)	1.597 (0.288)	9.632 (0.052)	0.777 (0.112)
		(b)	0.533 (0.281)	1.624 (0.290)	9.538 (0.061)	0.864 (0.129)
	Model 1	(a)	1.012 (0.287)	0.714 (0.225)	9.846 (0.061)	0.124 (0.073)
		(b)	1.302 (0.301)	0.854 (0.245)	9.797 (0.070)	0.287 (0.061)
500	Model 2	(a)	1.622 (0.564)	5.024 (0.821)	9.495 (0.302)	1.204 (0.592)
		(b)	2.176 (0.636)	6.015 (0.801)	9.452 (0.363)	1.512 (0.614)
	Model 3	(a)	0.361 (0.211)	1.419 (0.202)	9.781 (0.029)	0.626 (0.091)
		(b)	0.423 (0.235)	1.612 (0.236)	9.652 (0.037)	0.751 (0.111)

NOTE: Two different $\beta^*_{(1)}$: (a) $\beta^{*\top}_{(1)} = (5, 4, 3, 2, 1)$, (b) $\beta^{*\top}_{(1)} = (5, 2, 1, 0.8, 0.2)$; the error ε follows a $\mathrm{N}(0, 0.1)$ distribution. In 10,000 simulations we set $p = 10, q = 5, \tau = 0.95$. Standard deviations are given in brackets. Dev, Acc, Angle, and their standard deviations are reported in 10^{-1}; ASE and its standard deviations are reported in 10^{-2}.

follows a $\mathrm{N}(0, 0.1)$ distribution. In 10,000 simulations we set $p = 10, q = 5, \tau = 0.95$. Standard deviations are given in brackets. We notice that for the case (b), the estimation results are not better than (a) since the smaller values of $\beta^*_{(1)}$ in case (b) would be estimated as zeros, and the estimation of the link function would be affected as well. Figures 5 and 6 are the plots of the estimated link functions in these two cases.

Table 3 shows the criteria evaluated under the $p > n$ case. Here $\beta^{*\top}_{(1)} = (5, 5, 5, 5, 5)$, the error ε follows a $\mathrm{N}(0, 0.1)$ distribution. In 10,000 simulations we set $p = 200, q = 5, \tau = 0.05$. Standard deviations are given in brackets. We find that the errors are still moderate in the $p > n$ situation compared with Table 1. Figure 7 shows the graphs in this case.

4.4 Composite L_1-L_2 Regression

In this section, a combined L_1 and L_2 loss is considered and thus the corresponding optimization is formed as

$$\arg\min_{\beta, g(\cdot)} \Big[\sum_{i=1}^{n} \mathrm{w}_1 |Y_i - g\left(X_i^\top \beta\right)| + \mathrm{w}_2 \sum_{i=1}^{n} \{Y_i - g\left(X_i^\top \beta\right)\}^2 \omega_i(\beta) + n \sum_{l=1}^{p} \gamma_\lambda(|\beta_l|)|\beta_l| \Big]. \tag{4.4}$$

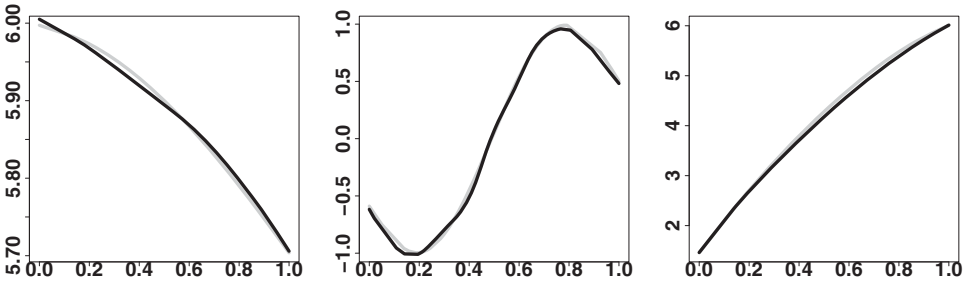


Figure 4. The true link functions (gray) and the estimated link functions (black) with $\beta^{*\top}_{(1)} = (5, 5, 5, 5, 5)$, and $\varepsilon \sim \mathrm{N}(0, 0.1)$, $n = 100$, $p = 10$, $q = 5$, $\tau = 0.5$, Model 1 (left) with $h = 0.55$, Model 2 (middle) with $h = 0.13$, and Model 3 (right) with $h = 0.65$.

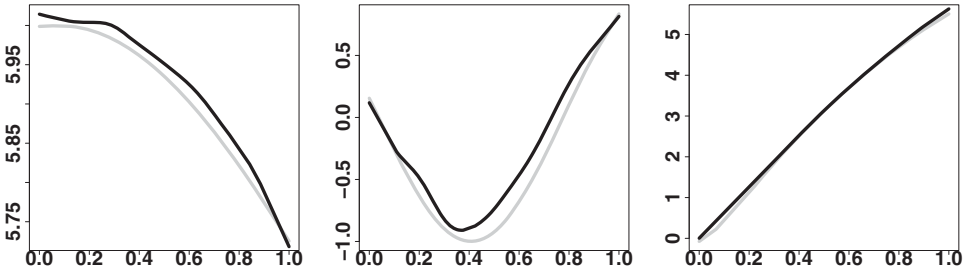


Figure 5. The true link functions (gray) and the estimated link functions (black) with $\beta^{*\top}_{(1)} = (5, 4, 3, 2, 1)$, and $\varepsilon \sim \mathrm{N}(0, 0.1)$, $n = 100$, $p = 10$, $q = 5$, $\tau = 0.95$, Model 1 (left) with $h = 0.31$, Model 2 (middle) with $h = 0.09$, and Model 3 (right) with $h = 0.8$.

Table 3. Criteria evaluated with different models $p \geq n$ case

n	$g(\cdot)$	Dev	Acc	Angle	ASE
100	Model 1	1.880 (0.753)	2.535 (0.847)	9.303 (0.157)	1.812 (0.239)
	Model 2	2.859 (0.954)	9.613 (1.411)	9.035 (0.835)	3.465 (0.936)
	Model 3	1.554 (0.635)	3.143 (0.866)	9.265 (0.095)	3.354 (0.297)
200	Model 1	1.865 (0.744)	1.818 (0.724)	9.331 (0.125)	1.103 (0.233)
	Model 2	2.433 (0.822)	8.499 (1.222)	9.112 (0.709)	2.224 (0.931)
	Model 3	1.415 (0.602)	2.001 (0.713)	9.303 (0.079)	2.915 (0.203)

NOTE: $\beta^{*\top}_{(1)} = (5, 5, 5, 5, 5)$; the error ε follows a $\mathrm{N}(0, 0.1)$ distribution. In 10, 000 simulations we set $p = 200$, $q = 5$, $\tau = 0.05$. Standard deviations are given in brackets. Dev, Acc, Angle, and their standard deviations are reported in 10^{-1}; ASE and its standard deviations are reported in 10^{-2}.

It can be further formulated as

$$\arg\min_{\beta, g(\cdot)} \Big[\sum_{i=1}^{n} \{ w_1 |Y_i - g\left(X_i^\top \beta\right)|^{-1} + w_2 \} |Y_i - g\left(X_i^\top \beta\right)|^2 \times \omega_i(\beta) + n \sum_{l=1}^{p} \gamma_\lambda(|\beta_l|)|\beta_l| \Big]. \quad (4.5)$$

Let $\mathrm{Res}_i^t \overset{\mathrm{def}}{=} Y_i - \hat{g}^t(X_i^\top \hat{\beta}^t)$ be the residual at tth step, and the final estimate can be acquired by the iteration between $g(\cdot)$ and β until convergence:

$$\arg\min_{\beta, g(\cdot)} \Big[\sum_{i=1}^{n} \{ w_1 |\mathrm{Res}_i^t|^{-1} + w_2 \} |Y_i - g\left(X_i^\top \beta\right)|^2 \omega_i(\hat{\beta}^{(t)}) + n \sum_{l=1}^{p} \gamma_\lambda(|\beta_l|)|\beta_l| \Big]. \quad (4.6)$$

Three different settings are conducted. The results are reported in Table 4. Figure 8 (the upper panel) shows the difference between the estimated and true $g(\cdot)$ functions. The level of estimation error is roughly the same as the previous level. Also the results would not change too much with respect to the error distributions and the increasing dimension of p, since only the dimension of q matters.

4.5 Composite L_1 Quantile Regression

We use majorize-minimization (MM) algorithm for a large-scale regression problem. Table 5 shows the estimation quality. Compared with the results in Table 1, the estimation efficiency is improved, even in the case of $p > n$. Figure 8 presents the plots of the estimated link functions for different models using both the composite L_1 regression and the $L_1 - L_2$ regression.

5. APPLICATION

In this section, we apply the proposed methodology to analyze risk for a specific firm conditioning on macro and other firm variables. More specifically, for small financial firms, we aim to detect the contagion effects and the potential risk contri-

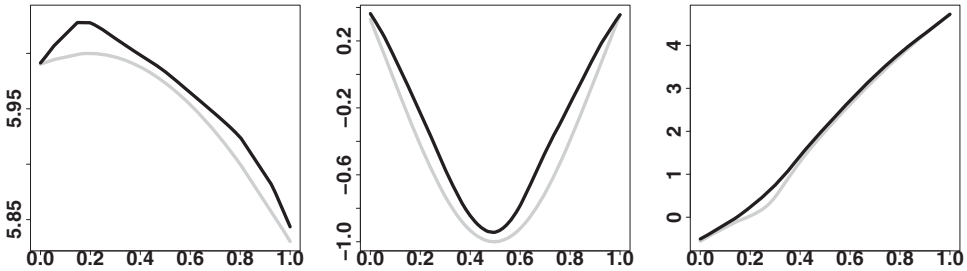


Figure 6. The true link functions (gray) and the estimated link functions (black) with $\beta^{*\top}_{(1)} = (5, 2, 1, 0.8, 0.2)$, and $\varepsilon \sim \mathrm{N}(0, 0.1)$, $n = 100$, $p = 10$, $q = 5$, $\tau = 0.95$, Model 1 (left) with $h = 0.21$, Model 2 (middle) with $h = 0.18$, and Model 3 (right) with $h = 0.25$.

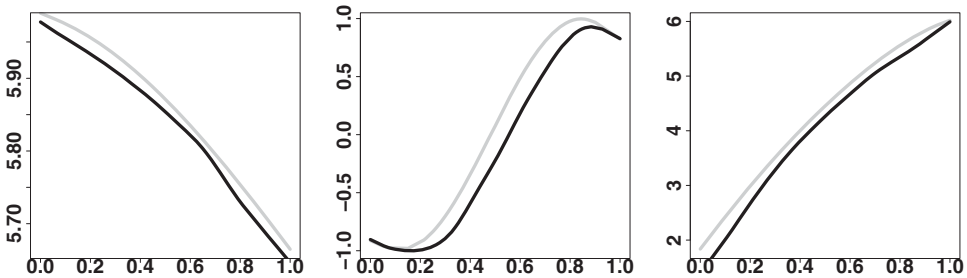


Figure 7. The true link functions (gray) and the estimated link functions (black) with $\beta^{*\top}_{(1)} = (5, 5, 5, 5, 5)$, and $\varepsilon \sim \mathrm{N}(0, 0.1)$, $n = 100$, $p = 200$, $q = 5$, $\tau = 0.05$, Model 1 (left) with $h = 0.81$, Model 2 (middle) with $h = 0.22$, and Model 3 (right) with $h = 0.57$.

Table 4. Simulation results under sparsity, nonsparsity, and large p cases

n	Model	Settings	ε	Dev	Acc	Angle	ASE
		$p = 10, q = 2$	N	1.033 (0.141)	1.037 (0.231)	9.888 (0.016)	0.223 (0.031)
			t	1.223 (0.230)	1.132 (1.237)	9.860 (0.021)	0.281 (0.047)
	Model 1	$p = 10, q = 7$	N	1.163 (0.201)	1.219 (0.211)	9.833 (0.023)	0.290 (0.049)
			t	1.444 (0.232)	1.298 (0.277)	9.805 (0.050)	0.318 (0.079)
		$p = 100, q = 5$	N	1.484 (0.303)	1.624 (1.426)	9.344 (0.091)	0.473 (0.216)
			t	1.576 (0.365)	1.845 (0.445)	9.311 (0.106)	0.534 (0.223)
100	Model 2	$p = 10, q = 2$	N	1.134 (0.277)	6.392 (0.381)	9.399 (0.125)	1.146 (0.216)
			t	1.235 (0.295)	6.442 (0.412)	9.391 (0.136)	1.241 (0.227)
		$p = 10, q = 7$	N	1.323 (0.346)	7.723 (0.682)	9.281 (0.287)	1.401 (0.321)
			t	1.706 (0.368)	7.953 (0.704)	9.259 (0.314)	1.577 (0.361)
		$p = 100, q = 5$	N	1.207 (0.483)	8.387 (0.891)	9.230 (0.359)	1.728 (0.673)
			t	1.994 (0.494)	8.543 (0.903)	9.142 (0.416)	1.751 (0.701)
	Model 3	$p = 10, q = 2$	N	0.880 (0.153)	1.254 (0.143)	9.968 (0.018)	0.550 (0.091)
			t	1.077 (0.175)	1.366 (0.145)	9.951 (0.023)	0.740 (0.102)
		$p = 10, q = 7$	N	1.285 (0.183)	1.553 (0.197)	9.950 (0.036)	0.838 (0.127)
			t	1.334 (0.195)	1.680 (0.257)	9.947 (0.048)	0.843 (0.139)
		$p = 100, q = 5$	N	1.369 (0.235)	2.023 (0.636)	9.377 (0.054)	1.304 (0.182)
			t	1.494 (0.383)	2.293 (0.652)	9.344 (0.063)	1.880 (0.197)
	Model 1	$p = 10, q = 2$	N	1.203 (0.132)	0.999 (0.193)	9.898 (0.010)	0.214 (0.031)
			t	1.338 (0.147)	1.019 (0.201)	9.835 (0.009)	0.237 (0.035)
		$p = 10, q = 7$	N	1.208 (0.166)	1.118 (0.218)	9.882 (0.012)	0.309 (0.046)
			t	1.457 (0.178)	1.236 (0.242)	9.802 (0.018)	0.306 (0.072)
		$p = 100, q = 5$	N	1.434 (0.183)	1.478 (0.396)	9.323 (0.063)	0.332 (0.152)
			t	1.482 (0.217)	1.646 (0.401)	9.315 (0.088)	0.491 (0.179)
500	Model 2	$p = 10, q = 2$	N	1.036 (0.222)	6.021 (0.311)	9.598 (0.133)	1.026 (0.211)
			t	1.133 (0.290)	6.129 (0.411)	9.435 (0.169)	1.198 (0.231)
		$p = 10, q = 7$	N	1.152 (0.229)	7.069 (0.518)	9.402 (0.212)	1.364 (0.288)
			t	1.468 (0.289)	7.188 (0.625)	9.382 (0.268)	1.473 (0.306)
		$p = 100, q = 5$	N	1.773 (0.461)	8.327 (0.794)	9.207 (0.281)	1.691 (0.652)
			t	1.872 (0.489)	8.376 (0.864)	9.141 (0.299)	1.706 (0.691)
	Model 3	$p = 10, q = 2$	N	0.746 (0.102)	1.023 (0.103)	9.590 (0.013)	0.498 (0.081)
			t	0.865 (0.169)	1.215 (0.128)	9.481 (0.020)	0.502 (0.099)
		$p = 10, q = 7$	N	0.992 (0.187)	1.436 (0.137)	9.487 (0.029)	0.578 (0.112)
			t	1.003 (0.193)	1.478 (0.186)	9.459 (0.032)	0.624 (0.131)
		$p = 100, q = 5$	N	1.209 (0.203)	1.646 (0.468)	9.381 (0.041)	0.847 (0.165)
			t	1.402 (0.353)	2.219 (0.579)	9.343 (0.053)	0.781 (0.194)

NOTE: N means errors follow i.i.d. $N(0, 0.1)$; t means t distribution with degree of 5. Dev, Acc, Angle, and their standard deviations are reported in 10^{-1}; ASE and its standard deviations are reported in 10^{-2}.

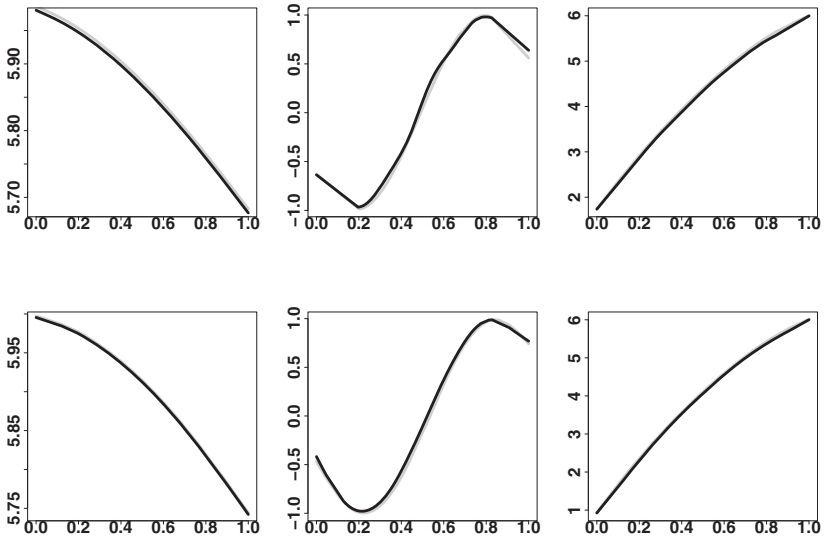


Figure 8. Plot of the true function $g(\cdot)$ (gray) and the estimation (black) with $n = 100$, $p = 10, q = 5$ and $\varepsilon \sim N(0, 0.1)$ in different $g(\cdot)$ functions. L_1–L_2 regression, $h = 0.6, 0.3, 0.4$ (upper pannel); composite quantile $h = 0.5, 0.2, 0.5$ (lower panel).

Table 5. Simulation results for composite L_1 quantile regression

n	Model	Settings	ε	Dev	Acc	Angle	ASE
		$p = 10, q = 2$	N	2.638 (0.053)	0.774 (0.149)	9.993 (0.013)	0.142 (0.022)
			t	1.038 (0.125)	0.899 (0.156)	9.991 (0.014)	0.145 (0.031)
	Model 1	$p = 30, q = 3$	N	1.148 (0.141)	1.072 (0.175)	9.828 (0.011)	0.169 (0.043)
			t	1.166 (0.106)	1.197 (0.193)	9.576 (0.012)	0.257 (0.063)
		$p = 120, q = 5$	N	1.183 (0.186)	1.207 (0.191)	9.421 (0.040)	0.332 (0.114)
			t	1.336 (0.215)	1.219 (0.201)	9.403 (0.063)	0.367 (0.119)
100	Model 2	$p = 10, q = 2$	N	1.119 (0.213)	4.001 (0.282)	9.592 (0.101)	1.112 (0.212)
			t	1.215 (0.241)	4.086 (0.323)	9.499 (0.117)	1.244 (0.218)
		$p = 30, q = 3$	N	1.335 (0.252)	5.154 (0.393)	9.595 (0.132)	1.304 (0.311)
			t	1.359 (0.282)	5.538 (0.462)	9.583 (0.168)	1.383 (0.381)
		$p = 120, q = 5$	N	1.742 (0.289)	6.703 (0.504)	9.382 (0.202)	1.453 (0.412)
			t	1.946 (0.320)	7.335 (0.611)	9.363 (0.310)	1.626 (0.503)
	Model 3	$p = 10, q = 2$	N	0.415 (0.086)	1.007 (0.100)	9.974 (0.011)	0.426 (0.041)
			t	0.512 (0.093)	1.032 (0.113)	9.968 (0.013)	0.493 (0.059)
		$p = 30, q = 3$	N	0.841 (0.143)	1.167 (0.139)	9.965 (0.013)	0.528 (0.060)
			t	0.953 (0.153)	1.235 (0.155)	9.962 (0.022)	0.560 (0.069)
		$p = 120, q = 5$	N	0.883 (0.161)	1.357 (0.168)	9.575 (0.034)	0.892 (0.104)
			t	0.903 (0.233)	1.946 (0.273)	9.553 (0.044)	0.949 (0.113)
	Model 1	$p = 10, q = 2$	N	0.935 (0.102)	0.609 (0.102)	9.998 (0.003)	0.114 (0.018)
			t	1.026 (0.134)	0.774 (0.124)	9.992 (0.005)	0.125 (0.029)
		$p = 30, q = 3$	N	1.132 (0.142)	0.852 (0.138)	9.993 (0.005)	0.133 (0.033)
			t	1.148 (0.116)	0.945 (0.165)	9.991 (0.006)	0.174 (0.049)
		$p = 120, q = 5$	N	1.157 (0.125)	1.144 (0.185)	9.543 (0.030)	0.247 (0.110)
			t	1.275 (0.166)	1.232 (0.196)	9.572 (0.046)	0.303 (0.115)
500	Model 2	$p = 10, q = 2$	N	1.104 (0.206)	3.908 (0.260)	9.691 (0.053)	1.009 (0.116)
			t	1.185 (0.214)	4.105 (0.273)	9.685 (0.055)	1.216 (0.151)
		$p = 30, q = 3$	N	1.286 (0.219)	4.239 (0.294)	9.552 (0.050)	1.309 (0.216)
			t	1.294 (0.278)	5.046 (0.347)	9.504 (0.127)	1.316 (0.231)
		$p = 120, q = 5$	N	1.727 (0.246)	5.675 (0.405)	9.459 (0.134)	1.448 (0.317)
			t	1.824 (0.289)	5.856 (0.581)	9.443 (0.168)	1.497 (0.413)
	Model 3	$p = 10, q = 2$	N	0.380 (0.076)	0.996 (0.087)	9.993 (0.010)	0.391 (0.040)
			t	0.508 (0.087)	1.022 (0.116)	9.990 (0.016)	0.446 (0.048)
		$p = 30, q = 3$	N	0.763 (0.092)	1.154 (0.125)	9.982 (0.016)	0.514 (0.051)
			t	0.846 (0.104)	1.265 (0.142)	9.971 (0.020)	0.546 (0.064)
		$p = 120, q = 5$	N	0.966 (0.113)	1.843 (0.193)	9.833 (0.022)	0.768 (0.087)
			t	1.124 (0.235)	1.898 (0.237)	9.742 (0.031)	0.830 (0.104)

NOTE: N means errors follow iid N$(0, 0.1)$; t means t distribution with degree of 5. Dev, Acc, Angle, Error, and their standard deviations are reported in 10^{-1}; ASE and its standard deviations are reported in 10^{-2}.

butions from larger firms and other market variables. As a result one identifies a risk index, which is expressed as a linear combination, composed of selected large firm returns and market prudential variables.

5.1 Data and Risk Calibration

The firm data are selected according to the ranking of NASDAQ. We take as an example, city national corp. (CYN) as our dependent variable. The remaining 199 financial institutions together with seven lagged macro variables are chosen as covariates. The list of these firms comes from the website: *http://www.nasdaq.com/screening/companies-by-industry.aspx?industry=Finance*. The daily stock prices of these 200 firms are from Yahoo Finance for the period from January 5, 2006, to October 30, 2015. The descriptive statistics of the company, the description of the macro variables, and the list of the firms (Tables A.2–A.4) can be found in the Appendix. To evaluate the risk exposure of the firm CYN, we adopt a modified two-step quantile regression procedure that involves our quantile single-index model in the second step. The first one is a quantile regression to calculate the VaR of all the covariates, respectively. For this propose, one performs QR of log returns of each covariate on all the lagged macro variables:

Table 6. The p-values for CaViaR test for $\widehat{\text{VaR}}$, $\widehat{\text{CoVaR}}_L$, and $\widehat{\text{CoVaR}}_{\text{SIM}}$ for CYN

p-Value	Overall	Crisis
$\widehat{\text{VaR}}$	1.2×10^{-6}	0.99
$\widehat{\text{CoVaR}}_L$	0.01	3.2×10^{-5}
$\widehat{\text{CoVaR}}_{SIM}$	0.46	0.93

NOTE: $T = 2335$ in overall period (20060710 − 20151030) and crisis period (20080915 − 20100208).

$$X_{i,t} = \alpha_i + \gamma_i^\top M_{t-1} + \varepsilon_{i,t}, \tag{5.1}$$

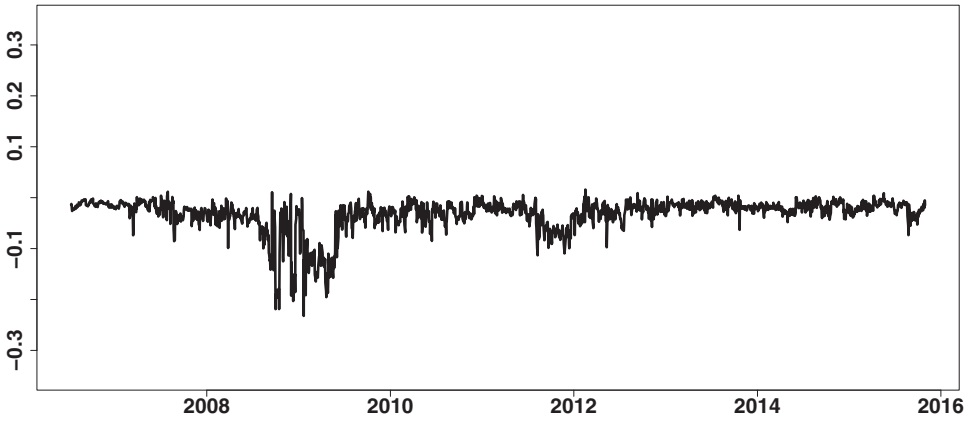


Figure 9. Log returns of JPM (gray) and VaR of log returns of JPM (black); $\tau = 0.05$, $T = 2335$, window size $n = 126$; refer to (5.2).

where $X_{i,t}$ represents the asset return of financial institution i at time t. Then the VaR of each firm with $F^{-1}_{\varepsilon_{i,t}}(\tau|M_{t-1}) = 0$ is obtained by

$$\widehat{\mathrm{VaR}}^{\tau}_{i,t} = \widehat{\alpha}_i + \widehat{\gamma}_i^{\top} M_{t-1}. \tag{5.2}$$

Now the second regression is performed using the proposed MACE method. The response variable is log returns of CYN, and the explanatory variables are potential risk factors that include the log returns of those covariates and the lagged macro variables:

$$X_{j,t} = g(S^{\top}\beta_{j|S}) + \varepsilon_{j,t}, \tag{5.3}$$

where $S \overset{\text{def}}{=} [M_{t-1}, R]$, R is a vector of log returns for different firms. $\beta_{j|S}$ is a $p \times 1$ vector. A detailed list of factors can be found in Tables A.2–A.4 in the Appendix.

With $F^{-1}_{\varepsilon_{j,t}}(\tau|S) = 0$ the CoVaR for firm j is estimated as

$$\widehat{\mathrm{CoVaR}}^{\tau}_{j|\widehat{S}} = \widehat{g}(\widehat{S}^{\top}\widehat{\beta}_{j|S}), \tag{5.4}$$

where $\widehat{S} \overset{\text{def}}{=} [M_{t-1}, \widehat{V}]$, with $\widehat{V}$ as the estimated VaR in (5.2).

To evaluate the preciseness of the proposed CoVaR risk measure, we launch a back-testing procedure. First, one calculates the violations over time, which is defined as the days on which the log returns are lower than the estimated VaR or CoVaR:

$$\hat{I}_{i,t} = \begin{cases} 1, & X_{i,t} < \widehat{\mathrm{VaR}}^{\tau}_{i,t}; \\ 0, & \text{otherwise}, \end{cases}$$

where theoretically $I_{i,t} - \tau$ should be a martingale difference sequence. Then we apply one version of the CaViaR test (see Berkowitz, Christoffersen, and Pelletier 2011), which adopts a logit model

$$I_{i,t} = \alpha + \beta_1 I_{i,t-1} + \beta_2 \mathrm{VaR}_{i,t} + u_{i,t},$$

where $u_{i,t}$ has a logistic distribution. The Wald test is then applied with null hypothesis: $\widehat{\beta}_1 = \widehat{\beta}_2 = 0$; see Franke, Härdle, and Hafner (2004) for more details.

5.2 Results

We use a moving window size of $n = 126$ (corresponding approximately to half a year of trading days) to calculate VaR of the log returns for the 199 firms, macro variables, and CYN. Figures 9 and 10 show one illustration of the estimated VaR of JPM (one covariate in the second step) and CYN, respectively. It can be seen that the estimated VaR traces the low values of returns closely, and becomes more volatile when the volatility of the returns is large.

With the VaR estimation in previous step, we show further the estimation of the CoVaR for CYN. The estimation

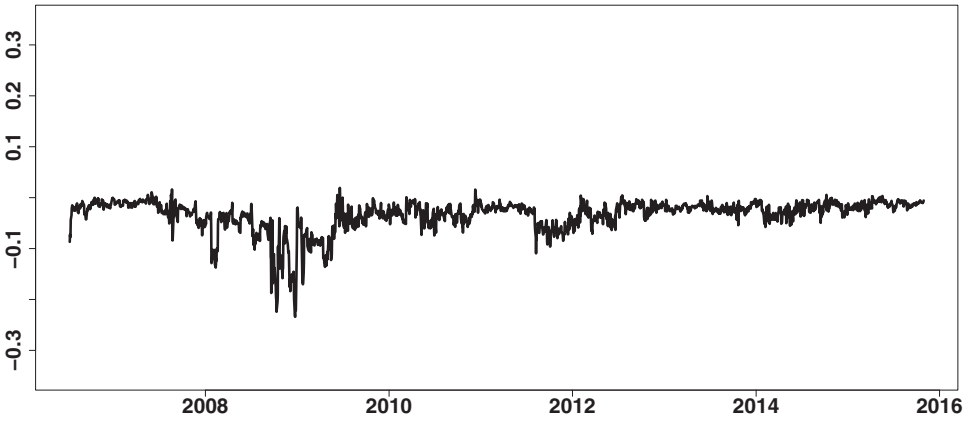


Figure 10. Log returns of CYN (gray) and VaR of log returns of CYN (black); $\tau = 0.05$, $T = 2335$, window size $n = 126$; refer to (5.2).

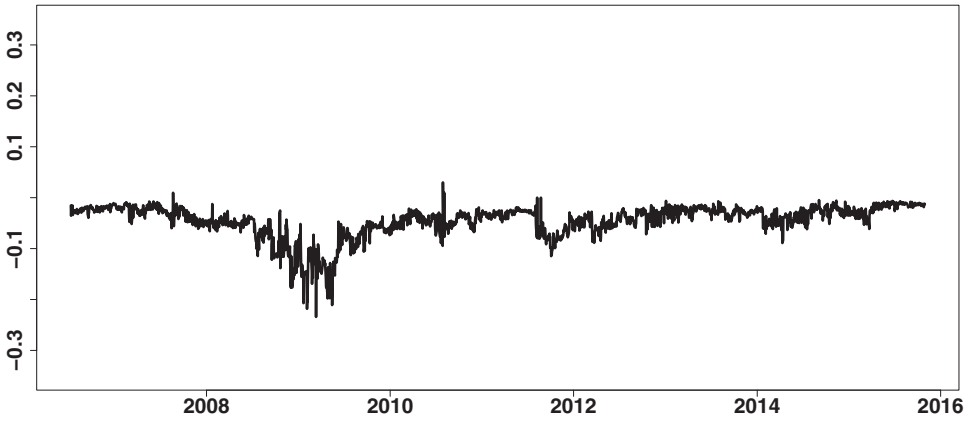


Figure 11. Log returns of CYN (gray) and the estimated CoVaR (black); $\tau = 0.05$, $T = 2335$, window size $n = 126$; refer to (5.4).

is conducted in a moving window of size 126. Our technique is applied with $\tau = 0.05$. We use $p = 206$ covariates, and the CoVaR for CYN is estimated with different variables selected in each window. Figure 11 shows the estimation results. We further summarize the selected variables in different windows.

Figure 12 summarizes the selection frequency of the firms and macro variables for all the windows. The variable 187, "Radian Group Inc. (RDN)," is the most frequently selected variable with frequency 752, which indicates the most relevant risk driver for CYN.

To compare the performance of our proposed measure with existing measures, we further apply CaViaR test for backtesting. Figure 13 shows the $\hat{I}_{i,t}$ sequence of $\widehat{\text{VaR}}$ (estimated value at risk measure) of CYN; there are a total of 23 violations. With $T = 2335$, the violation proportion is then $\widehat{\tau} = 0.009$.

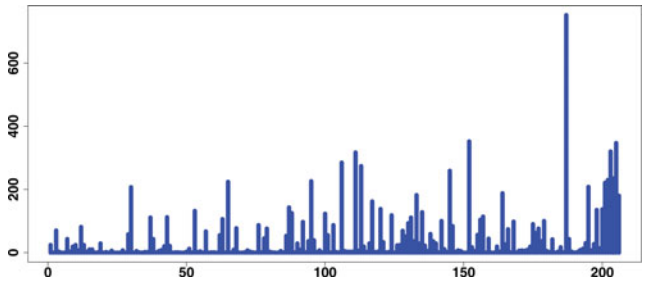


Figure 12. The frequency of the firms and macro variables. The X-axis: $1 - 206$ variables, and the Y-axis: the frequency of the variables selected in the moving window estimation. The variable 187, that is, "Radian Group Inc. (RDN)" is the most frequently selected variable with frequency 752.

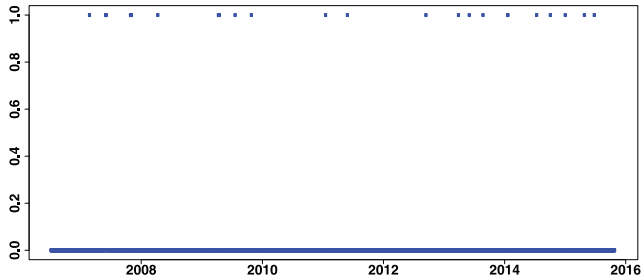


Figure 13. The violations (i.e., $\{t : \hat{I}_{i,t} = 1\}$) of $\widehat{\text{VaR}}$ for CYN(the dots above), in total 23 violations, $T = 2335$, $\widehat{\tau} = 0.009$.

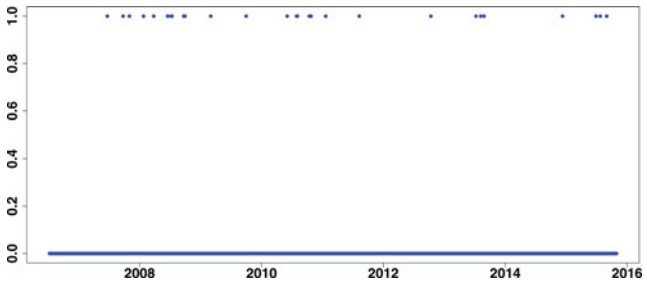


Figure 14. The violations (i.e., $\{t : \hat{I}_{i,t} = 1\}$) of $\widehat{\text{CoVaR}}_{\text{sim}}$ of CYN(the dots above), in total 28 violations, $T = 2335$, $\widehat{\tau} = 0.012$.

From Figure 14 we get the $\hat{I}_{i,t}$ sequence of $\widehat{\text{CoVaR}}$ of CYN; there are 28 violations out of $T = 2335$, which means $\widehat{\tau} = 0.012$.

The p-values of the CaViaR tests are then shown in Table 6, in which we compare our measure $\widehat{\text{CoVaR}}_{\text{SIM}}$ (CoVaR estimated from single-index model) with the measure attained solely by doing linear quantile variable selection, that is, $\widehat{\text{CoVaR}}_{\text{L}}$ (see, e.g., Belloni and Chernozhukov 2011). For the overall period, only for $\widehat{\text{CoVaR}}_{\text{SIM}}$, the null hypothesis can not be rejected. Therefore, $\widehat{\text{VaR}}$ and $\widehat{\text{CoVaR}}_{\text{L}}$ algorithms do not perform so well in an overall period. During crisis times, the null hypothesis of $\widehat{\text{VaR}}$ and $\widehat{\text{CoVaR}}_{SIM}$ cannot be rejected; therefore, both $\widehat{\text{VaR}}$ and $\widehat{\text{CoVaR}}_{\text{SIM}}$ algorithms perform well during the crisis periods, but $\widehat{\text{CoVaR}}_{\text{L}}$'s performance is not favorable.

APPENDIX

A.1 Proof

Proofs are available in the online supplementary materials.

Table A.1. Descriptive statistics of CYN

	Mean	SD	Skewness	Kurtosis
Overall period	−0.0001	0.0237	0.2821	14.0036
In crisis	-9.247×10^{-5}	0.0312	0.1326	8.9544

A.2 Application

The macro variables are the same as suggested by Adrian and Brunnermeier (2011) and Chao, Härdle, and Wang (2012). The macro variables and the corresponding source are listed as follows:

Table A.2. The financial firms

The financial firms	
1. Wells Fargo & Co (WFC)	15. Franklin Resources Inc. (BEN)
2. JP Morgan Chase & Co (JPM)	16. The Travelers Companies, Inc. (TRV)
3. Bank of America Corp (BAC)	17. AFLAC Inc. (AFL)
4. Citigroup Inc (C)	18. Prudential Financial, Inc. (PRU)
5. American Express Company (AXP)	19. State Street Corporation (STT)
6. U.S. Bancorp (USB)	20. The Chubb Corporation (CB)
7. The Goldman Sachs Group, Inc. (GS)	21. BB&T Corporation (BBT)
8. American International Group, Inc. (AIG)	22. Marsh & McLennan Companies, Inc. (MMC)
9. MetLife, Inc. (MET)	23. The Allstate Corporation (ALL)
10. Capital One Financial Corp. (COF)	24. Aon plc (AON)
11. BlackRock, Inc. (BLK)	25. CME Group Inc. (CME)
12. Morgan Stanley (MS)	26. The Charles Schwab Corporation (SCHW)
13. PNC Financial Services Group Inc. (PNC)	27. T. Rowe Price Group, Inc. (TROW)
14. The Bank of New York Mellon Corporation (BK)	28. Loews Corporation (L)
29. SunTrust Banks, Inc. (STI)	44. Lincoln National Corporation (LNC)
30. Fifth Third Bancorp (FITB)	45. Affiliated Managers Group Inc. (AMG)
31. Progressive Corp. (PGR)	46. Cincinnati Financial Corp. (CINF)
32. M&T Bank Corporation (MTB)	47. Equifax Inc. (EFX)
33. Ameriprise Financial Inc. (AMP)	48. Alleghany Corp. (Y)
34. Northern Trust Corporation (NTRS)	49. Unum Group (UNM)
35. Invesco Ltd. (IVZ)	50. Comerica Incorporated (CMA)
36. Moody's Corp. (MCO)	51. W.R. Berkley Corporation (WRB)
37. Regions Financial Corp. (RF)	52. Fidelity National Financial, Inc. (FNF)
38. The Hartford Financial Services Group, Inc. (HIG)	53. Huntington Bancshares Incorporated (HBAN)
39. TD Ameritrade Holding Corporation (AMTD)	54. Raymond James Financial Inc. (RJF)
40. Principal Financial Group Inc. (PFG)	55. Torchmark Corp. (TMK)
41. SLM Corporation (SLM)	56. Markel Corp. (MKL)
42. KeyCorp (KEY)	57. Ocwen Financial Corp. (OCN)
43. CNA Financial Corporation (CNA)	58. Arthur J Gallagher & Co. (AJG)

Table A.3. The financial firms

The financial firms	
59. Hudson City Bancorp, Inc. (HCBK)	74. Commerce Bancshares, Inc. (CBSH)
60. People's United Financial Inc. (PBCT)	75. Signature Bank (SBNY)
61. SEI Investments Co. (SEIC)	76. Jefferies Group, Inc. (JEF)
62. Nasdaq OMX Group Inc. (NDAQ)	77. Rollins Inc. (ROL)
63. Brown & Brown Inc. (BRO)	78. Morningstar Inc. (MORN)
64. BOK Financial Corporation (BOKF)	79. East West Bancorp, Inc. (EWBC)
65. Zions Bancorp. (ZION)	80. Waddell & Reed Financial Inc. (WDR)
66. HCC Insurance Holdings Inc. (HCC)	81. Old Republic International Corporation (ORI)
67. Eaton Vance Corp. (EV)	82. ProAssurance Corporation (PRA)
68. Erie Indemnity Company (ERIE)	83. Assurant Inc. (AIZ)
69. American Financial Group Inc. (AFG)	84. Hancock Holding Company (HBHC)
70. Dun & Bradstreet Corp. (DNB)	85. First Niagara Financial Group Inc. (FNFG)
71. White Mountains Insurance Group, Ltd. (WTM)	86. SVB Financial Group (SIVB)
72. Cullen-Frost Bankers, Inc. (CFR)	87. First Horizon National Corporation (FHN)
73. Legg Mason Inc. (LM)	88. E-TRADE Financial Corporation (ETFC)
89. SunTrust Banks, Inc. (STI)	104. Valley National Bancorp (VLY)
90. Mercury General Corporation (MCY)	105. KKR Financial Holdings LLC (KFN)
91. Associated Banc-Corp (ASBC)	106. Synovus Financial Corporation (SNV)
92. Credit Acceptance Corp. (CACC)	107. Texas Capital BancShares Inc. (TCBI)
93. Protective Life Corporation (PL)	108. American National Insurance Co. (ANAT)
94. Federated Investors, Inc. (FII)	109. Washington Federal Inc. (WAFD)
95. CNO Financial Group, Inc. (CNO)	110. First Citizens Bancshares Inc. (FCNCA)
96. Popular, Inc. (BPOP)	111. Kemper Corporation (KMPR)
97. Bank of Hawaii Corporation (BOH)	112. UMB Financial Corporation (UMBF)
98. Fulton Financial Corporation (FULT)	113. Stifel Financial Corp. (SF)
99. AllianceBernstein Holding L.P. (AB)	114. CapitalSource Inc. (CSE)
100. TCF Financial Corporation (TCB)	115. Portfolio Recovery Associates Inc. (PRAA)
101. Susquehanna Bancshares, Inc. (SUSQ)	116. Janus Capital Group, Inc. (JNS)
102. Capitol Federal Financial, Inc. (CFFN)	117. MBIA Inc. (MBI)
103. Webster Financial Corp. (WBS)	118. Healthcare Services Group Inc. (HCSG)

Table A.4. The financial firms

The financial firms	
119. The Hanover Insurance Group Inc. (THG)	134. BancorpSouth, Inc. (BXS)
120. F.N.B. Corporation (FNB)	135. Privatebancorp Inc. (PVTB)
121. FirstMerit Corporation (FMER)	136. United Bankshares Inc. (UBSI)
122. FirstMerit Corporation (FMER)	137. Old National Bancorp. (ONB)
123. RLI Corp. (RLI)	138. International Bancshares Corporation (IBOC)
124. StanCorp Financial Group Inc. (SFG)	139. First Financial Bankshares Inc. (FFIN)
125. Trustmark Corporation (TRMK)	140. Westamerica Bancorp. (WABC)
126. IberiaBank Corp. (IBKC)	141. Northwest Bancshares, Inc. (NWBI)
127. Cathay General Bancorp (CATY)	142. Bank of the Ozarks, Inc. (OZRK)
128. National Penn Bancshares Inc. (NPBC)	143. Huntington Bancshares Incorporated (HBAN)
129. Nelnet, Inc. (NNI)	144. Euronet Worldwide Inc. (EEFT)
130. Wintrust Financial Corporation (WTFC)	145. Community Bank System Inc. (CBU)
131. Umpqua Holdings Corporation (UMPQ)	146. CVB Financial Corp. (CVBF)
132. GAMCO Investors, Inc. (GBL)	147. MB Financial Inc. (MBFI)
133. Sterling Financial Corp. (STSA)	148. ABM Industries Incorporated (ABM)
149. Glacier Bancorp Inc. (GBCI)	164. Citizens Republic Bancorp, Inc (CRBC)
150. Selective Insurance Group Inc. (SIGI)	165. Horace Mann Educators Corp. (HMN)
151. Park National Corp. (PRK)	166. DFC Global Corp. (DLLR)
152. Flagstar Bancorp Inc. (FBC)	167. Navigators Group Inc. (NAVG)
153. FBL Financial Group Inc. (FFG)	168. Boston Private Financial Holdings, Inc. (BPFH)
154. Astoria Financial Corporation (AF)	169. American Equity Investment Life Holding Co. (AEL)
155. World Acceptance Corp. (WRLD)	170. BlackRock Limited Duration Income Trust (BLW)
156. First Midwest Bancorp Inc. (FMBI)	171. Columbia Banking System Inc. (COLB)
157. PacWest Bancorp (PACW))	172. Safety Insurance Group Inc. (SAFT)
158. First Financial Bancorp. (FFBC)	173. National Financial Partners Corp. (NFP)
159. BBCN Bancorp, Inc. (BBCN)	174. NBT Bancorp, Inc. (NBTB)
160. Provident Financial Services, Inc. (PFS)	175. Tower Group Inc. (TWGP)
161. FBL Financial Group Inc. (FFG)	176. Encore Capital Group, Inc. (ECPG)
162. WisdomTree Investments, Inc. (WETF)	177. Pinnacle Financial Partners Inc. (PNFP)
163. Hilltop Holdings Inc. (HTH)	178. First Commonwealth Financial Corp. (FCF)

(*Continued on next column*)

Table A.4. The financial firms (*Continued*)

The financial firms	
179. BancFirst Corporation (BANF)	190. Berkshire Hills Bancorp Inc. (BHLB)
180. Independent Bank Corp. (INDB)	191. Brookline Bancorp, Inc. (BRKL)
181. Infinity Property and Casualty Corp. (IPCC)	192. National Western Life Insurance Company (NWLI)
182. Central Pacific Financial Corp. (CPF)	193. Tompkins Financial Corporation (TMP)
183. Kearny Financial Corp. (KRNY)	194. BGC Partners, Inc. (BGCP)
184. Chemical Financial Corporation (CHFC)	195. Epoch Investment Partners, Inc. (EPHC)
185. Banner Corporation (BANR)	196. United Fire Group, Inc (UFCS)
186. State Auto Financial Corp. (STFC)	197. 1st Source Corporation (SRCE)
187. Radian Group Inc. (RDN)	198. Citizens Inc. (CIA)
188. SCBT Financial Corporation (SCBT)	199. S&T Bancorp Inc. (STBA)
189. WesBanco Inc. (WSBC)	

1. VIX, which measures the implied volatility in the market.
2. The short-term liquidity spread, which is calculated by the difference between the 3-month Treasury repo rate and 3-month Treasury constant maturities.
3. The daily change in the 3-month Treasury constant maturities, which can be defined as the difference between the current day and the previous day of 3-month Treasury constant maturities.
4. The change in the slope of the yield curve, which is defined by the difference between the 10-year Treasury constant maturities and the 3-month Treasury constant maturities.
5. The change in the credit spread between 10-year BAA corporate bonds and the 10-year Treasury constant maturities.
6. The daily S&P500 index returns.
7. The daily Dow Jones U.S. Real Estate index returns.

The repo data can be obtained from the Datastream database, and the 10-year Treasury constant maturities and BAA corporate bonds data can be found in the website of the Federal Reserve Board H.15: *http://www.federalreserve.gov/releases/h15/data.htm*. Other data are available in Yahoo Finance. The macro variables' data are available from January 4, 2006, to October 29, 2015, with a daily frequency.

Tables A.1 shows the descriptive statistics of this series. The mean of CYN in the the overall period (i.e., January 6, 2006, to October 30, 2015) is -0.000118, which is higher than that (-0.000092) in the crisis period (i.e., from September 15, 2008, to February 8, 2010). The volatility in the crisis period is higher than it in the overall period. The p-values of the Jarque Bera test indicate that log returns of CYN are not normally distributed. We also perform a unit root test, which suggests that the log returns of CYN are stationary. The mentioned two test results for the other firms show that all these series are not normally distributed, but are likely to be stationary.

SUPPLEMENTARY MATERIALS

In the supplementary materials we provide proofs for theorems in Section 3.

ACKNOWLEDGMENTS

We thank the editor, the associate editor, and the referees for valuable comments. The financial support from the Deutsche Forschungsgemeinschaft via CRC 649 "Ökonomisches Risiko," Humboldt-Universität zu Berlin, IRTG 1792, and the Research Grants Council of Hong Kong via G-HK012/10 is gratefully acknowledged. We also gratefully acknowledge the funding from DAAD ID 50746311. Y. Fan would like to acknowledge the National Natural Science Foundation of China (Grant No. 11501354).

[Received May 2014. Revised April 2016.]

REFERENCES

Adrian, T., and Brunnermeier, M. K. (2011), "CoVaR," Technical Report, National Bureau of Economic Research. [1,13]

Belloni, A., and Chernozhukov, V. (2011), "L1-Penalized Quantile Regression in High-Dimensional Sparse Models," *The Annals of Statistics*, 39, 82–130. [12]

Berkowitz, J., Christoffersen, P., and Pelletier, D. (2011), "Evaluating Value-at-Risk Models With Desk-Level Data," *Management Science*, 57, 2213–2227. [11]

Bradic, J., Fan, J., and Wang, W. (2011), "Penalized Composite Quasi-Likelihood for Ultrahigh Dimensional Variable Selection," *Journal of the Royal Statistical Society*, Series B, 73, 325–349. [2,3]

Chao, S.-K., Härdle, W. K., and Wang, W. (2012), "Quantile Regression in Risk Calibration," Technical Report, SFB 649 discussion paper, Humboldt University, Berlin. [13]

Engle, R. F., and Manganelli, S. (2004), "CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles," *Journal of Business & Economic Statistics*, 22, 367–381. [2]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [3]

Fan, J., and Lv, J. (2010), "A Selective Overview of Variable Selection in High Dimensional Feature Space," *Statistica Sinica*, 20, 101. [4]

Franke, J., Härdle, W., and Hafner, C. M. (2004), *Statistics of Financial Markets* (Vol. 2), Berlin: Springer. [11]

Härdle, W., and Stoker, T. M. (1989), "Investigating Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American Statistical Association*, 84, 986–995. [2]

Hautsch, N., Schaumburg, J., and Schienle, M. (2014), "Financial Network Systemic Risk Contributions," *Review of Finance*, doi: 10.1093/rof/rfu010 [1]

Huber, P. J. (1985), "Projection Pursuit," *The Annals of Statistics*, 13, 435–475. [2]

Kai, B., Li, R., and Zou, H. (2010), "Local Composite Quantile Regression Smoothing: An Efficient and Safe Alternative to Local Polynomial Regression," *Journal of the Royal Statistical Society*, Series B, 72, 49–69. [2]

——— (2011), "New Efficient Estimation and Variable Selection Methods for Semiparametric Varying-Coefficient Partially Linear Models," *Annals of Statistics*, 39, 305. [2]

Kong, E., and Xia, Y. (2007), "Variable Selection for the Single-Index Model," *Biometrika*, 94, 217–229. [2]

——— (2012), "A Single-Index Quantile Regression Model and Its Estimation," *Econometric Theory*, 28, 730–768. [2]

Leng, C., Xia, Y., and Xu, J. (2008), "An Adaptive Estimation Method for Semiparametric Models and Dimension Reduction," in *WSPC-Proceedings*, pp. 1–14. [3]

Li, Y., and Zhu, J. (2008), "L1-Norm Quantile Regression," *Journal of Computational and Graphical Statistics*, 17, 163–185. [5]

Newey, W. K., and Powell, J. L. (1987), "Asymmetric Least Squares Estimation and Testing," *Econometrica*, 55, 819–847. [2]

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995), "An Effective Bandwidth Selector for Local Least Squares Regression," *Journal of the American Statistical Association*, 90, 1257–1270. [5]

Schnabel, S. K., and Eilers, P. H. (2009), "Optimal Expectile Smoothing," *Computational Statistics & Data Analysis*, 53, 4168–4177. [2]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [3]

Wang, H. J., Li, D., and He, X. (2012), "Estimation of High Conditional Quantiles for Heavy-Tailed Distributions," *Journal of the American Statistical Association*, 107, 1453–1464. [4]

Wang, Q., and Yin, X. (2008), "A Nonlinear Multi-Dimensional Variable Selection Method for High Dimensional Data: Sparse MAVE," *Computational Statistics & Data Analysis*, 52, 4512–4520. [2]

Wu, T. Z., Yu, K., and Yu, Y. (2010), "Single-Index Quantile Regression," *Journal of Multivariate Analysis*, 101, 1607–1621. [2]

Xia, Y., Tong, H., Li, W., and Zhu, L.-X. (2002), "An Adaptive Estimation of Dimension Reduction Space," *Journal of the Royal Statistical Society*, Series B, 64, 363–410. [2]

Yu, K., and Jones, M. (1998), "Local Linear Quantile Regression," *Journal of the American Statistical Association*, 93, 228–237. [3,5]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society*, Series B, 68, 49–67. [3]

Zeng, P., He, T., and Zhu, Y. (2012), "A Lasso-Type Approach for Estimation and Variable Selection in Single Index Models," *Journal of Computational and Graphical Statistics*, 21, 92–109. [2]

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [3]

Zou, H., and Yuan, M. (2008), "Composite Quantile Regression and the Oracle Model Selection Theory," *The Annals of Statistics*, 36, 1108–1126. [2]

Journal of Multivariate Analysis 162 (2017) 1–15

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Statistical inference for generalized additive partially linear models

Rong Liu [a,*], Wolfgang K. Härdle [b,c], Guoyi Zhang [d]

[a] Department of Mathematics and Statistics, University of Toledo, Toledo, OH, United States
[b] Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Germany
[c] School of Business, Singapore Management University, Singapore
[d] Department of Mathematics and Statistics, The University of New Mexico, NM, United States

ARTICLE INFO

Article history:
Received 23 September 2016
Available online 18 August 2017

Keywords:
B-spline
Default
Empirical likelihood
Kernel estimator
Link function
Mixing

ABSTRACT

The class of Generalized Additive Models (GAMs) is a powerful tool which has been well studied. It helps to identify additive regression structure that can be determined even more sharply via test procedures when some component functions have a parametric form. Generalized Additive Partially Linear Models (GAPLMs) enjoy the simplicity of GLMs and the flexibility of GAMs because they combine both parametric and nonparametric components. We use the hybrid spline-backfitted kernel estimation method, which combines the best features of both spline and kernel methods, to make fast, efficient and reliable estimation under an α-mixing condition. In addition, simultaneous confidence corridors (SCCs) for testing overall trends and empirical likelihood confidence regions for parameters are provided under an independence condition. The asymptotic properties are obtained and simulation results support the theoretical properties. As an illustration, we use GAPLM methodology to improve the accuracy ratio of the default predictions for 19,610 German companies. The quantlet for this paper are available on https://github.com.

1. Introduction

The class of Generalized Additive Models (GAMs) provides an effective semiparametric regression tool for high-dimensional data; see [6]. For a response Y and a predictor vector $\mathbf{X} = (X_1, \ldots, X_d)^\top$, the pdf of Y_i conditional on $\mathbf{X}_i$ with respect to a fixed σ-finite measure is from an exponential family, viz.

$$f(Y_i \mid \mathbf{X}_i, \phi) = \exp\left[\{Y_i m(\mathbf{X}_i) - b\{m(\mathbf{X}_i)\}\}/a(\phi) + h(Y_i, \phi)\right].$$

The function b is a given function which relates $m(\mathbf{x})$ to the conditional variance function $\sigma^2(\mathbf{x}) = \mathrm{var}(Y \mid \mathbf{X} = \mathbf{x})$ via the equation $\sigma^2(\mathbf{x}) = a(\phi) b''\{m(\mathbf{x})\}$, in which $a(\phi)$ is a nuisance parameter that quantifies overdispersion. For theoretical developments, it is not necessary to assume that the data $(Y_1, \mathbf{X}_1^\top), \ldots, (Y_n, \mathbf{X}_n^\top)$ come from such an exponential family, but only that the conditional mean and variance are linked by the relation

$$\mathrm{var}(Y \mid \mathbf{X} = \mathbf{x}) = a(\phi) b''[(b')^{-1}\{\mathrm{E}(Y \mid \mathbf{X} = \mathbf{x})\}].$$

* Corresponding author.
E-mail addresses: rong.liu@utoledo.edu (R. Liu), haerdle@wiwi.hu-berlin.de (W.K. Härdle), gzhang@unm.edu (G. Zhang).

http://dx.doi.org/10.1016/j.jmva.2017.07.011
0047-259X/

More specifically, the model is

$$\mathrm{E}\,(Y \mid \mathbf{X}) = b'\left\{c + \sum_{\alpha=1}^{d} m_\alpha(X_\alpha)\right\}, \tag{1}$$

where b' is the derivative of function b. Model ((1)) can be used, e.g., in scoring methods and analyzing default of companies; here $Y = 1$ denotes default and $b' = e^y/1 + e^y$ is the link function. Fitting Model (1) to such a default data set leads to estimated component functions $\hat{m}_1, \ldots, \hat{m}_d$; see, e.g., [11,25]. Plotting these functions with simultaneous confidence corridors (SCCs) as developed by [25], one can check the functional form and therefore obtain simpler parameterizations of $m_1, \ldots, m_d$.

The typical approach is to perform a preliminary (nonparametric) analysis on the influence of the component functions, and one may improve the model by introducing parametric components. This will lead to simplification, more interpretability and higher precision in statistical calibration. With these thoughts in mind, GAMs can be extended to Generalized Additive Partially Linear Models (GAPLM), in which

$$\mathrm{E}\,(Y \mid \mathbf{T}, \mathbf{X}) = b'\,\{m\,(\mathbf{T}, \mathbf{X})\}, \tag{2}$$

with $m\,(\mathbf{T}, \mathbf{X}) = \boldsymbol{\beta}^\top \mathbf{T} + \sum_{\alpha=1}^{d_2} m_\alpha(X_\alpha)$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{d_1})^\top$, $\mathbf{T} = \left(T_0, \ldots, T_{d_1}\right)^\top$, and $\mathbf{X} = \left(X_1, \ldots, X_{d_2}\right)^\top$, where $T_0 = 1$ and $T_k \in \mathbb{R}$ for all $k \in \{1, \ldots, d_1\}$. In this paper, we assume that

$$\mathrm{var}\,(Y \mid \mathbf{T} = \mathbf{t}, \mathbf{X} = \mathbf{x}) = a\,(\phi)\,b''[\left(b'\right)^{-1}\{\mathrm{E}\,(Y \mid \mathbf{T} = \mathbf{t}, \mathbf{X} = \mathbf{x})\}].$$

We can write (2) in the usual regression form $Y_i = b'\,\{m\,(\mathbf{T}_i, \mathbf{X}_i)\} + \sigma\,(\mathbf{T}_i, \mathbf{X}_i)\,\varepsilon_i$ with white noise ε_i that satisfies $\mathrm{E}\,(\varepsilon_i \mid \mathbf{T}_i, \mathbf{X}_i) = 0$, $\mathrm{E}(\varepsilon_i^2 \mid \mathbf{T}_i, \mathbf{X}_i) = 1$. For identifiability, we impose the condition

$$\forall_{\alpha \in \{1,\ldots,d_2\}} \quad \mathrm{E}\,\{m_\alpha(X_\alpha)\} = 0. \tag{3}$$

As in most works on nonparametric smoothing, estimation of the functions $m_1, \ldots, m_{d_2}$ is conducted on compact sets. Without loss of generality, let the compact set be $\varkappa = [0, 1]^{d_2}$.

Some estimation methods for Model (2) have been proposed, but are either computationally expensive or lacking theoretical justification. The kernel-based backfitting and marginal integration methods, e.g., in [5,9,24], are computationally expensive. More advanced non- and semi-parametric models (without link function) have also been studied, e.g., partially linear models and varying-coefficient models; see [10,14,16,20,23]. In [20], a nonconcave penalized quasi-likelihood method was proposed with polynomial spline smoothing for estimation of $m_1, \ldots, m_{d_2}$, and deriving quasi-likelihood based estimators for the linear parameter $\boldsymbol{\beta} \in \mathbb{R}^{1+d_1}$.

To our knowledge, [20] is a pilot paper since it establishes the asymptotic normality of the estimators for the parametric components in GAPLMs with independent observations. However, the asymptotic normality of the estimators of the nonparametric component functions $m_1, \ldots, m_{d_2}$ and SCCs remains to be proved. Recently, [12] studied more complicated Generalized Additive Coefficient Models by using a two-step spline method, but an iid assumption is required for the asymptotic properties of the estimation and inference of m_α, and the asymptotic normality of parameter estimates has not been shown either. Nonparametric analysis of deviance tools was developed in [4], which can be used to test the significance of the nonparametric term in generalized partially linear models with univariate nonparametric component function. Empirical likelihood based confidence regions for the parameter $\boldsymbol{\beta}$ and point-wise confidence intervals for the nonparametric term in generalized partially linear models were also provided in [8].

The spline-backfitted kernel (SBK) estimation introduced in [21] combines the advantages of both kernel and spline methods and the result is balanced in terms of theory, computation, and interpretation. The basic idea is to pre-smooth the component functions by spline estimation and then use the kernel method to improve the accuracy of the estimation on a specific m_α. In this paper, we extend the SBK method to calibrate Model (2) with additive nonparametric components and as a result, we obtain oracle efficiency and asymptotic normality of the estimators for both the parametric and nonparametric components under α- mixing condition, which complicates the derivation of theoretical properties. With the stronger iid assumption, we provide an empirical likelihood (EL) based confidence region for the parameter $\boldsymbol{\beta}$ due to the advantages of EL such as increase in coverage accuracy, easy implementation, avoiding estimating variances and Studentizing automatically; see [8]. In addition, we provide SCCs for the nonparametric component functions based on the maximal deviation distribution in [2], so that one can test the hypothesis of the shape for nonparametric terms.

The paper is organized as follows. In Section 2, we discuss the details of (2). In Section 3, the oracle estimator and its asymptotic properties are introduced. In Section 4, the SBK estimator is introduced and the asymptotics for both the parametric and nonparametric component estimations is given. In addition, SCCs for testing overall trends and entire shapes are considered. In Section 5, we apply the methods to simulated and real data examples. All technical proofs are given in Appendix.

2. Model assumptions

The space of α-centered square integrable functions on $[0, 1]$ is defined as in [18], viz.

$$\mathcal{H}_\alpha^0 = \{g : \mathrm{E}\{g(X_\alpha)\} = 0, \mathrm{E}\{g^2(X_\alpha)\} < \infty\}.$$

Next define the model space $\mathcal{M}$, a collection of functions on $\mathbb{R}^{d_2}$ as

$$\mathcal{M} = \left\{ g(\mathbf{x}) = \sum_{\alpha=1}^{d_2} g_\alpha(\mathbf{x}) : g_\alpha \in \mathcal{H}_\alpha^0 \right\}.$$

The constraints that $\mathrm{E}\{g_\alpha(X_\alpha)\} = 0$ for all $\alpha \in \{1, \ldots, d_2\}$ ensure the unique additive representation of m_α as expressed in (3). Denote the empirical expectation by E_n, i.e., $\mathrm{E}_n(\varphi) = \sum_{i=1}^n \varphi(\mathbf{X}_i)/n$. For functions $g_1, g_2 \in \mathcal{M}$, the theoretical and empirical inner products are defined respectively as $\langle g_1, g_2\rangle = \mathrm{E}\{g_1(\mathbf{X}) g_2(\mathbf{X})\}$, $\langle g_1, g_2\rangle_n = \mathrm{E}_n\{g_1(\mathbf{X}) g_2(\mathbf{X})\}$. The corresponding induced norms are $\|g_1\|_2^2 = \mathrm{E}\{g_1^2(\mathbf{X})\}$, $\|g_1\|_{2,n}^2 = \mathrm{E}_n\{g_1^2(\mathbf{X})\}$. More generally, we set $\|g\|_r^r = \mathrm{E}|\{g(\mathbf{X})|^r\}$.

In the paper, for any compact interval $[a, b]$, we denote the space of pth order smooth functions as $C^{(p)}[a, b] = \{g : g^{(p)} \in C[a, b]\}$, and the class of Lipschitz continuous functions for constant $C > 0$ as

$$\mathrm{Lip}([a, b], C) = \{g : \forall_{x,x' \in [a,b]} |g(x) - g(x')| \le C|x - x'|\}.$$

For any vector $\mathbf{x} = (x_1, \ldots, x_d)^\top$, we denote the supremum and p norm as $|\mathbf{x}| = \max_{1\le\alpha\le d}|x_\alpha|$ and $\|\mathbf{x}\|_p = (\sum_{\alpha=1}^d x_\alpha^p)^{1/p}$, respectively. In particular, we use $\|\mathbf{x}\|$ to denote the Euclidean norm, i.e., $p = 2$. We need the following assumptions.

(A1) For every $\alpha \in \{1, \ldots, d_2\}$, one has $m_\alpha \in C^{(1)}[0, 1]$; furthermore, $m_1 \in C^{(2)}[0, 1]$ and there exists a constant $C_m > 0$ such that, for all $\alpha \in \{2, \ldots, d_2\}$, $m'_\alpha \in \mathrm{Lip}([0, 1], C_m)$.

(A2) The inverse link function b' satisfies $b' \in C^2(\mathbb{R})$, $b''(\theta) > 0$, $\theta \in \mathbb{R}$ and $C_b > \max_{\theta\in\Theta} b''(\theta) \ge \min_{\theta\in\Theta} b''(\theta) > c_b$ for constants $C_b > c_b > 0$.

(A3) The conditional variance function $\sigma^2(\mathbf{x})$ is measurable and bounded. The errors $\epsilon_1, \ldots, \epsilon_n$ are such that $\mathrm{E}(\varepsilon_i \mid \mathcal{F}_i) = 0$, $\mathrm{E}(|\varepsilon_i|^{2+\eta}) \le C_\eta$ for some $\eta \in (1/2, \infty)$ with the sequence of σ-fields: $\mathcal{F}_i = \sigma\{(\mathbf{X}_j) : j \le i, \varepsilon_j, j \le i - 1\}$ for all $i \in \{1, \ldots, n\}$.

(A4) The density function f of $(X_1, \ldots, X_{d_2})$ is continuous and $0 < c_f \le \inf_{\mathbf{x}\in\chi} f(\mathbf{x}) \le \sup_{\mathbf{x}\in\chi} f(\mathbf{x}) \le C_f < \infty$. The marginal densities f_α of X_α have continuous derivatives on $[0, 1]$ and are uniformly bounded from above by C_f and from below by c_f.

(A5) There exist constants $K_0, \lambda_0 \in (0, +\infty)$ such that $\alpha(n) \le K_0 e^{-\lambda_0 n}$ holds for all $n \in \mathbb{N}$, with the α-mixing coefficients for the sequence $\mathbf{Z}_1 = (\mathbf{T}_1^\top, \mathbf{X}_1^\top, \varepsilon_1)^\top, \ldots, \mathbf{Z}_n = (\mathbf{T}_n^\top, \mathbf{X}_n^\top, \varepsilon_i)^\top$ defined, for every integer $k \ge 1$, by

$$\alpha(k) = \sup_{B\in\sigma\{\mathbf{Z}_s, s\le t\}, C\in\sigma\{\mathbf{Z}_s, s\ge t+k\}} |\Pr(B \cap C) - \Pr(B)\Pr(C)|.$$

(A5') The variables $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are mutually independent and identically distributed.

(A6) There exist constants $0 < c_\delta < C_\delta < \infty$ and $0 < c_{\mathbf{Q}} < C_{\mathbf{Q}} < \infty$ such that $c_\delta \le \mathrm{E}(|T_k|^{2+\delta} \mid \mathbf{X} = \mathbf{x}) \le C_\delta$ for some $\delta > 0$, and $c_{\mathbf{Q}} I_{d_1\times d_1} \le \mathrm{E}(\mathbf{T}\mathbf{T}^\top \mid \mathbf{X} = \mathbf{x}) \le C_{\mathbf{Q}} I_{d_1\times d_1}$.

Assumptions (A1), (A2) and (A4) are standard in the GAM literature; see [19,22]. Assumptions (A3) and (A5) are the same for weakly dependent data as in [11,21], and Assumption (A6) is the same with (C5) in [20]. When categorical predictors are present, we can create dummy variables in $\mathbf{T}_i$ and Assumption (A6) is still satisfied.

3. Oracle estimators

The aim of our analysis is to provide precise estimators for the component functions m_α and parameters $\boldsymbol{\beta}$. Without loss of generality, we may focus on m_1. If all the unknown $\boldsymbol{\beta}$ and other $m_2, \ldots, m_{d_2}$ were known, we are in a comfortable situation since the multidimensional modeling problem has reduced to one dimension. As in [17] define, for each $x_1 \in [h, 1 - h]$ and $a \in A$, a local quasi log-likelihood function

$$\tilde{\ell}_{m_1}(a, x_1) = \frac{1}{n}\sum_{i=1}^n [Y_i\{a + m(\mathbf{T}_i, \mathbf{X}_{i\_1})\} - b\{a + m(\mathbf{T}_i, \mathbf{X}_{i\_1})\}] K_h(X_{i1} - x_1)$$

with $m(\mathbf{T}_i, \mathbf{X}_{i\_1}) = \boldsymbol{\beta}^\top \mathbf{T}_i + \sum_{\alpha=2}^{d_2} m_\alpha(\mathbf{X}_{i\alpha})$ and $K_h(u) = K(u/h)/h$ a kernel function K with bandwidth h satisfying the following condition.

(A7) The kernel function $K \in C^1[-1, 1]$ is a symmetric pdf and $h = h_n$ satisfies $h = \mathcal{O}\{n^{-1/5}(\ln n)^{-1/5}\}$, $h^{-1} = \mathcal{O}\{n^{1/5}(\ln n)^\delta\}$ for some constant $\delta > 1/5$.

Since all the $\boldsymbol{\beta}$ and $m_2, \ldots, m_{d_2}$ are known as obtained from the oracle, one can obtain the so-called oracle estimator

$$\tilde{m}_{K,1}(x_1) = \text{argmax}_{a\in A}\tilde{\ell}_{m_1}(a, x_1). \tag{4}$$

Denote $\|K\|_2^2 = \int K^2(u)\,du$, $\mu_2(K) = \int K(u)\,u^2 du$ and introduce the scale function

$$D_1(x_1) = f_1(x_1)\mathrm{E}\left\{b''\{m(\mathbf{T},\mathbf{X})\} \mid X_1 = x_1\right\}, \tag{5}$$

and the bias function

$$\begin{aligned}\text{bias}_1(x_1) = \mu_2(K)\Big[&m_1''(x_1)f_1(x_1)\mathrm{E}\left[b''\{m(\mathbf{T},\mathbf{X})\} \mid X_1 = x_1\right] \\ &+ m_1'(x_1)\frac{\partial}{\partial x_1}\left\{f_1(x_1)\mathrm{E}\left[b''\{m(\mathbf{T},\mathbf{X})\} \mid X_1 = x_1\right]\right\} \\ &- \left\{m_1'(x_1)\right\}^2 f_1(x_1)\mathrm{E}\left[b'''\{m(\mathbf{T},\mathbf{X})\} \mid X_1 = x_1\right]\Big].\end{aligned} \tag{6}$$

Lemma 1. *Under Assumptions* (A1)–(A7), *for any* $x_1 \in [h, 1-h]$, *as* $n \to \infty$, *the oracle kernel estimator* $\tilde{m}_{K,1}(x_1)$ *given in* (4) *satisfies*

$$\sup_{x_1\in[h,1-h]}|\tilde{m}_{K,1}(x_1) - m_1(x_1)| = \mathcal{O}_{a.s.}(\ln n/\sqrt{nh}),$$

$$\sqrt{nh}\{\tilde{m}_{K,1}(x_1) - m_1(x_1) - \text{bias}_1(x_1)h^2/D_1(x_1)\} \rightsquigarrow \mathcal{N}[0, D_1(x_1)^{-1}v_1^2(x_1)D_1(x_1)^{-1}],$$

with $v_1^2(x_1) = f_1(x_1)E\{\sigma^2(\mathbf{T},\mathbf{X}) \mid X_1 = x_1\}\|K\|_2^2$.

Lemma 1 is proved in [11]. The above oracle idea applies to the parametric part as well. Define the log-likelihood function

$$\tilde{\ell}_{\boldsymbol{\beta}}(\mathbf{a}) = \frac{1}{n}\sum_{i=1}^{n}[Y_i\{\mathbf{a}^\top\mathbf{T}_i + m(\mathbf{X}_i)\} - b\{\mathbf{a}^\top\mathbf{T}_i + m(\mathbf{X}_i)\}], \tag{7}$$

where $m(\mathbf{X}_i) = \sum_{\alpha=1}^{d_2} m_\alpha(X_{i\alpha})$. The infeasible estimator of $\boldsymbol{\beta}$ is $\tilde{\boldsymbol{\beta}} = \text{argmax}_{\mathbf{a}\in\mathbb{R}^{1+d_1}}\tilde{\ell}_{\boldsymbol{\beta}}(\mathbf{a})$. Clearly, $\nabla\tilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = \mathbf{0}$. To maximize (7), we have

$$\frac{1}{n}\sum_{i=1}^{n}[Y_i\mathbf{T}_i - b'\{\mathbf{a}^\top\mathbf{T}_i + m(\mathbf{X}_i)\}\mathbf{T}_i] = \mathbf{0},$$

then the empirical likelihood ratio is

$$\tilde{R}(\mathbf{a}) = \max\left\{\prod_{i=1}^{n} np_i : \sum_{i=1}^{n} p_i Z_i(\mathbf{a}) = \mathbf{0}, p_i \geq 0, \sum_{i=1}^{n} p_i = 1\right\}$$

where $Z_i(\mathbf{a}) = \left[Y_i - b'\left\{\mathbf{a}^\top\mathbf{T}_i + m(\mathbf{X}_i)\right\}\right]\mathbf{T}_i$.

Theorem 1. (i) *Under Assumptions* (A1)–(A6), *as* $n \to \infty$,

$$\left|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} - [Eb''\{m(\mathbf{T},\mathbf{X})\}\mathbf{T}\mathbf{T}^\top]^{-1}\frac{1}{n}\sum_{i=1}^{n}\sigma(\mathbf{T}_i,\mathbf{X}_i)\varepsilon_i\mathbf{T}_i\right| = \mathcal{O}_{a.s.}\{(\ln n)^2/n\},$$

$$\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightsquigarrow \mathcal{N}\left[\mathbf{0}, a(\phi)[Eb''\{m(\mathbf{T},\mathbf{X})\}\mathbf{T}\mathbf{T}^\top]^{-1}\right].$$

(ii) *Under Assumptions* (A1)–(A4), (A5') *and* (A6), $-2\ln\{\tilde{R}(\boldsymbol{\beta})\} \rightsquigarrow \chi^2_{d_1}$.

Although the oracle estimators $\tilde{\boldsymbol{\beta}}$ and $\tilde{m}_{K,1}(x_1)$ enjoy the desirable theoretical properties in Theorem 1 and Lemma 1, they are not feasible statistics as their computation is based on the knowledge of unavailable component functions $m_2, \ldots, m_{d_2}$.

4. Spline-backfitted kernel estimators

In practice, $m_2, \ldots, m_{d_2}$ are of course unknown and need to be approximated. We obtain the spline-backfitted kernel estimators by using estimations of $m_2, \ldots, m_{d_2}$ and the unknown $\boldsymbol{\beta}$ by splines and we employ them to estimate $m_1(x_1)$ as in (4). First, we introduce the linear spline basis as in [10]. Let $0 = \xi_0 < \xi_1 < \cdots < \xi_N < \xi_{N+1} = 1$ denote a sequence of equally spaced points, called interior knots, on $[0, 1]$. Denote by $H = 1/(N+1)$ the width of each subinterval $\left[\xi_J, \xi_{J+1}\right]$ for each $j \in \{0, \ldots, N\}$ and denote the degenerate knots $\xi_{-1} = 0$, $\xi_{N+2} = 1$. We need the following assumption.

(A8) The number of interior knots $N \sim n^{1/4}\ln n$, i.e., $c_N n^{1/4}\ln n \leq N \leq C_N n^{1/4}\ln n$ for some constants $c_N, C_N > 0$.

Following [11], for each $j \in \{0, \ldots, N\}$, define the linear B-spline basis as follows:

$$b_J(x) = (1 - |x - \xi_J|/H)_+ = \begin{cases} (N+1)\,x - J + 1 \text{if } \xi_{J-1} \le x \le \xi_J, \\ J + 1 - (N+1)\,x \text{if } \xi_J \le x \le \xi_{J+1}, \\ 0 \qquad\quad \text{otherwise.} \end{cases}$$

Let also the space of α-empirically centered linear spline functions on $[0, 1]$ be defined, for each $\alpha \in \{1, \ldots, d_2\}$, as

$$G^0_{n,\alpha} = \left\{ g_\alpha : g_\alpha(X_\alpha) = \sum_{J=0}^{N+1} \lambda_J b_J(X_\alpha),\ \mathrm{E}_n\{g_\alpha(X_\alpha)\} = 0 \right\},$$

and let the space of additive spline functions on $\boldsymbol{\chi}$ be

$$G^0_n = \left\{ g(\mathbf{x}) = \sum_{\alpha=1}^{d_2} g_\alpha(X_\alpha) : g_\alpha \in G^0_{n,\alpha} \right\}.$$

Define the log-likelihood function be given, for any $g \in G^0_n$, by

$$\hat{L}(\boldsymbol{\beta}, g) = \frac{1}{n} \sum_{i=1}^{n} [Y_i\{\boldsymbol{\beta}^\top \mathbf{T}_i + g(\mathbf{X}_i)\} - b\{\boldsymbol{\beta}^\top \mathbf{T}_i + g(\mathbf{X}_i)\}], \tag{8}$$

which according to Lemma 14 of [19], has a unique maximizer with probability approaching 1. The multivariate function $m(\mathbf{x})$ is then estimated by the additive spline function $\hat{m}(\mathbf{x})$ with

$$\hat{m}(\mathbf{t}, \mathbf{x}) = \hat{\boldsymbol{\beta}}^\top \mathbf{t} + \hat{m}(\mathbf{x}) = \mathrm{argmax}_{g \in G^0_n} \hat{L}(\boldsymbol{\beta}, g).$$

Since $\hat{m}(\mathbf{x}) \in G^0_n$, one can write $\hat{m}(\mathbf{x}) = \sum_{\alpha=1}^{d_2} \hat{m}_\alpha(x_\alpha)$ for $\hat{m}_\alpha(X_\alpha) \in G^0_{n,\alpha}$. Next define the log-likelihood function

$$\hat{\ell}_{m_1}(a, x_1) = \frac{1}{n} \sum_{i=1}^{n} \left[Y_i \left\{ a + \hat{m}(\mathbf{T}_i, \mathbf{X}_{i\_1}) \right\} - b \left\{ a + \hat{m}(\mathbf{T}_i, \mathbf{X}_{i\_1}) \right\} \right] K_h(X_{i1} - x_1), \tag{9}$$

where $\hat{m}(\mathbf{T}_i, \mathbf{X}_{i\_1}) = \hat{\boldsymbol{\beta}}^\top \mathbf{T}_i + \sum_{\alpha=2}^{d_2} \hat{m}_\alpha(X_{i\alpha})$. Define the SBK estimator as

$$\hat{m}_{\mathrm{SBK},1}(x_1) = \mathrm{argmax}_{a \in A} \hat{\ell}_{m_1}(a, x_1). \tag{10}$$

Theorem 2. *Under Assumptions* (A1)–(A8), *as* $n \to \infty$, $\hat{m}_{\mathrm{SBK},1}(x_1)$ *is oracally efficient,*

$$\sup_{x_1 \in [0,1]} |\hat{m}_{\mathrm{SBK},1}(x_1) - \tilde{m}_{K,1}(x_1)| = \mathcal{O}_{a.s.}(n^{-1/2} \ln n).$$

The following corollary is a consequence of Lemma 1 and Theorem 2.

Corollary 1. *Under Assumptions* (A1)–(A8), *as* $n \to \infty$, *the SBK estimator* $\hat{m}_{\mathrm{SBK},1}(x_1)$ *given in* (10) *satisfies*

$$\sup_{x_1 \in [h, 1-h]} |\hat{m}_{\mathrm{SBK},1}(x_1) - m_1(x_1)| = \mathcal{O}_{a.s.}(\ln n / \sqrt{nh})$$

and for any $x_1 \in [h, 1-h]$, *with* $\mathrm{bias}_1(x_1)$ *as in* (6) *and* $D_1(x_1)$ *in* (5)

$$\sqrt{nh}\,\{\hat{m}_{\mathrm{SBK},1}(x_1) - m_1(x_1) - \mathrm{bias}_1(x_1)h^2/D_1(x_1)\} \rightsquigarrow \mathcal{N}[0, D_1(x_1)^{-1} v_1^2(x_1) D_1(x_1)^{-1}].$$

Denote $a_h = \sqrt{-2 l_{n,h}}$, $C(K) = \left\|K'\right\|_2^2 \|K\|_2^{-2}$ and for any $\alpha \in (0, 1)$, the quantile

$$Q_h(\alpha) = a_h + a_h^{-1}[\ln\{\sqrt{C(K)}/(2\pi)\} - \ln\{-\ln\sqrt{1-\alpha}\}].$$

Also with $D_1(x_1)$ and $v_1^2(x_1)$ given in (5), define $\sigma_n(x_1) = n^{-1/2} h^{-1/2} v_1(x_1) D_1^{-1}(x_1)$.

Theorem 3. *Under Assumptions* (A1)–(A4), (A5'), (A6)–(A8), *as* $n \to \infty$,

$$\lim_{n \to \infty} \Pr\left\{ \sup_{x_1 \in [h, 1-h]} \left| \hat{m}_{\mathrm{SBK},1}(x_1) - m_1(x_1) \right| / \sigma_n(x_1) \le Q_h(\alpha) \right\} = 1 - \alpha.$$

A $100 \times (1-\alpha)\,\%$ *simultaneous confidence band for* $m_1(x_1)$ *is* $\hat{m}_{\mathrm{SBK},1}(x_1) \pm \sigma_n(x_1) Q_h(\alpha)$.

In fact, $\hat{\boldsymbol{\beta}}$ obtained by maximizing (8) is equivalent to $\hat{\boldsymbol{\beta}}_{\mathrm{SBK}} = \mathrm{argmax}_{\mathbf{a} \in \mathbb{R}^{1+d_1}} \hat{\ell}_{\boldsymbol{\beta}}(\mathbf{a})$ with

$$\hat{\ell}_{\boldsymbol{\beta}}(\mathbf{a}) = \frac{1}{n} \sum_{i=1}^{n} [Y_i\{\mathbf{a}^\top \mathbf{T}_i + \hat{m}(\mathbf{X}_i)\} - b\{\mathbf{a}^\top \mathbf{T}_i + \hat{m}(\mathbf{X}_i)\}]$$

in which $\hat{m}(\mathbf{X}_i) = \sum_{\alpha=1}^{d_2} \hat{m}_\alpha(X_{i\alpha})$. The empirical likelihood ratio is

$$\hat{R}(\mathbf{a}) = \max\left\{\prod_{i=1}^{n} np_i : \sum_{i=1}^{n} p_i \hat{Z}_i(\mathbf{a}) = \mathbf{0}, p_1 \geq 0, \ldots, p_n \geq 0, \sum_{i=1}^{n} p_i = 1\right\}$$

where $\hat{Z}_i(\mathbf{a}) = \left[Y_i - b'\left\{\mathbf{a}^\top \mathbf{T}_i + \hat{m}(\mathbf{X}_i)\right\}\right]\mathbf{T}_i$. Similar to Theorem 2, the main result shows that the difference between $\hat{\boldsymbol{\beta}}$ and its infeasible counterpart $\tilde{\boldsymbol{\beta}}$ is asymptotically negligible.

Theorem 4. (i) *Under Assumptions* (A1)–(A6) *and* (A8), *as* $n \to \infty$, $\hat{\boldsymbol{\beta}}$ *is oracally efficient, i.e.,* $\sqrt{n}(\hat{\beta}_k - \tilde{\beta}_k) \xrightarrow{p} 0$ *for all* $k \in \{0, \ldots, d_1\}$ *and hence*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightsquigarrow \mathcal{N}[\mathbf{0}, a(\phi)[Eb''\{m(\mathbf{T}, \mathbf{X})\}\mathbf{T}\mathbf{T}^\top]^{-1}].$$

(ii) *Under Assumptions* (A1)–(A4), (A5'), (A6) *and* (A8), *as* $n \to \infty$, $\sup|-2\ln\hat{R}(\boldsymbol{\beta}) + 2\ln\tilde{R}(\boldsymbol{\beta})| = \mathcal{O}_p(1)$, *and hence* $-2\ln\{\hat{R}(\boldsymbol{\beta})\} \rightsquigarrow \chi^2_{d_1}$.

As a reviewer pointed out, an obvious advantage of GAPLM over GAM is the capability of including categorical predictors. Since m_α is not a function of $\mathbf{T}$ in GAPLM, we can simply create dummy variables to represent the categorical effects and use spline estimation. [13] proposed spline estimation combined with categorical kernel functions to handle the case when function m_α depends on categorical predictors.

5. Examples

We have applied the SBK procedure to both simulated (Example 1) and real (Example 2) data and implemented our algorithms with the following rule-of-thumb number of interior knots

$$N = N_n = \min(\lfloor n^{1/4}\ln n\rfloor + 1, \lfloor n/4d - 1/d\rfloor - 1),$$

which satisfies (A8), i.e., $N = N_n \sim n^{1/4}\ln n$, and ensures that the number of parameters in the linear least squares problem is less than $n/4$, i.e., $1 + d(N+1) \leq n/4$. The bandwidth of h_α is computed as in [11] in an asymptotically optimal way.

5.1. Example 1

The data are generated from the model

$$\Pr(Y = 1 \mid \mathbf{T} = \mathbf{t}, \mathbf{X} = \mathbf{x}) = b'\left\{\boldsymbol{\beta}^\top \mathbf{T} + \sum_{\alpha=1}^{d_2} m_\alpha(X_\alpha)\right\}, \quad b'(x) = \frac{e^x}{1+e^x}$$

with $d_1 = 2, d_2 = 5, \boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\top = (1, 1, 1,)^\top$, $m_1(x) = m_2(x) = m_3(x) = \sin(2\pi x)$, $m_4(x) = \Phi(6x - 3) - 0.5$ and $m_5(x) = x^2 - 1/3$, where Φ is the standard normal cdf. The predictors are generated by transforming the following vector autoregression (VAR) equation for $0 \leq r_1, r_2 < 1$ and all $i \in \{1, \ldots, n\}$, viz. $\mathbf{Z}_0 = \mathbf{0}$, and

$$\mathbf{Z}_i = r_1\mathbf{Z}_{i-1} + \boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \Sigma), \Sigma = (1 - r_2)\mathbf{I}_{d\times d} + r_2\mathbf{1}_d\mathbf{1}_d^\top, \quad d = d_1 + d_2,$$
$$\mathbf{T}_i = \left(1, Z_{i1}, \ldots, Z_{id_1},\right)^\top, X_{i\alpha} = \Phi\left(\sqrt{1 - r_1^2}Z_{i\alpha}\right), \quad 1 + d_1 \leq \alpha \leq d_1 + d_2,$$

with stationary $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{id})^\top \sim \mathcal{N}[0, (1 - r_1^2)^{-1}\Sigma]$, $\mathbf{1}_d = (1, \ldots, 1)^\top$ and $\mathbf{I}_{d\times d}$ is the $d \times d$ identity matrix. The X is transformed from Z to satisfy Assumption (A4). In this study, we selected four scenarios: (a) $r_1 = 0$, $r_2 = 0$; (b) $r_1 = 0.5$, $r_2 = 0$; (c) $r_1 = 0, r_2 = 0.5$; (d) $r_1 = 0.5, r_2 = 0.5$. The parameter r_1 controls the dependence between observations and r_2 controls the correlation between variables. In the selected scenarios, $r_1 = 0$ indicates independent observations and $r_1 = 0.5$ α-mixing observations, $r_2 = 0$ indicates independent variables and $r_2 = 0.5$ correlated variables within each observation. Define the empirical relative efficiency of $\hat{\beta}_1$ with respect to $\tilde{\beta}_1$ as $\mathrm{EFF}_r(\hat{\beta}_1) = \{\mathrm{MSE}(\tilde{\beta}_1)/\mathrm{MSE}(\hat{\beta}_1)\}^{1/2}$.

Table 1 shows the mean of bias, variances, MSEs and EFFs of $\hat{\beta}_1$ for $R = 1000$ with sample sizes $n \in \{500, 1000, 2000, 4000\}$. The results show that the estimator works as the asymptotic theory indicates, see Theorem 4(i).

Fig. 1 shows the kernel densities of $\hat{\beta}_1$s for $n \in \{500, 1000, 2000, 4000\}$ from 1000 replications, again the theoretical properties are supported.

Table 2 shows the simulation results of the empirical likelihood confidence interval for β with $n \in \{500, 1000, 2000, 4000\}$, and $r_1 = 0, r_2 = 0$ from 1000 replications. The mean and standard deviation of $-2\ln\{\hat{R}(\boldsymbol{\beta})\} + 2\ln\tilde{R}\{(\boldsymbol{\beta})\}$ (DIFF) support the oracle efficiency in Theorem 4 (ii). The performance of empirical likelihood confidence interval are compared with the wald-type one and it is clear that they have similar performance but empirical likelihood confidence interval has better coverage ratio and shorter average length.

Next for $\alpha \in \{1, \ldots, 5\}$, let $X^i_{\alpha,\min}, X^i_{\alpha,\max}$ denote the smallest and largest observations of the variable X_α in the ith replication, respectively. The component functions $m_1, \ldots, m_5$ are estimated on equally spaced points such that $0 = x_0 <$

Table 1
The mean of 10 × bias, 100 × variances, 100 × MSEs and EFFs of $\hat{\beta}_1$ from 1000 replications.

r	n	$10 \times \overline{\text{BIAS}}$	$100 \times \overline{\text{VARIANCE}}$	$100 \times \overline{\text{MSE}}$	$\overline{\text{EFF}}\left(\hat{\beta}_1\right)$
	500	1.509	2.018	4.298	0.8436
$r_1 = 0$	1000	0.727	1.197	1.726	0.8749
$r_2 = 0$	2000	0.408	0.626	0.793	0.9189
	4000	0.240	0.282	0.339	0.9534
	500	1.473	3.136	5.306	0.8392
$r_1 = 0.5$	1000	0.834	1.287	1.983	0.8873
$r_2 = 0$	2000	0.476	0.674	0.901	0.9294
	4000	0.260	0.202	0.270	0.9665
	500	1.327	3.880	5.642	0.8475
$r_1 = 0$	1000	0.699	1.851	2.339	0.8856
$r_2 = 0.5$	2000	0.665	0.739	1.182	0.9353
	4000	0.390	0.290	0.442	0.9479
	500	1.635	4.230	6.903	0.8203
$r_1 = 0.5$	1000	0.901	1.190	2.002	0.8758
$r_2 = 0.5$	2000	0.529	0.806	1.086	0.9304
	4000	0.209	0.366	0.410	0.9483

Table 2
Coverage ratios and average length of the empirical likelihood confidence interval (EL) and Wald-type confidence interval for β_1 for $n = 500, 1000, 2000, 4000$ with $r_1 = 0$ from 1000 replications. DIFF$= -2\ln\{\hat{R}(\boldsymbol{\beta})\} + 2\ln\{\tilde{R}(\boldsymbol{\beta})\}$ is the difference between $-2\ln\{\hat{R}(\boldsymbol{\beta})\}$ and $-2\ln\{\tilde{R}(\boldsymbol{\beta})\}$.

		$n = 500$	$n = 1000$	$n = 2000$	$n = 4000$
Coverage ratio	EL	0.923	0.941	0.946	0.951
	Wald	0.918	0.934	0.944	0.948
Average length	EL	1.2675	0.9474	0.7105	0.5339
	Wald	1.4073	1.0447	0.7480	0.5625
DIFF	MEAN	0.1213	0.1023	0.0981	0.0726
	SD	0.5199	0.4703	0.3667	0.3242

$\cdots < x_{100} = 1$ and the estimator of m_α in the rth sample as $\hat{m}_{\text{SBK},\alpha,r}$. The (mean) average squared errors (ASE and MASE) are:

$$\text{ASE}(\hat{m}_{\text{SBK},\alpha,r}) = \frac{1}{101}\sum_{t=0}^{100}\left\{\hat{m}_{\text{SBK},\alpha,r}(x_t) - m_\alpha(x_t)\right\}^2,$$

$$\text{MASE}(\hat{m}_{\text{SBK},\alpha}) = \frac{1}{R}\sum_{r=1}^{R}\text{ASE}(\hat{m}_{\text{SBK},\alpha,r}).$$

In order to examine the efficiency of $\hat{m}_{\text{SBK},\alpha}$ relative to the oracle estimator $\tilde{m}_{K,\alpha}(x_\alpha)$, both are computed using the same data-driven bandwidth $\hat{h}_{\alpha,\text{opt}}$, described in Section 5 of [11]. Define the empirical relative efficiency of $\hat{m}_{\text{SBK},\alpha}$ with respect to $\tilde{m}_{K,\alpha}$ as

$$\text{EFF}_r\left(\hat{m}_{\text{SBK},\alpha}\right) = \left[\frac{\sum_{t=0}^{100}\left\{\tilde{m}_{K,\alpha}(x_t) - m_\alpha(x_t)\right\}^2}{\sum_{t=0}^{100}\left\{\hat{m}_{\text{SBK},\alpha,r}(x_t) - m_\alpha(x_t)\right\}^2}\right]^{1/2}.$$

EFF measures the relative efficiency of the SBK estimator to the oracle estimator. For increasing sample size, it should increase to 1 by Theorem 2. Table 3 shows the MASEs of $\tilde{m}_{K,1}$, $\hat{m}_{\text{SBK},1}$ and the average of EFFs from 1000 replications for $n \in \{500, 1000, 2000, 4000\}$. It is clear that the MASEs of both SBK estimator and the oracle estimator decrease when sample sizes increase, and the SBK estimator performs as well asymptotically as the oracle estimator, see Theorem 2.

To have an impression of the actual function estimates, for $r_1 = 0, r_2 = 0.5$ with sample size $n \in \{500, 1000, 2000, 4000\}$, we have plotted the SBK estimators and their 95% asymptotic SCCs (red solid lines), point-wise confidence intervals (red dashed lines), oracle estimators (blue dashed lines) for the true functions m_1 (thick black lines) in Fig. 2. Here we use $r_1 = 0$ because we want to give the 95% asymptotic SCCs, which need the observations be iid to satisfy Assumption (A5'). As expected by theoretical results, the estimation is closer to the real function and the confidence band is narrower as sample size increasing.

To compare the prediction performance of GAM and GAPLM, we introduce CAP and AR first. For any score function S, one defines its alarm rate $F(s) = \Pr(S \le s)$ and the hit rate $F_{\text{D}}(s) = \Pr(S \le s \mid D)$ where D represents the conditioning event of "default". Define the Cumulative Accuracy Profile (CAP) curve, for each $u \in (0, 1)$, as

$$\text{CAP}(u) = F_{\text{D}}\{F^{-1}(u)\}, \tag{11}$$

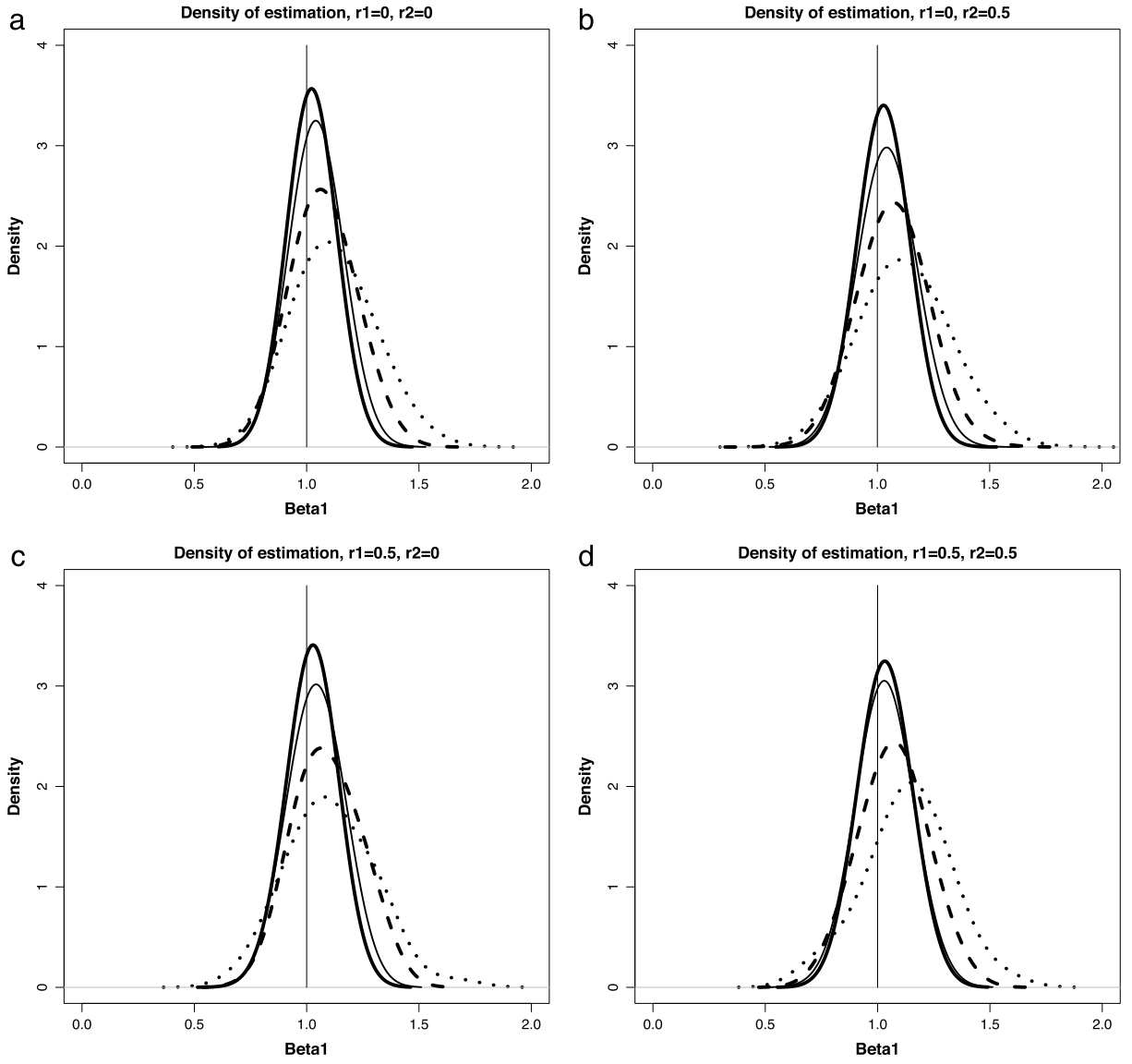


Fig. 1. Plots of densities for $\hat{\beta}_1$ with $n = 500$ (dotted line), $n = 1000$ (dashed line), $n = 2000$ (thin solid line), $n = 4000$ (thick solid line) for (a) $r_1 = 0, r_2 = 0$, (b) $r_1 = 0, r_2 = 0.5$, (c) $r_1 = 0.5, r_2 = 0$, (d) $r_1 = 0.5, r_2 = 0.5$ from 1000 replications.

which is the percentage of default-infected obligators that are found among the first (according to their scores) $100 \times u\%$ of all obligators. A perfect rating method assigns all lowest scores to exactly the defaulters, so its CAP curve linearly increases up and then stays at 1; in other words, $\mathrm{CAP_P}(u) = \min(u/p, 1)$ for all $u \in (0, 1)$, where p denotes the unconditional default probability. In contrast, a noninformative rating method with zero discriminatory power displays a diagonal line $\mathrm{CAP_N}(u) = u$ for all $u \in (0, 1)$. The CAP curve of a given scoring method S always locates between these two extremes and gives information about its performance.

The area between the CAP curve and the noninformative diagonal $\mathrm{CAP_N}(u) \equiv u$ is a_R, whereas a_P is the area between the perfect CAP curve $\mathrm{CAP_P}(u)$ and the noninformative diagonal $\mathrm{CAP_N}(u)$. Thus the CAP can be measured for example by Accuracy Ratio (AR): the ratio of a_R and a_P, viz.

$$\mathrm{AR} = \frac{a_R}{a_P} = \frac{2}{1-p}\left\{\int_0^1 \mathrm{CAP}(u)\,du - 1\right\},$$

where $\mathrm{CAP}(u)$ is given in (11). The AR takes value in $[0, 1]$, with value 0 corresponding to the noninformative scoring, and 1 the perfect scoring method. A higher AR indicates an overall higher discriminatory power of a method. Table 4 shows the average and standard deviations of the ARs from 1000 replications using k-fold cross-validation with $k \in \{2, 10, 100\}$ for

Table 3
The 100×MASEs of $\tilde{m}_{\mathrm{K},1}$, $\hat{m}_{\mathrm{SBK},1}$ and $\overline{\mathrm{EFF}}$s for $n \in \{500, 1000, 2000, 4000\}$ from 1000 replications.

r	n	$100 \times \mathrm{MASE}\left(\tilde{m}_{\mathrm{K},\alpha}\right)$	$100 \times \mathrm{MASE}\left(\hat{m}_{\mathrm{SBK},\alpha}\right)$	$\overline{\mathrm{EFF}}\left(\hat{m}_{\mathrm{SBK},1}\right)$
	500	4.482	4.603	0.9501
$r_1 = 0$	1000	2.418	2.503	0.9809
$r_2 = 0$	2000	1.582	1.613	0.9854
	4000	1.212	1.247	0.9923
	500	4.060	4.322	0.9445
$r_1 = 0.5$	1000	2.592	2.649	0.9767
$r_2 = 0$	2000	1.746	1.714	0.9832
	4000	1.194	1.218	0.9936
	500	4.845	6.348	0.8827
$r_1 = 0$	1000	2.935	3.559	0.8755
$r_2 = 0.5$	2000	1.951	2.177	0.9494
	4000	1.515	1.648	0.9795
	500	5.656	7.114	0.8722
$r_1 = 0.5$	1000	2.804	3.570	0.8951
$r_2 = 0.5$	2000	1.886	2.089	0.9478
	4000	1.525	1.634	0.9744

Table 4
The mean and standard deviation (in parentheses) of Accuracy Ratio (AR) values for GLM, GAM, GAPLM for $r_1 = 0$, $r_2 = 0$ from 1000 replications.

n		$k = 2$	$k = 10$	$k = 100$
500	GLM	0.6287 (0.0436)	0.6412 (0.0397)	0.6438 (0.0390)
	GAM	0.6222 (0.0732)	0.6706 (0.0393)	0.6756 (0.0400)
	GAPLM	0.6511 (0.0479)	0.6828 (0.0377)	0.6861 (0.0391)
1000	GLM	0.6429 (0.0282)	0.6476 (0.0268)	0.6488 (0.0268)
	GAM	0.6735 (0.0438)	0.6863 (0.0326)	0.6929 (0.0261)
	GAPLM	0.6861 (0.0298)	0.6968 (0.0254)	0.7001 (0.0258)
2000	GLM	0.6474 (0.0204)	0.6513 (0.0195)	0.6519 (0.0188)
	GAM	0.6842 (0.0615)	0.6984 (0.0286)	0.7000 (0.0185)
	GAPLM	0.6984 (0.0204)	0.7067 (0.0178)	0.7057 (0.0178)
4000	GLM	0.6507 (0.0134)	0.6522 (0.0136)	0.6529 (0.0132)
	GAM	0.6889 (0.0243)	0.6968 (0.0403)	0.7079 (0.0164)
	GAPLM	0.7056 (0.0130)	0.7110 (0.0124)	0.7119 (0.0119)

$r_1 = 0, r_2 = 0$ and $n \in \{500, 1000, 2000, 4000\}$. In each replication, we randomly divide the set of observations into k equal size folds and use the remaining $k - 1$ folds as training data set to make prediction for each fold. After we obtain all the predictions for each observation in the data set, we compute the CAP and AR based on above formula. It is clear that GAPLM has best predication accuracy.

Finally, to show the estimation performance when $\mathbf{T}$ has categorical variables, we generate data using the same model above but add one more categorical variable, i.e., $d_1 = 3$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^\top = (1, 1, 1, 1)^\top$, $T_3 = \{0, 1\}$ with probability 0.5 for $T_3 = 1$ and independent with the other variables T and X. Table 5 shows the bias, variances, MSEs and EFFs of $\hat{\beta}_3$ for $R = 1000$ with sample sizes $n \in \{500, 1000, 2000, 4000\}$. The results show that the estimator works as the asymptotic theory indicates.

5.2. Example 2

The credit reform database, provided by the Research Data Center (RDC) of the Humboldt Universität zu Berlin, was studied using a GAM in [11]. The data set contains $d = 8$ financial ratios, which are shown in Table 6, of 19,610 German companies (18,610 solvent and 1000 insolvent). The time period ranges from 1997 to 2002 and in the case of the insolvent companies the information was gathered two years before the insolvency took place. The last annual report of a company before it went bankrupt receives the indicator $Y = 1$ and for the rest (solvent) $Y = 0$. In the original data set, the variables are labeled as Z_α. In order to satisfy the Assumption (A4) in [11], we need the transformation $X_{i\alpha} = F_{n\alpha}(Z_{i\alpha})$ for all $\alpha \in \{1, \ldots, 8\}$, where $F_{n\alpha}$ is the empirical cdf of the data $X_{1\alpha}, \ldots, X_{n\alpha}$. See [3,11] for more details of this data set.

Using a GAM and the SBK method, we clearly see via the SCCs that the shape of $m_2(x_2)$ is linear. Fig. 3(a) shows that a linear line is covered by the SCCs of $\hat{m}_2$. We additionally show the SCCs for another component function of ln(Total_Assets) in Fig. 3(b). The SCCs do not cover a linear line. In fact, among the eight financial ratios considered, only x_2 yields a linear influence. To improve the precision in statistical calibration and interpretability, we can use GAPLM with parametric $m_2(x_2) = \beta_2 x_2$.

For the RDC data, the in-sample AR value obtained from GAPLM is 62.89%, which is very close to the AR value 63.05% obtained from GAM in [11] and higher than the AR value 60.51% obtained from SVM in [3]. To compare the prediction performance, we use the AR introduced in Example 1. Then we randomly divide the data set into $k \in \{2, 10\}$ folds and

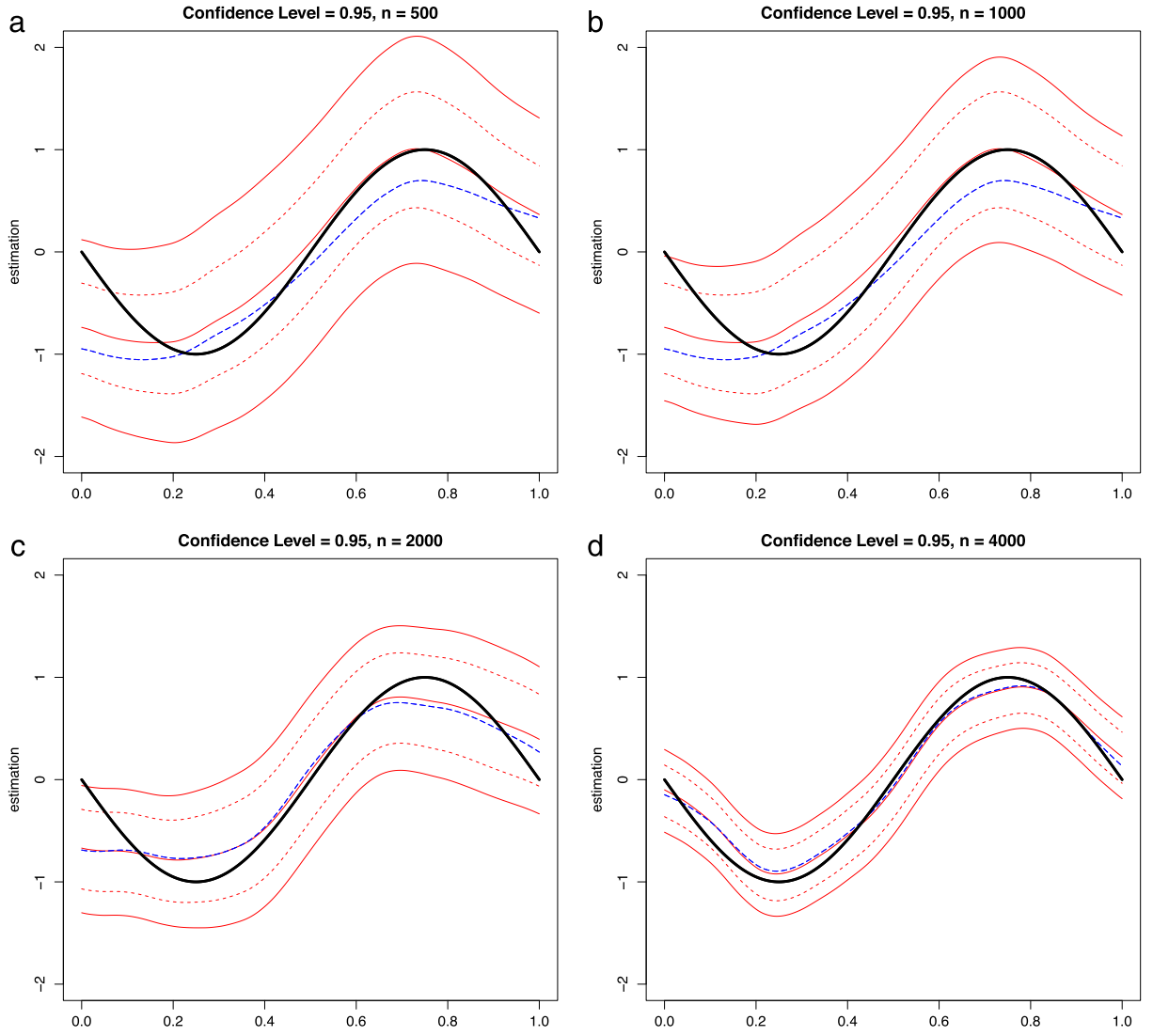


Fig. 2. Plots of m_1 (thick black line), $\tilde{m}_{\mathrm{K},1}$ (blue dashed line), asymptotic 95% point-wise confidence intervals (red dashed line), $\hat{m}_{\mathrm{SBK},1}$ and 95% simultaneous confidence bands (red solid line) for $r_1 = 0$, $r_2 = 0.5$ and (a) $n = 500$, (b) $n = 1000$, (c) $n = 2000$, (d) $n = 4000$.

obtain the prediction for each observation using the remaining $k - 1$ folds as training set. Based on the prediction of all the observations, we can compute prediction AR value. Table 7 shows the mean and standard deviation of the prediction AR values from 100 replications. GAPLM has higher prediction AR value than GAM for 99 replications when $k = 2$ and 100 times when $k = 10$. It is clear that GAPLM has best prediction accuracy due to the better statistical calibration.

Acknowledgments

Financial support from the Deutsche Forschungsgemeinschaft (DFG) via SFB 649 "Economic Risk", and International Research Training Group (IRTG) 1792 was gratefully acknowledged. The authors thank the Editor-in-Chief, the Associate Editor and two referees for their comments and suggestions which have led to substantial improvement of this work.

Appendix A

A.1. Preliminaries

In the proofs that follow, we use "$\mathcal{U}$" and "$\mathcal{U}$" to denote sequences of random variables that are uniformly "$\mathcal{O}$" and "$\mathcal{O}$ " of certain order. Denote the theoretical inner product of b_J and 1 with respect to the αth marginal density $f_\alpha(X_\alpha)$ as

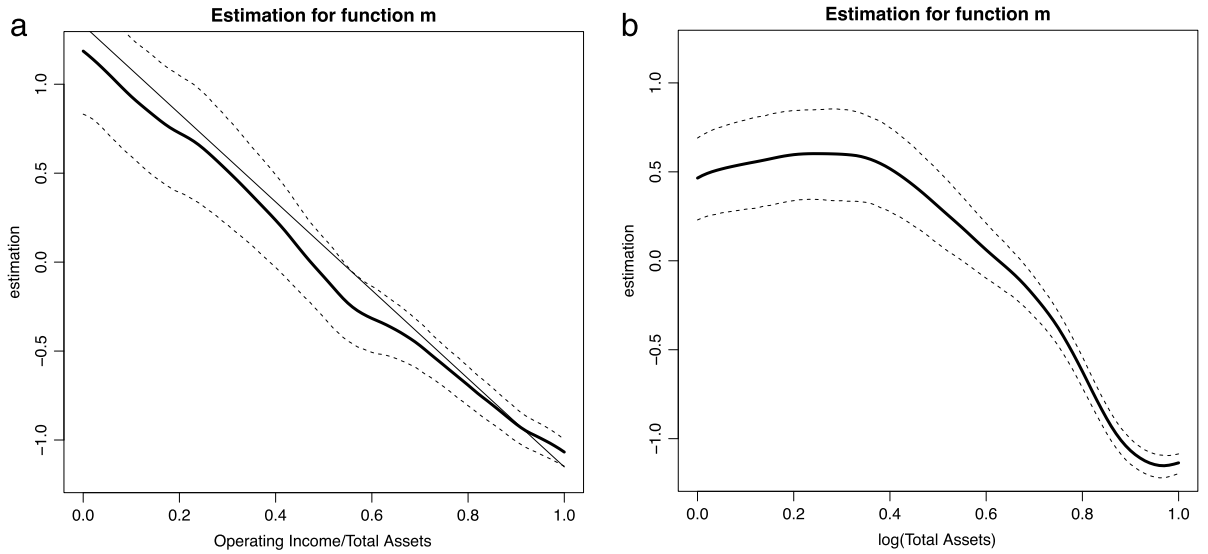


Fig. 3. Plots of estimations of component functions (a) $\hat{m}_{\mathrm{SBK},2}(x_2)$ and (b) $\hat{m}_{\mathrm{SBK},8}(x_8)$ and asymptotic 95% simultaneous confidence bands.

Table 5
The mean of 10 × bias, 100 × variances, 100 × MSEs and EFFs of $\hat{\beta}_3$ from 1000 replications.

r	n	$10 \times \overline{\mathrm{BIAS}}$	$100 \times \overline{\mathrm{VARIANCE}}$	$100 \times \overline{\mathrm{MSE}}$	$\overline{\mathrm{EFF}}(\hat{\beta}_3)$
	500	1.476	10.129	12.309	0.7634
$r_1 = 0$	1000	0.770	4.437	5.031	0.8343
$r_2 = 0$	2000	0.448	1.846	2.047	0.8929
	4000	0.315	0.937	1.037	0.9572
	500	1.336	10.329	12.115	0.7445
$r_1 = 0.5$	1000	0.833	4.221	4.916	0.8267
$r_2 = 0$	2000	0.423	1.952	2.132	0.8832
	4000	0.302	0.944	1.036	0.9436
	500	1.441	10.154	12.231	0.7556
$r_1 = 0$	1000	0.803	4.446	5.114	0.8430
$r_2 = 0.5$	2000	0.489	2.136	2.376	0.8785
	4000	0.328	0.924	1.032	0.9572
	500	1.475	11.014	13.190	0.7794
$r_1 = 0.5$	1000	0.812	4.464	5.124	0.8314
$r_2 = 0.5$	2000	0.524	1.970	2.245	0.8852
	4000	0.302	0.966	1.058	0.9529

Table 6
Definitions of financial ratios.

Ratio No.	Definition	Ratio No.	Definition
Z_1	Net_Income/Sales	Z_5	Cash/Total_Assets
Z_2	Operating_Income/Total_Assets	Z_6	Inventories/Sales
Z_3	Ebit/Total_Assets	Z_7	Accounts_Payable/Sales
Z_4	Total_Liabilities/Total_Assets	Z_8	ln(Total_Assets)

Table 7
The mean and standard deviation (in parentheses) of AR values for GLM, GAM, GAPLM for k-fold cross-validation with $k \in \{2, 10\}$ from 1000 replications.

	$k = 2$	$k = 10$
GLM	0.5627 (0.0271)	0.5751 (0.00162)
GAM	0.5888 (0.0405)	0.6123 (0.00219)
GAPLM	0.5928 (0.0408)	0.6164 (0.00196)

$c_{J,\alpha} = \langle b_J(X_\alpha), 1\rangle = \int b_J(X_\alpha) f_\alpha(X_\alpha) dx_\alpha$ and define the centered B-spline basis $b_{J,\alpha}(x_\alpha)$ and the standardized B-spline basis, for each $J \in \{1, \ldots, N+1\}$, as

$$b_{J,\alpha}(X_\alpha) = b_J(X_\alpha) - \frac{c_{J,\alpha}}{c_{J-1,\alpha}} b_{J-1}(X_\alpha), \quad B_{J,\alpha}(X_\alpha) = \frac{b_{J,\alpha}(x_\alpha)}{\|b_{J,\alpha}\|_2},$$

so that $\mathrm{E}\{B_{J,\alpha}(X_\alpha)\} = 0$ and $\mathrm{E}\{B^2_{J,\alpha}(X_\alpha)\} = 1$. Theorem A.2 in [21] shows that under Assumptions . (A1)–(A5) and (A7), constants $c_0(f), C_0(f), c_1(f)$ and $C_1(f)$ exist depending on the marginal densities $f_1, \ldots, f_d$ such that $c_0(f) H \le c_{J,\alpha} \le C_0(f) H$ and

$$c_1(f) H \le \|b_{J,\alpha}\|_2^2 \le C_1(f) H. \tag{A.1}$$

Lemma A.1 (*[1], p. 149*). *For any $m \in C^1[0,1]$ with $m' \in \mathrm{Lip}([0,1], C_\infty)$, there exist a constant $C_\infty > 0$ and a function $g \in G_n^{(0)}[0,1]$ such that $\|g - m\|_\infty \le C_\infty H^2$.*

A.2. Oracle estimators

Proof of Theorem 1. (i) According to the Mean Value Theorem, a vector $\bar{\boldsymbol{\beta}}$ between $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\beta}}$ exists such that $(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \nabla^2 \tilde{\ell}_{\boldsymbol{\beta}}(\bar{\boldsymbol{\beta}}) = \nabla \tilde{\ell}_{\boldsymbol{\beta}}(\tilde{\boldsymbol{\beta}}) - \nabla \tilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = -\nabla \tilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\beta})$ since $\nabla \tilde{\ell}_{\boldsymbol{\beta}}(\tilde{\boldsymbol{\beta}}) = \mathbf{0}$, where

$$-\nabla^2 \tilde{\ell}_{\boldsymbol{\beta}}(\bar{\boldsymbol{\beta}}) = n^{-1} \sum_{i=1}^n b''\{\bar{\boldsymbol{\beta}}^\top T_i + m(\mathbf{X}_i)\} \mathbf{T}_i \mathbf{T}_i^\top > c_b c_{\mathbf{Q}} \mathbf{I}_{d_1 \times d_1}$$

with $c_b > 0$ according to (A2), and then the infeasible estimator is $\tilde{\boldsymbol{\beta}} = \mathrm{argmax}_{\mathbf{a} \in \mathbb{R}^{1+d_1}} \tilde{\ell}_{\boldsymbol{\beta}}(\mathbf{a})$.

$$\nabla \tilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n [Y_i \mathbf{T}_i - b'\{\boldsymbol{\beta}^\top \mathbf{T}_i + m(\mathbf{X}_i)\} \mathbf{T}_i] = \frac{1}{n} \sum_{i=1}^n \sigma(\mathbf{T}_i, \mathbf{X}_i) \varepsilon_i \mathbf{T}_i.$$

We have $|n^{-1} \sum_{i=1}^n \sigma(\mathbf{T}_i, \mathbf{X}_i) \varepsilon_i \mathbf{T}_i| = \mathcal{O}_{a.s.}(n^{-1/2} \ln n)$ by Bernstein's Inequality as Lemma A.2 in [11], so $|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}| = \mathcal{O}_{a.s.}(n^{-1/2} \ln n)$ according to $\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} = -\{\nabla^2 \tilde{\ell}_{\boldsymbol{\beta}}(\bar{\boldsymbol{\beta}})\}^{-1} \nabla \tilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\beta})$. Then

$$\nabla^2 \tilde{\ell}_{\boldsymbol{\beta}}(\bar{\boldsymbol{\beta}}) \xrightarrow{a.s.} \nabla^2 \tilde{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = -\frac{1}{n} \sum_{i=1}^n b''\{\boldsymbol{\beta}^\top \mathbf{T}_i + m(\mathbf{X}_i)\} \mathbf{T}_i \mathbf{T}_i^\top,$$

which converges to $-\mathrm{E}[b''\{m(\mathbf{T}, \mathbf{X})\} \mathbf{T}\mathbf{T}^\top]$ almost surely at the rate of $n^{-1/2} \ln n$. So

$$\left| \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} - [\mathrm{E} b''\{m(\mathbf{T}, \mathbf{X})\} \mathbf{T}\mathbf{T}^\top]^{-1} \frac{1}{n} \sum_{i=1}^n \sigma(\mathbf{T}_i, \mathbf{X}_i) \varepsilon_i \mathbf{T}_i \right| = \mathcal{O}_{a.s.}\{n^{-1} (\ln n)^2\}.$$

Since $n^{-1} \sum_{i=1}^n \sigma(\mathbf{T}_i, \mathbf{X}_i) \varepsilon_i \mathbf{T}_i \rightsquigarrow \mathcal{N}[\mathbf{0}, a(\phi) [\mathrm{E} b''\{m(\mathbf{T}, \mathbf{X})\} \mathbf{T}\mathbf{T}^\top]^{-1}]$ by the Central Limit Theorem, an application of Slutsky's Lemma completes the proof of Theorem 1(i).

(ii) The proof is trivial based on the properties of empirical likelihood ratio for GLMs; see Theorem 3.2 in [15] and Corollary 1 in [7]. □

A.3. Spline-backfitted kernel estimators

In this section, we present the proofs of Theorems 2–4. We write any $g \in G_n^0$ as $g = \boldsymbol{\lambda}^\top \mathbf{B}(\mathbf{X}_i)$ with vector $\boldsymbol{\lambda}_g = \left(\lambda_{J,\alpha}\right)^\top_{1 \le J \le N+1, 1 \le \alpha \le d_2} \in \mathbb{R}^{(N+1)d_2}$ the dimension of the additive spline space G_n^0, and

$$\mathbf{B}(\mathbf{x}) = \left(B_{1,1}(x_1), \ldots, B_{N+1,1}(x_1), \ldots, B_{1,d_2}\left(x_{d_2}\right), \ldots, B_{N+1,d_2}\left(x_{d_2}\right)\right)^\top.$$

Denote $\mathbf{B}(\mathbf{t}, \mathbf{x}) = \left(1, t_1, \ldots, t_{d_1}, B_{1,1}(x_1), \ldots, B_{N+1,1}(x_1), \ldots, B_{1,d_2}\left(x_{d_2}\right), \ldots, B_{N+1,d_2}\left(x_{d_2}\right)\right)^\top$,

$$\boldsymbol{\lambda} = (\boldsymbol{\lambda}_{\boldsymbol{\beta}}^\top, \boldsymbol{\lambda}_{\mathbf{g}}^\top)^\top = \left(\lambda_0, \lambda_k, \lambda_{J,\alpha}\right)^\top_{1 \le J \le N+1, 1 \le \alpha \le d_2, 1 \le k \le d_1} \in \mathbb{R}^{N_d}$$

with $N_d = 1 + d_1 + (N+1) d_2$ and

$$\hat{L}(\boldsymbol{\lambda}_\beta, g) = \hat{L}(\boldsymbol{\lambda}) = \frac{1}{n} \sum_{i=1}^n [Y_i\{\boldsymbol{\lambda}^\top \mathbf{B}(\mathbf{T}_i, \mathbf{X}_i)\} - b\{\boldsymbol{\lambda}^\top \mathbf{B}(\mathbf{T}_i, \mathbf{X}_i)\}],$$

which yields the gradient and Hessian formulas

$$\nabla\hat{L}(\boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n}[Y_i\mathbf{B}(\mathbf{T}_i,\mathbf{X}_i) - b'\{\boldsymbol{\lambda}^\top\mathbf{B}(\mathbf{T}_i,\mathbf{X}_i)\}\mathbf{B}(\mathbf{T}_i,\mathbf{X}_i)],$$

$$\nabla^2\hat{L}(\boldsymbol{\lambda}) = -\frac{1}{n}\sum_{i=1}^{n} b''\{\boldsymbol{\lambda}^\top\mathbf{B}(\mathbf{T}_i,\mathbf{X}_i)\}\mathbf{B}(\mathbf{T}_i,\mathbf{X}_i)\mathbf{B}(\mathbf{T}_i,\mathbf{X}_i)^\top.$$

The multivariate function $m(\mathbf{t},\mathbf{x})$ is estimated by

$$\hat{m}(\mathbf{t},\mathbf{x}) = \hat{\beta}_0 + \sum_{k=1}^{d_1}\hat{\beta}_k t_k + \sum_{\alpha=1}^{d_2}\hat{m}_\alpha(X_\alpha) = \hat{\boldsymbol{\lambda}}^\top\mathbf{B}(\mathbf{t},\mathbf{x}),$$

$$\hat{\boldsymbol{\lambda}} = (\hat{\boldsymbol{\lambda}}_{\boldsymbol{\beta}}^\top, \hat{\boldsymbol{\lambda}}_{\mathbf{g}}^\top)^\top = (\hat{\boldsymbol{\beta}}^\top, \hat{\boldsymbol{\lambda}}_{\mathbf{g}}^\top)^\top = (\hat{\beta}_k, \hat{\lambda}_{J,\alpha})^\top_{0\le k\le d_1, 1\le\alpha\le d_2, 1\le J\le N+1} = \operatorname{argmax}_{\boldsymbol{\lambda}}\hat{L}(\boldsymbol{\lambda}).$$

Lemma 14 of Stone [19] ensures that with probability approaching 1, $\hat{\boldsymbol{\lambda}}$ exists uniquely and that $\nabla\hat{L}(\hat{\boldsymbol{\lambda}}) = \mathbf{0}$. In addition, Lemma A.1 and (A1) provide a vector $\bar{\boldsymbol{\lambda}} = (\boldsymbol{\beta}^\top, \bar{\boldsymbol{\lambda}}_{\mathbf{g}}^\top)^\top$ and an additive spline function $\bar{m}$ such that

$$\bar{m}(\mathbf{x}) = \bar{\boldsymbol{\lambda}}_{\mathbf{g}}^\top\mathbf{B}(\mathbf{x}), \quad \|\bar{m} - m\|_\infty \le C_\infty H^2. \tag{A.2}$$

We first establish technical lemmas before proving Theorems 2 and 4.

Lemma A.2. *Under Assumptions* (A1)–(A6) *and* (A8), *as* $n\to\infty$,

$$|\nabla\hat{L}(\bar{\boldsymbol{\lambda}})| = \mathcal{O}_{a.s.}(H^2 + n^{-1/2}\ln n), \quad \|\nabla\hat{L}(\bar{\boldsymbol{\lambda}})\| = \mathcal{O}_{a.s.}(H^{3/2} + H^{-1/2}n^{-1/2}\ln n).$$

Proof. See Online Supplement. □

Define the following matrices:

$$\mathbf{V} = \mathrm{E}\mathbf{B}(\mathbf{T},\mathbf{X})\mathbf{B}(\mathbf{T},\mathbf{X})^\top, \quad \mathbf{S} = \mathbf{V}^{-1}, \quad \mathbf{V}_n = n^{-1}\sum_{i=1}^{n}\mathbf{B}(\mathbf{T}_i,\mathbf{X}_i)\mathbf{B}(\mathbf{T}_i,\mathbf{X}_i)^\top, \quad \mathbf{S}_n = \mathbf{V}_n^{-1},$$

$$\mathbf{V}_b = \mathrm{E}b''\{m(\mathbf{T},\mathbf{X})\}\mathbf{B}(\mathbf{T},\mathbf{X})\mathbf{B}(\mathbf{T},\mathbf{X})^\top = \begin{bmatrix} v_{b,00} & v_{b,0,k} & v_{b,0,J,\alpha} \\ v_{b,0,k'} & v_{b,k,k'} & v_{b,J,\alpha,k'} \\ v_{b,0,J',\alpha'} & v_{b,J',\alpha',k} & v_{b,J,\alpha,J',\alpha'} \end{bmatrix}_{N_d\times N_d}$$

where $N_d = (N+1)d_2 + 1 + d_1$, and

$$\mathbf{S}_b = \mathbf{V}_b^{-1} = \begin{bmatrix} s_{b,00} & s_{b,0,k} & s_{b,0,J,\alpha} \\ s_{b,0,k'} & s_{b,k,k'} & s_{b,J,\alpha,k'} \\ s_{b,0,J',\alpha'} & s_{b,J',\alpha',k} & b,J,\alpha,J',\alpha' \end{bmatrix}_{N_d\times N_d}. \tag{A.3}$$

For any vector $\boldsymbol{\lambda}\in\mathbb{R}^{N_d}$, denote

$$\mathbf{V}_b(\boldsymbol{\lambda}) = \mathrm{E}b''\{\boldsymbol{\lambda}^\top\mathbf{B}(\mathbf{T},\mathbf{X})\}\mathbf{B}(\mathbf{T},\mathbf{X})\mathbf{B}(\mathbf{T},\mathbf{X})^\top, \quad \mathbf{S}_b(\boldsymbol{\lambda}) = \mathbf{V}_b^{-1}(\boldsymbol{\lambda})$$

$$\mathbf{V}_{n,b}(\boldsymbol{\lambda}) = -\nabla^2\hat{L}(\boldsymbol{\lambda}), \quad \mathbf{S}_{n,b}(\boldsymbol{\lambda}) = \mathbf{V}_{n,b}^{-1}(\boldsymbol{\lambda}). \tag{A.4}$$

Lemma A.3. *Under Assumptions* (A2) *and* (A4), *one has*

$$c_{\mathbf{V}}\mathbf{I}_{N_d} \le \mathbf{V} \le C_{\mathbf{V}}\mathbf{I}_{N_d}, \quad c_{\mathbf{S}}\mathbf{I}_{N_d} \le \mathbf{S} \le C_{\mathbf{S}}\mathbf{I}_{N_d}, \quad c_{\mathbf{V},b}\mathbf{I}_{N_d} \le \mathbf{V}_b \le C_{\mathbf{V},b}\mathbf{I}_{N_d}, \quad c_{\mathbf{S},b}\mathbf{I}_{N_d} \le \mathbf{S}_b \le C_{\mathbf{S},b}\mathbf{I}_{N_d}.$$

Under Assumptions (A2), (A4), (A5) *and* (A8), *as* $n\to\infty$ *with probability increasing to* 1

$$c_{\mathbf{V}}\mathbf{I}_{N_d} \le \mathbf{V}_n(\boldsymbol{\lambda}) \le C_{\mathbf{V}}\mathbf{I}_{N_d}, \quad c_{\mathbf{S}}\mathbf{I}_{N_d} \le \mathbf{S}_n(\boldsymbol{\lambda}) \le C_{\mathbf{S}}\mathbf{I}_{N_d} \quad c_{\mathbf{V},b}\mathbf{I}_{N_d} \le \mathbf{V}_{n,b}(\boldsymbol{\lambda}) \le C_{\mathbf{V},b}\mathbf{I}_{N_d}, \quad c_{\mathbf{S},b}\mathbf{I}_{N_d} \le \mathbf{S}_{n,b}(\boldsymbol{\lambda}) \le C_{\mathbf{S},b}\mathbf{I}_{N_d}.$$

Proof. Using Lemma A.7 in [14] and the boundedness of the function b'. □

Define three vectors $\boldsymbol{\Phi}_b$, $\boldsymbol{\Phi}_v$, $\boldsymbol{\Phi}_r$ as

$$\boldsymbol{\Phi}_b = \left(\Phi_{b,J,\alpha}\right)^\top_{0\le k\le d_1, 1\le\alpha\le d_2, 1\le J\le N+1} = -\mathbf{S}_b n^{-1}\sum_{i=1}^{n}\left[b'\{m(\mathbf{T}_i,\mathbf{X}_i)\} - b'\{\bar{m}(\mathbf{T}_i,\mathbf{X}_i)\}\right]\mathbf{B}(\mathbf{T}_i,\mathbf{X}_i),$$

$$\boldsymbol{\Phi}_v = \left(\Phi_{v,J,\alpha}\right)^\top_{0\le k\le d_1, 1\le\alpha\le d_2, 1\le J\le N+1} = -\mathbf{S}_b n^{-1}\sum_{i=1}^{n}\left[\sigma(\mathbf{T}_i,\mathbf{X}_i)\varepsilon_i\right]\mathbf{B}(\mathbf{T}_i,\mathbf{X}_i),$$

and

$$\boldsymbol{\Phi}_r = \left(\Phi_{r,J,\alpha}\right)^{\top}_{0\le k\le d_1, 1\le\alpha\le d_2, 1\le J\le N+1} = \hat{\boldsymbol{\lambda}} - \bar{\boldsymbol{\lambda}} - \boldsymbol{\Phi}_b - \boldsymbol{\Phi}_v.$$

Lemma A.4. *Under Assumptions* (A1)–(A6) *and* (A8), *as* $n \to \infty$,

$$\|\hat{\boldsymbol{\lambda}} - \bar{\boldsymbol{\lambda}}\| = \mathcal{O}_{a.s.}(H^{3/2} + H^{-1/2}n^{-1/2}\ln n), \tag{A.5}$$

$$\|\boldsymbol{\Phi}_r\| = \mathcal{O}_p(H^{-3/2}n^{-1}\ln n), \tag{A.6}$$

$$\|\boldsymbol{\Phi}_b\| = \mathcal{O}_{a.s.}(H^2), \quad \|\boldsymbol{\Phi}_v\| = \mathcal{O}_{a.s.}(H^{-1/2}n^{-1/2}\ln n).$$

Proof. See Online Supplement. □

Lemma A.5. *Under Assumptions* (A1)–(A6) *and* (A8), *as* $n \to \infty$,

$$\left\|\hat{m} - \bar{m}\right\|_{2,n} + \left\|\hat{m} - \bar{m}\right\|_2 = \mathcal{O}_{a.s.}(H^{3/2} + H^{-1/2}n^{-1/2}\ln n),\ \left\|\hat{m} - m\right\|_{2,n} + \left\|\hat{m} - m\right\|_2 = \mathcal{O}_{a.s.}(H^{3/2} + H^{-1/2}n^{-1/2}\ln n).$$

Proof. Lemma A.3 implies

$$\left\|\hat{m} - \bar{m}\right\|_{2,n} + \left\|\hat{m} - \bar{m}\right\|_2 \le 2C_{\mathbf{V}}\|\hat{\boldsymbol{\lambda}}_{\mathbf{g}} - \bar{\boldsymbol{\lambda}}_g\| = \mathcal{O}_{a.s.}(H^{3/2} + H^{-1/2}n^{-1/2}\ln n).$$

The claim follows from the fact that $\|\bar{m} - m\|_\infty + \|\bar{m} - m\|_2 + \|\bar{m} - m\|_{2,n} = \mathcal{O}(H^2)$ by (A.2). □

Proof of Theorem 2. According to (9) and the Mean Value Theorem, a $\bar{m}_{\mathrm{K},1}(x_1)$ between $\hat{m}_{\mathrm{SBK},1}(x_1)$ and $\tilde{m}_{\mathrm{K},1}(x_1)$ exists such that

$$\hat{\ell}'_{m_1}\{\hat{m}_{\mathrm{SBK},1}(x_1), x_1\} - \hat{\ell}'\{\tilde{m}_{\mathrm{K},1}(x_1), x_1\} = \hat{\ell}''_{m_1}\{\bar{m}_{\mathrm{K},1}(x_1), x_1\}\{\hat{m}_{\mathrm{SBK},1}(x_1) - \tilde{m}_{\mathrm{K},1}(x_1)\}.$$

Then according to $\hat{\ell}'_{m_1}\{\hat{m}_{\mathrm{SBK},1}(x_1), x_1\} = 0$, one has

$$\hat{m}_{\mathrm{SBK},1}(x_1) - \tilde{m}_{\mathrm{K},1}(x_1) = -\frac{\hat{\ell}'_{m_1}\{\tilde{m}_{\mathrm{K},1}(x_1), x_1\}}{\hat{\ell}''_{m_1}\{\bar{m}_{\mathrm{K},1}(x_1), x_1\}}.$$

The theorem then follows Lemmas A.15 and A.16 in [11] with small modification including variable **T**. □

Proof of Theorem 3. It follows Theorem 2 and the same proof of Theorem 1 in [25]. □

Proof of Theorem 4. See Online Supplement. □

Appendix B. Supplementary data

gaplmsbk.R: R package containing code to perform SBK estimation for component functions in generalized additive partially linear model available at https://github.com.

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jmva.2017.07.011.

References

[1] C. de Boor, A Practical Guide To Splines, Springer, New York, 2001.
[2] W.K. Härdle, Asymptotic maximal deviation of *M*-smoothers, J. Multivariate Anal. 29 (1989) 163–179.
[3] W.K. Härdle, L. Hoffmann, R. Moro, Learning Machines Supporting Bankruptcy Prediction, in: Cizek, Härdle, Weron (Eds.), Statistical Tools in Finance and Insurance, second ed., Springer, Berlin, 2011.
[4] W.K. Härdle, L. Huang, Analysis of deviance for hypothesis testing in generalized partially linear models, J. Bus. Econom. Statist. (2017) accepted. DOI: http://dx.doi.org/10.1080/07350015.2017.1330693.
[5] W.K. Härdle, E. Mammen, M. Müller, Testing parametric versus semiparametric modelling in generalized linear models, J. Amer. Statist. Assoc. 93 (1998) 1461–1474.
[6] T.J. Hastie, R.J. Tibshirani, Generalized Additive Models, Chapman & Hall, London, 1990.
[7] E. Kolaczyk, Empirical likelihood for generalized linear models, Statist. Sinica 4 (1994) 199–218.
[8] H. Liang, Y. Qin, X. Zhang, D. Ruppert, Empirical-likelihood-based inferences for generalized partially linear models, Scand. J. Stat. 36 (2009) 433–443.
[9] O.B. Linton, J.P. Nielsen, A kernel method of estimating structured nonparametric regression based on marginal integration, Biometrika 82 (1995) 93–100.
[10] R. Liu, L. Yang, Spline-backfitted kernel smoothing of additive coefficient model, Econom. Theory 26 (2010) 29–59.
[11] R. Liu, L. Yang, W.K. Härdle, Oracally efficient two-step estimation of generalized additive model, J. Amer. Statist. Assoc. 108 (2013) 619–631.
[12] S. Ma, R.J. Carroll, H. Liang, S. Xu, Estimation and inference in generalized additive coefficient models for nonlinear interactions with high-dimensional covariates, Ann. Statist. 43 (2015) 2102–2131.
[13] S. Ma, S. Racine, L. Yang, Spline regression in the presence of xategorical predictors, J. App. Econom. 30 (2015) 705–717.
[14] S. Ma, L. Yang, Spline-backfitted kernel smoothing of partially linear additive nodel, J. Statist. Plann. Inference 141 (2011) 204–219.

[15] A. Owen, Empirical Likelihood, Chapman & Hall/CRC, London, 2001.
[16] B. Park, E. Mammen, W.K. Härdle, S. Borak, Time series modelling with semiparametric factor dynamics, J. Amer. Statist. Assoc. 104 (2009) 284–298.
[17] T. Severini, J. Staniswalis, Quasi-likelihood estimation in semiparametric models, J. Amer. Statist. Assoc. 89 (1994) 501–511.
[18] C.J. Stone, Additive regression and other nonparametric models, Ann. Statist. 13 (1985) 689–705.
[19] C.J. Stone, The dimensionality reduction principle for generalized additive models, Ann. Statist. 14 (1986) 590–606.
[20] L. Wang, X. Liu, H. Liang, R.J. Carroll, Estimation and variable selection for generalized additive partial linear models, Ann. Statist. 39 (2011) 1827–1851.
[21] L. Wang, L. Yang, Spline-backfitted kernel smoothing of nonlinear additive autoregression model, Ann. Statist. 35 (2007) 2474–2503.
[22] L. Xue, H. Liang, Polynomial spline estimation for a generalized additive coefficient model, Scand. J. Stat. 37 (2010) 26–46.
[23] L. Xue, L. Yang, Additive coefficient modeling via polynomial spline, Statist. Sinica 16 (2006) 1423–1446.
[24] L. Yang, S. Sperlich, W.K. Härdle, Derivative estimation and testing in generalized additive models, J. Statist. Plann. Inference 115 (2003) 521–542.
[25] S. Zheng, R. Liu, L. Yang, W.K. Härdle, Statistical inference for generalized additive models: Simultaneous confidence corridors and variable selection, TEST 25 (2016) 607–626.

Research Article

Russ A. Moro*, Wolfgang K. Härdle and Dorothea Schäfer

Company rating with support vector machines

DOI: 10.1515/strm-2012-1141
Received October 18, 2012; revised December 1, 2016; accepted March 2, 2017

Abstract: This paper proposes a rating methodology that is based on a non-linear classification method, a support vector machine, and a non-parametric isotonic regression for mapping rating scores into probabilities of default. We also propose a four data set model validation and training procedure that is more appropriate for credit rating data commonly characterised with cyclicality and panel features. Tests on representative data covering fifteen years of quarterly accounts and default events for 10,000 US listed companies confirm superiority of non-linear PD estimation. Our methodology demonstrates the ability to identify companies of diverse credit quality from Aaa to Caa–C.

Keywords: Bankruptcy, company rating, probability of default, support vector machines

MSC 2010: 62-07, 62G05, 62P05

1 Introduction

Banking throughout the world, both central and commercial, is based on credit or trust in the debtor's ability to fulfil obligations. Facing increasing pressure from markets and regulators, banks build their trust to an ever increasing degree on statistical techniques for corporate bankruptcy prediction known as *rating* or *scoring*. Their main purpose is to estimate the financial situation of a company and, if possible, the probability that a company will default on its obligations within a certain period.

Application of statistical models to corporate bankruptcy was made popular after the introduction of discriminant analysis (DA) by Altman [1]. Later the logit and probit models were suggested by Martin [30] and Ohlson [32]. Similar to them is the hazard or survival analysis [15]. All these models belong to the class of Generalised Linear Models (GLM) and could also be interpreted using a latent (score) variable. Their core decision element is a linear score function (graphically represented as a hyperplane in a multidimensional space) separating successful and failing companies. The company score is computed as a value of that function. In the case of the probit and logit models the score is – via a link function – directly transformed into a probability of default (PD). The major disadvantage of these popular approaches is the enforced linearity of the score and, in the case of logit and probit models, the prespecified form of the link function (logit and cumulative Gaussian) between PDs and the linear combination of predictors. For more details about rating models, see [2].

In this paper we introduce an alternative way of assessing a company's creditworthiness. The proposed rating methodology is based on a non-linear classification method, the support vector machine (SVM), and a non-parametric technique for mapping rating scores into probabilities of default (see Section 5). The latter is an indispensable part of our rating methodology because the SVM score, as well as scores produced by

***Corresponding author: Russ A. Moro:** Department of Economics and Finance, Brunel University London, Uxbridge UB8 3PH, United Kingdom, e-mail: russ.moro@brunel.ac.uk
Wolfgang K. Härdle: Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Spandauer Str. 1, 10178 Berlin, Germany, e-mail: stat@wiwi.hu-berlin.de
Dorothea Schäfer: German Institute for Economic Research, Mohrenstr. 58, 10117 Berlin, Germany, e-mail: dschaefer@diw.de

a majority of other statistical techniques, is represented on an arbitrary scale not related to PD necessitating additional calibration. Any attempts of applying parametric techniques for calibrating the score, for example, the same probit or logit transformations, are unsatisfactory since the relationship between the score and PD is usually complex and unknown a priori. We believe that by providing a suitable methodology for mapping credit scores into PDs we can facilitate the adoption of non-parametric statistical learning techniques in credit analysis.

Furthermore, we address the prevailing model validation and testing approaches and raise the question of their applicability to credit rating. It is common to "bundle" credit rating with other data for model comparison without addressing data specific characteristics such as high cyclicality or time dependence, and the presence of strong panel data features [4, 29]. In addition, relatively small and cleaned samples that are used in most cases do not necessarily represent the data that financial institutions encounter in their analysis. Hajek and Michalak [17] summarise the features of seventeen credit rating studies, seven of which apply SVM among other non-parametric techniques. None of them employed a sample larger than 3,900 companies with the median size of only 852 companies. In contrast, our data cover all major US listed companies over the period of fifteen years, i.e. represent almost entire capitalisation of the US stock market.

In particular, we are skeptical about an almost universal adoption of random sampling and cross-validation in the context of company rating. For example, Chen–Shih [7], Lee [26] and Yu–Yao–Wang–Lai [39] perform a random split, while Huang–Chen–Hsu–Chen–Wu [21] and Chen–Li [6] use k-fold validation to create training and validation sets. The structure of credit data is such that a default event is always in the future compared to the financial reports used for model training. However, with cross-validation it is very likely that some defaults to be predicted will happen contemporaneously or prior to the reporting date of other companies selected in the training set. The use of future information for model training raises serious doubts if such an exercise can be called forecasting, especially in presence of default correlations caused by business cycles.

The situation becomes even more exacerbated if we consider a typical design of credit data in which every company is represented several times with its annual or quarterly accounts. Then, with cross-validation, the status of a company will often be predicted with the data for the same company from the future when more information has already been revealed. The same company is likely to appear multiple times both in the training and validation sets with very similar characteristics and identical labels. Since company accounts change relatively slowly, forecasting in such a situation resembles in-sample forecasting and becomes an easy task for a non-linear statistical learning technique such as SVM. The result will likely be overfitting and a lower forecasting accuracy out of time. This also affects the prospects of adopting learning machines as a viable alternative due to understating their performance. We address this issue by proposing a four set test design better reflecting credit rating reality.

We focus on the cross-sectional analysis of the data as opposed to the time series approach (e.g. the Merton's model [31]). This is justified by the fact that company accounts are released only with a quarterly frequency generating an insufficient amount of data for applying a robust time series analysis. Moreover, assumptions of a stochastic process are prone to misspecification [5].

The SVM is based on the principle of a safe separation of solvent and insolvent companies in such a way that the distance between the classes is maximised while misclassifications are penalised [38]. The method allows using kernel techniques [35] and, therefore, non-linear separating surfaces in contrast to classical DA, logit and probit models that rely on linear ones. SVM can be considered as a generalised linear method with input variables mapped to a high-dimensional feature space in which classification is performed. Figure 1 illustrates the qualitative step forward that characterises this paper. The straight line is the linear hyperplane separating solvent and insolvent companies based on a logistic regression. The curve is calculated with an SVM. It is evident that the non-linear separation outperforms the linear one and translates into a better classification performance. An important feature of SVM is also its automatic rather than manual surface shape identification.

Finally, we investigate a relative impact of SVM characteristics responsible for its higher forecasting accuracy. This issue is usually neglected in the literature. When compared with a logistic regression as a benchmark, SVM is different in two major aspects. First, it can map non-linear dependencies. Second, it employs

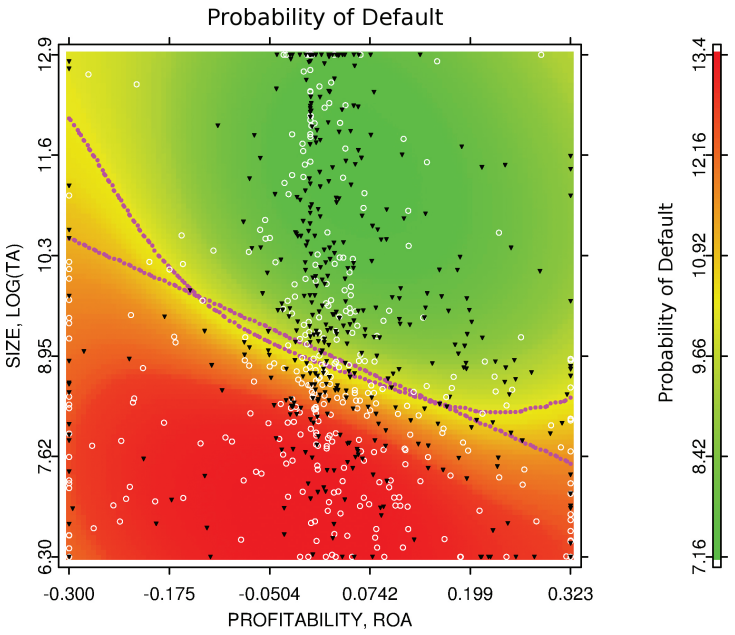


Figure 1. A classification example. The boundary, here corresponding to a 10 % PD, between the classes of solvent (black triangles) and insolvent companies (white rectangles) was estimated using a logistic regression (a straight lines) and a non-linear SVM (a non-linear curve). It is clear that SVM offers superior classification by heavier penalising companies with extremely low or high profitability.

the principle of margin maximisation. To explore their relative importance, we introduce another benchmark, a linear SVM which lacks the non-linearity feature.

Our study has potential implications for supervisory agencies, banks and firms. In tests designed to resemble as close as possible the actual rating procedures, we illustrate that non-linearity in the data significantly influences accuracy. Our assessments show the magnitude of the impact of simplified quantitative models on PD estimates and, therefore, on capital requirements.

The rest of the paper proceeds as follows. The next section describes the non-linear scoring methodology. Data are presented in Section 3 and the test design, model testing and comparison in Section 4. Then a non-parametric technique for calibrating a single firm's score in terms of PD is introduced in Section 5. Finally, Section 6 concludes.

2 Non-linear scoring methodology

When following a linear approach, we automatically impose, through a modelling bias, a monotonic relationship between financial and economic indicators and PDs. A typical example is the monotonically decreasing dependence of PD from liquidity measured as a ratio of cash over total assets (CASH/TA, Figure 2). This ratio characterises the ability of a firm to perform financial transactions quickly and with a low cost. PD is low for highly liquid firms and high for the firms with low liquidity. However, there is a non-monotonic dependence of PD from such important indicators as company size (LOGTA) or sales growth (SG). In the latter case, excessively high growth rates turn out to be as disadvantageous for a firm as low or negative growth rates. This is understandable because extremely high growth rates may be unsustainable and likely indicative of high volatility and, as a consequence, a higher PD. This result is completely in accordance with the Merton's model [31], but we need a non-linear technique to discover such a dependence in the data.

Non-monotonic dependencies in the data were confirmed by Fernandes [13], Manning [28] and Sjur–Hol–van der Wijst [36]; also see a summary in [17], as well as reflected in some commercial models [10, 12]. There is a strong reason to believe that a successful credit rating model must be able to identify and map

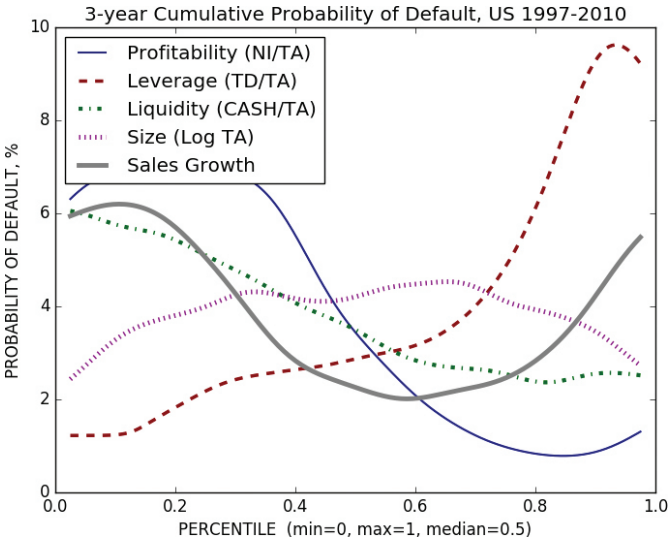


Figure 2. Three year cumulative probabilities of default estimated on quarterly accounts of the US listed companies for 1996–2007 as a function of financial ratios. Here a kernel regression was used.

non-linear dependencies without a priory knowledge about their type, e.g. decreasing, increasing, U-shaped, etc. as evident in Figure 2. Moreover, the model must also be applicable in a high-dimensional case.

SVM, as a non-linear non-parametric statistical learning technique, satisfies all these criteria and has demonstrated very good accuracy in many applications, such as optical character recognition, medical diagnostics, electrical load forecasting, face recognition, high-energy physics, etc. It has as a solution a flexible scoring function and is controlled by adjusting only a few parameters. The SVM solution is stable, i.e. changes slowly in response to a slow change of the data, since the method is based on a convex optimisation problem [37]. This criteria is particularly important for credit rating, where a sudden and unexplained change in output due to switching from one local solution to another can create legal liabilities for a financial institution that performs rating. For SVM software implementation, see [22].

The purpose of classification methods is to separate insolvent ($y = 1$) from solvent ($y = -1$) companies described with a d-dimensional vector of characteristics x, usually financial ratios. The SVM separates the two groups with the maximum distance (margin) between them [35, 38]. The score for x is computed as

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i K(x, x_i) + b.$$

In our case the kernel $K(x, x_i)$ is up to a coefficient, a Gaussian density function with the argument $\|x - x_i\|$ that measures the proximity of an observation x of an unknown class to the observation x_i whose class y_i is known. The closer x and x_i are, the larger $K(x, x_i)$ is; therefore, the score $f(x)$ is primarily defined by the observations that are close to x:

$$K(x, x_i) = e^{\frac{\|x - x_i\|^2}{2r^2}}. \tag{1}$$

The n factors α_i (Lagrange multipliers) are free parameters that are the solution of the classical hinge-loss SVM dual optimisation problem (2) and have a higher magnitude for misclassified observations:

$$\max_{\alpha_i} \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j), \qquad \text{subject to} \qquad 0 \le \alpha_i \le \frac{C}{n}, \quad \sum_{i=1}^{n} \alpha_i y_i = 0. \tag{2}$$

The Gaussian SVM relies on two complexity parameters that have to be calibrated, r in (1) and C in (2). The former is a radial basis coefficient that determines the minimum size of data features that can be replicated with a kernel. The latter one called capacity determines the relative importance of in-sample misclassifications vs. generalisation ability. SVMs with a kernel depending only on the distance $\|x - x_i\|$ are often called radial basis function (RBF) SVMs. For a linear SVM, $K(x, x_i) = x^\top x_i$.

3 Data

For this study we use the company and credit event data collected and maintained by the Risk Management Institute of the National University of Singapore (RMI NUS). The data are available for researchers visiting RMI NUS. At our disposal we have 322,134 quarterly accounts of 10,969 US listed companies covering the period of fifteen years, 1996–2010, excluding the financial and insurance sector and funds. They represent almost the entire capitalisation of the US stock market and are a representative sample of US listed companies.

From them 1,220 companies experienced at least one default event, which in our case is either liquidation under Chapter 7 or restructuring under Chapter 11 of the US Bankruptcy Code. After liquidation a company is delisted and excluded from the sample. We treat companies under Chapter 11 similarly, i.e. a default event has a permanent effect on a firm. This is justified by the fact that firms with a default history are likely to have a repeated default. Therefore, when we talk about a probability of default (PD) we always mean a forward looking conditional probability of default given that a company has never experienced it (for the forward-looking estimation of PD, see [9]).

We constructed financial ratios common in the literature and selected from them only the ones that could be computed at least for 40 % of all accounts in the data. The resulting list contained twenty financial ratios (Table 1). In addition, a new binary default indicator y was generated, which is 1 if default happens within a medium term horizon of three years from the account date, and −1 otherwise. Thus, we adopt a three year cumulative definition of default. We use distress events from 1996–2010 but accounts only from 1996–2007 to ensure that all defaults within every three year post-account period are recorded. The resulting data contain 266,749 quarterly accounts, from which 10,175 received the label $y = 1$, i.e. the company will be in default

Ratio	*N*	mean	std	min	25 %	med	75 %	max
NI/TA	258, 974	−0.86	249	$-1.3 \cdot 10^5$	−0.049	0.003	0.018	$3.2 \cdot 10^3$
NI/S	248, 553	−6.51	253	$-7.0 \cdot 10^4$	−0.181	0.016	0.069	$1.9 \cdot 10^4$
OI/TA	258, 259	−5.60	2,733	$-1.4 \cdot 10^6$	−0.040	0.009	0.029	$2.2 \cdot 10^4$
OI/S	245, 934	−108.08	50,801	$-2.5 \cdot 10^7$	−0.140	0.039	0.112	$4.1 \cdot 10^4$
TL/TA	259, 631	3.19	111	−4,900	0.284	0.504	0.714	$3.7 \cdot 10^4$
TD/TA	260, 925	5.73	477	−6,100	0.015	0.187	0.389	$1.5 \cdot 10^5$
CL/TA	258, 856	2.70	109	−4,900	0.138	0.228	0.380	$3.7 \cdot 10^4$
INT/TD	183, 663	0.34	88	−15	0.014	0.019	0.027	$3.8 \cdot 10^4$
CASH/TA	259, 519	0.158	0.61	−0.45	0.018	0.066	0.199	$2.9 \cdot 10^2$
CASH/CL	259, 445	1.48	18.0	−34	0.061	0.258	0.901	$4.5 \cdot 10^3$
QA/CL	257, 669	3.32	63	−1,500	0.731	1.308	2.544	$2.1 \cdot 10^4$
CA/CL	259, 503	3.89	63	−1,500	1.091	1.874	3.303	$2.1 \cdot 10^4$
WC/TA	258, 598	−2.14	108	−37,000	0.023	0.218	0.450	$4.9 \cdot 10^3$
TA/S	245, 138	53.78	5,777	−10,000	2.544	4.021	7.614	$2.5 \cdot 10^6$
INV/S	242, 733	1.06	35	−5,400	0.017	0.306	0.682	$8.9 \cdot 10^3$
AR/S	244, 589	0.97	27	−160	0.348	0.570	0.776	$1.1 \cdot 10^4$
AP/S	244, 055	3.20	188	−2,900	0.178	0.299	0.519	$8.0 \cdot 10^4$
SG	226, 636	4.39	653	−100	−0.059	0.086	0.290	$2.5 \cdot 10^5$
LOGTA	259, 646	4.52	2.6	−6.9	2.911	4.600	6.287	13.6
LOGS	248, 464	3.14	2.7	−6.9	1.528	3.287	4.977	11.7

Table 1. Summary statistics of the financial ratios estimated for the 1996–2007 quarterly accounts. The average 3 year cumulative default rate is 3.81 %. The financial ratios include four profitability ratios: NI/TA – return on assets (ROA), NI/S – net income margin, OI/TA, OI/S; three leverage ratios: TL/TA, TD/TA, CL/TA; a cost structure ratio: INT/TD – average interest rate; five liquidity ratios: CASH/TA, CASH/CL, QA/CL, CA/CL, WC/TA; four activity ratios: TA/S, INV/S, AR/S, AP/S; a growth indicator: SG – annual sales growth; two size indicators: LOGTA, LOGS. Here the abbreviations are: TA – total assets, NI – net income, S – sales, OI – operating income, TL – total liabilities, TD – total debt, CL – current liabilities, INT – interest rate expense, CASH – cash and cash-like equivalents, QA – quick assets, WC – working capital, INV – inventories, AR – accounts receivable, AP – accounts payable, SG – annual sales growth, LOGTA – natural logarithm of total assets in million dollars, LOGS – natural logarithm of sales in million dollars.

within the next three years. Summary statistics for the twenty financial ratios for 1996–2007 are reported in Table 1.

Almost all financial ratios contain significant outliers or errors which are reported as minimum and maximum. For example, the negative value of the turnover TA/S implies that total assets or sales are negative. Likewise, extremely high maximum profitability margins NI/S are hard to explain. If they are used for model calibration, the result can be a complete misspecification of the model. At the same time, the medians are consistent with our understanding of what a median financial ratio of an established listed company can be. For example, the median interest rate INT/TD is close to 2 %.

To avoid the bias introduced by outliers we cap all variables and subsequently normalise them to ensure that they have the same variance (see Section 4). All preprocessing utilises only the information available up to the moment when the model is calibrated, thus strictly following the out-of-time principle of model testing. This is in contrast to many publications where all data are preprocessed at once with summary statistics capturing the future data distribution.

Extreme outliers need to be capped or cleaned prior to calibration since the SVM, as well as the majority of other statistical learning techniques, are sensitive to outliers. It is often not clear how adequate is data pre-processing used in the literature since it is often not discussed. Some of the applied techniques, such as min-max normalisation [23], are unable to address the problem. We draw attention of our readers to the necessity of outlier capping and normalisation due to the nature of company data, and demonstrate in the next section how this can be done as an integral part of credit rating.

4 Test design and model comparison

Our judgments about model accuracy are based on a widely accepted metric, the accuracy ratio (AR). It will be used for model comparison. AR, that can take any value between 0, when a model has no discriminatory power, and 1, when a model has 100 % accuracy, is a rescaled version of the area under curve (AUC) characteristic, whereas $AR = 2 \cdot AUC - 1$ (see [11]). AUC can be computed as the area under a receiver operating characteristic (ROC) curve such as in Figure 3 (see [20]).

AR (or AUC) has an important property: it is invariant to any strictly monotonic transformation of the score and thus better reflects pure performance of a classification model unaffected by subsequent manipulations with the score. In contrast, popular performance metrics, such as the hit rate, alpha- or beta-errors, are sensitive to the threshold score value separating solvent from insolvent companies. To address the shortcomings of AR as an accuracy measure, we avoid making conclusions about model superiority when ROC curves cross or are not concave [19].

Our principle behind the design of model validation and testing procedures is their proximity to the real life situations in which credit rating models are used by financial institutions. We also address severe deficiencies related to the use of random sampling and cross-validation techniques in the context of credit rating and propose a four data set design that better reflects a time structure of credit data. We apply a four set design in a two-stage out-of-sample out-of-time procedure.

The first stage is model *calibration* and *validation* which consists of estimating model parameters that deliver the highest accuracy on a validation set (VALS) after the model has been calibrated on a calibration set (CALS). For CALS we use quarterly accounts from 1996–1997 and corporate defaults from 1996–2000; and for VALS quarterly accounts from 2001–2002 and corporate defaults from 2001–2005. The accounts are matched for every company and every account receives a label $y = 1$ if a default happens within the next three years and -1 otherwise. If a default has already happened in the past, the account is excluded from our analysis. Finally, we exclude observations with missing values. The longer period for defaults extending three years after the last account is required to capture all defaults with the maximum horizon of three years and to avoid a bias in PD estimation. With this design VALS is strictly out-of-time compared to CALS.

All variables are capped or winsorised with the values $Q1 - 1.5 \cdot IQR$ as minimum and $Q3 + 1.5 \cdot IQR$ as maximum, where Q1 is the first quartile or 25 % percentile, Q3 is the third quartile or 75 % percentile and

IQR = Q3 − Q1 is the interquartile range. After capping we normalise all variables by subtracting their median and dividing by their standard deviation that was estimated on the capped data. This brings all variables to the same scale avoiding excessive influence of some. Strictly following the out-of-time principle of testing, we estimate Q1, Q3, median and standard deviation only on CALS and use them both for companies in CALS and VALS. This is essential because in reality it is common to rate only few companies in the rating period, for which a reliable evaluation of summary statistics is impossible.

Model calibration includes variable selection and estimation of the model complexity parameters C and r for a non-linear RBF SVM and C for a linear SVM. Since it is practically impossible to try all combinations of variables in order to choose the one that yields the overall best performance, we apply a forward selection procedure. Variable selection is based on a logistic regression. We use the same set of variables with all models to ensure that any improvement in performance is due to the model and can not be attributed to a difference in variables.

First we try all univariate models using one of the twenty variables reported in Table 1. The variable that demonstrates the highest AR is selected. Next, all bivariate models are tried that include the previously selected variable, and the second variable is identified. The process continues until all twenty variables have been included in the final model. A combination of variables that has demonstrated the highest AR is retained. In the order of their selection, this combination of variables is

(1) TL/TA, leverage,
(2) NI/TA – returns on assets, profitability,
(3) CASH/TA, liquidity,
(4) INT/TD – average interest rate, cost structure,
(5) OI/TA, profitability,
(6) LOGTA, size,
(7) CL/TA, leverage,
(8) NI/S – profit margin, profitability,
(9) OI/S, profitability,
(10) AR/S, activity.

The ten selected variables represent the six main groups of financial ratios. TL/TA captures the total level of indebtness and CL/TA the level of debt with the maturity of one year or less, in other words the debt that needs to be refinanced soon. NI/TA, OI/TA, NI/S and OI/S are various profitability indicators measuring the ability of capital to generate profits and the share of profits in gross sales. CASH/TA characterises the cash position of a company. INT/TD or the average interest rate determines the cost of external financing. LOGTA is the company size measured as the logarithm of total assets (the use of logarithm is essential for avoiding significant heteroscedasticity caused by the presence of very large and relatively small companies in the data). AR/S characterises the ability of a company to collect its payments.

The maximum accuracy ratio (AR) achieved by the logistic regression with these ten variables during the selection procedure is 54.7. It rapidly increases when the first variables are added and decreases at a lower rate after the maximum is passed. The dependence of AR from the number of selected inputs is flat near its maximum: all models having between four and fourteen and between six and twelve input variables have AR higher than 53.0 and 54.0, respectively.

Once the model inputs have been selected, we proceed with the calibration of the parameters C and r that appear in the formulation of a non-linear SVM. They are estimated on a 12×12 grid. The next value of C is twice as large as the previous one and the next value of r is 1.5 times as large. After the initialisation of the grid, if the solution is not achieved on its internal nodes but on the boundary, the grid is shifted in the direction of that boundary until the solution becomes internal. As the solution we take the combination of C and r that deliver the highest AR on VALS. The procedure is similar for a linear SVM when only one parameter C needs to be calibrated. The grid has only one dimension in this case. The highest AR = 61.1 is achieved with a non-linear SVM for $C = 2{,}560$ and $r = 1.51$. For a linear SVM the maximum AR = 53.3 is achieved with $C = 200$. Too high complexities when C is high or r is small lead to overfitting and low generalisation ability. Both situations should be avoided.

	Set-up 1	Set-up 2	Set-up 3
Validation	TRAS X-Val	CALS-VALS	CALS-VALS
Testing	TRAS-TESS	TRAS-TESS	TRAS*-TESS
Logit	46.0	46.0	50.9
Linear SVM	49.4	51.1	52.2
Non-linear SVM	55.2	60.2	61.3

Table 2. Forecasting accuracy for the three tested models, a logistic regression, linear and non-linear SVM, demonstrated in different set-ups. CALS contains the data from 1996–2000 with 27,758 accounts, from which 621 belong to companies expecting default; VALS and TRAS from 2001–2005 with 27,274 (1,473) accounts; an enlarged training set (TRAS*) from 1996–2005 with 105,056 (5,323) accounts; TESS from 2006–2010 with 27,046 (1,284) accounts. All differences between models are significant at a ≤5 % level according to the Mann-Whitney test. The difference in AR that corresponds here to a 5 % significance level is approximately 0.5 or less.

The second stage of our testing procedure is model training and testing. In real applications the training set (TRAS) can contain any information available until the present moment, it can include both CALS and VALS as well. For our TRAS we choose company quarterly accounts from 2001–2002 matched with the default events from 2001–2005 following the same procedure described for calibration and validation. In this case the TRAS is the same as the VALS but this does not have always to be the case. The testing set (TESS) is composed of the accounts from 2006–2007 and default events from 2006–2010. The ten selected variables are capped, normalised and the missing values excluded following the same protocol as for the pair CALS–VALS. The number of accounts in every data set after cleaning missing values is reported in the caption of Table 2. The purpose of having a special TRAS different from CALS is to perform training as close as possible to the date when the trained model is applied to TESS, for minimising an inevitable change in data distribution.

After having validated the models using the pair CALS–VALS, we train the models on TRAS and evaluate their performance on TESS. The accuracies are reported in the second column in Table 2 (Set-up 2). We can observe a significant increase in accuracy of a non-linear SVM to AR = 60.2 from AR = 46.0 for Logit. A linear SVM also performs better than Logit demonstrating AR = 51.1. All differences are significant at a ≤ 5 % level according to the Mann–Whitney test [27]. The receiver operating characteristic (ROC) curves for the three models are presented in Figure 3. A non-linear SVM clearly dominates the other models. The comparison between a linear SVM and Logit cannot be made decisively since their ROC curves cross.

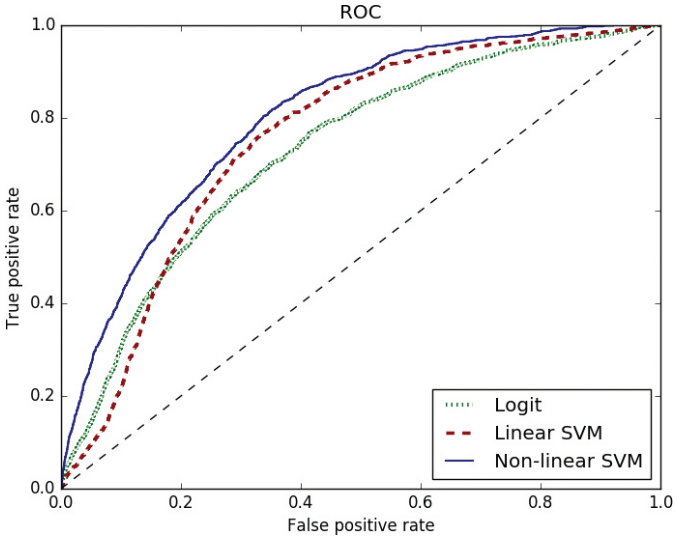


Figure 3. ROC curves for a logistic regression, linear and non-linear SVM. The superiority of a non-linear SVM is evident: its ROC curve is above the ROC of the other two models. The advantage of a linear SVM compared to a logistic regression is less clear since their ROC curves cross.

Next, we compare our test design based on four data sets with a common in the literature design when a model is cross-validated on TRAS. A cross-validation procedure is out-of-sample with respect to individual accounts, but it is not out-of-time. Moreover, it is not out-of-sample with respect to companies. We expect this to lead to a biased estimation of model parameters and overfitting due to cyclicality and a panel nature of the data. To demonstrate the shortcomings of the widely adopted approach, we cross-validate all models on TRAS, using here a 5-fold cross-validation. Subsequently, the models are trained on TRAS and tested on TESS. The results are reported in the first column of Table 2 (Set-up 1). As expected, both non-linear and linear SVM are affected by overfitting and their generalisation ability is impaired, which is evident from a lower AR compared to the second column. The AR of a non-linear SVM drops from 60.2 to 55.2, and of a linear SVM from 51.1 to 49.4.

In the final set-up, we extend the composition of TRAS, which now includes all accounts from 1996–2002 matched with defaults from 1996–2005. A substantially larger size of TRAS results in a better forecasting accuracy for all three models. Their relative ranking is, however, preserved (see the last column in Table 2). The performance of a linear SVM is closer to the performance of a logistic regression than a non-linear SVM. This confirms our hypothesis that it is the non-linearity feature that is mostly responsible for a higher accuracy.

In all tests we observe a significant superiority of a non-linear SVM, such as in the last test (AR = 61.3 vs. 50.9 for a logistic regression and 52.2 for a linear SVM). However, in contrast to a logistic regression that does not have any complexity parameters to be calibrated, a non-linear RBF SVM has two parameters and can be overfitted if the calibration is based on a design that is drastically different from the one applied for testing, for example, cross-validation for calibration vs. out-of-time forecasting for testing. Therefore, for SVMs and other learning machines with a variable complexity, it is important to employ a similar out-of-time design for determining model complexity parameters resembling the out-of-time set-up in which the models are used in practice. Failure to do so will likely penalise SVMs and other learning machines and hinder their acceptance as a viable credit rating technique. On the contrary, the application of a correctly calibrated non-linear SVM instead of linear alternatives would allow issuing more credit without increasing risk because of a better separation between solvent and insolvent companies.

5 Conversion of scores into PDs

The conversion of rating scores into PDs provides us with a link to the existing rating classes reported by rating agencies such as Moody's and S&P. In a logistic model a sigmoid function is used to estimate PD assuming a logistic distribution of the latent variable. The same model is also applicable for converting an SVM score into PD [33]. Other methodologies based on discriminant analysis or binning are different from the logistic regression in assuming a Gaussian or piecewise uniform distribution, respectively. Such assumptions, however, are often not compatible with reality. For example, a company score computed with an SVM has arbitrary scaling not related to PD, it only ranks companies from the least to the most prone to default. Moreover, some techniques such as binning can produce non-monotonic PDs as a function of the score.

With Bayesian inference, which can be applied to SVM and other learning machines, PDs can be derived inside the modelling framework through posterior class probabilities [14, 24, 25, 34]. Although the posterior distribution is not explicitly postulated in this case, the prior distribution of the model parameters is, usually as Gaussian. This also imposes restrictions, which can be unrealistic, on the relationship between PDs and scores.

In order to calibrate scores in terms of PDs we propose using an isotonic regression based on the Pool Adjacent Violators Algorithm (PAVA). It does not specify a functional dependence a priori, only imposing a monotonicity constraint, which is a desirable feature [3]; also see [8] for a software implementation of the algorithm.

For data $\{x_i, y_i\}_{i=1}^n$ with $x_1 \le x_2 \le \cdots \le x_n$ the PAVA produces a non-decreasing function $f(x)$, i.e.

$$f(x_1) \le f(x_2) \le \cdots \le f(x_n),$$

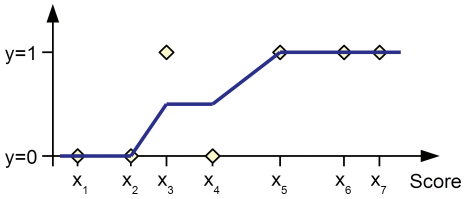


Figure 4. Calibration of the score in terms of PD with the pool adjacent violators algorithm (PAVA) [3]. The dots correspond to solvent (here y=0) or approaching insolvency (y=1) companies. The line denotes the estimated PD.

that solves the following problem:

$$\hat{f} = \arg\min_{f \in \mathbb{R}^n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 \qquad \text{subject to} \qquad x_i \le x_j, \quad f(x_i) \le f(x_j) \quad \text{for all } i \le j. \tag{3}$$

The solution of this problem is pooling (averaging) adjacent observations that violate monotonicity. The PAVA name comes from this procedure. The algorithm is fast and scales with the rate n (see [16]). Here $\{x_i\}_{i=1}^n$ are the scores computed with SVM or any other technique for the accounts in TRAS and $\{y_i\}_{i=1}^n$ are the labels of those accounts, however with a different encoding: $y = 1$ if a company experiences a default within the next three years and $y = 0$ if not. With this labelling the output $\hat{f}_i$, $i = 1, 2, \ldots, n$, of PAVA is PDs for the accounts in TESS, i.e. $\widehat{\mathrm{PD}}_i = \hat{f}_i$.

After the score has been converted into PD for TRAS, the final stage of our analysis is the estimation of PD of any company in TESS that has been rated and received a score x. We do this by looking up the closest value to x among the scores of the companies in TRAS $\{x_i\}_{i=1}^n$ and using PD of the company with the closest score:

$$\widehat{\mathrm{PD}}(x) = \widehat{\mathrm{PD}}(x_i) \quad \text{subject to} \quad x_i = \arg\min_{i=1,\ldots,n} |x - x_i|.$$

Figure 5 represents PDs estimated for the companies of the testing set for the scores reported along the horizontal axis as percentiles. The scores were produced with a non-linear SVM trained on the 2001–2005 TRAS, and calibrated on the CALS–VALS pair including years 1996–2000 and 2001–2006, respectively. Since extreme scores estimated by PAVA tend to be biased (it is easy to see from expression (3) that the minimum PD can be 0 and the maximum 1), we capped PDs at 1 % and 99 % percentiles of TRAS. The horizontal lines correspond to the historical average three-year cumulative default rates for the whole letter Moody's rating classes. Our methodology allows rating companies in a wide range of rating classes with the minimum PD around 0.00 % corresponding to the top Aaa class and the highest PD exceeding 25 % which corresponds to the lowest Caa–C rating classes preceding default, i.e. it covers the whole spectrum of rating classes.

6 Conclusion

In this paper we introduce a rating methodology based on a non-linear SVM combined with a non-parametric isotonic regression for mapping scores into probabilities of default (PD). The latter is essential, since the absolute value of the SVM score has no relationship with PD. It also can be used in combination with other statistical learning techniques, e.g. neural networks.

We confirm that a rating model based on a non-linear RBF SVM dominates both traditional linear parametric methods such as a logistic regression and a linear SVM. Most of the improvement in forecasting accuracy can be attributed to its non-linearity feature. The improvement is significant. For example, in terms of AR, 60.2 for a non-linear SVM vs. 51.1 for a linear SVM and 46.0 for a logistic regression.

However, in order to achieve high performance, the complexity parameters of a non-linear SVM must be correctly calibrated in a set-up that resembles the actual rating procedure, i.e. it must be out-of-sample and out-of-time. We show, that random sampling and cross-validation procedures routinely applied to calibrating model parameters and not satisfying the out-of-time criterion are likely to lead to overfitting and deterioration

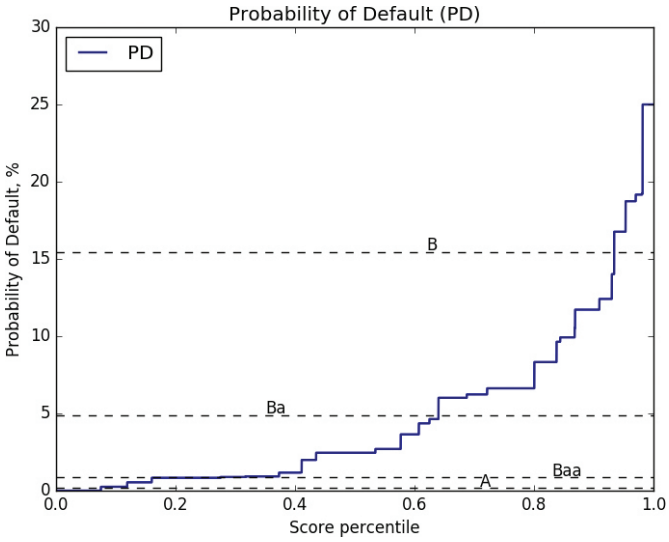


Figure 5. Non-linear SVM scores for the companies of TESS calibrated in terms of PD. The horizontal lines represent three-year cumulative average historical PDs for the Moody's rating classes [18].

of generalisation ability, and as a consequence, lower accuracy. For example, when an out-of-time calibration procedure was replaced with cross-validation, the AR of a non-linear SVM dropped from 60.2 to 55.2. We suggest a four data set testing procedure, whereas the calibration and validation sets used to estimate the complexity parameters resemble in their set-up the training and testing set pair used to evaluate the performance of a model.

Combined with a technique for mapping scores into PDs and correctly calibrated using the four data design, a non-linear SVM can be a potent alternative to other rating methodologies. The property of SVM to automatically determine the type of dependence between the input financial indicators and PD is an additional advantage. Combined with a higher accuracy, it can help to reduce the costs of rating and, as a result, lead to offering rating services in the areas where it was too costly earlier, for example for smaller companies, partially replacing human judgement.

Acknowledgment: We are grateful to Jin-Chuan Duan and Oliver Chen of RMI NUS and Laura Auria and Ralf Körner of the Deutsche Bundesbank for their irreplaceable assistance and valuable suggestions. We also would like to thank participants of the International Risk Management Conferences for their comments and discussion.

Funding: The authors gratefully acknowledge the support of this project by the Risk Management Institute of the National University of Singapore (RMI NUS). We thank RMI NUS for providing access to their database of financial statements and default events. Russ A. Moro was financially supported by the RMI NUS and Wolfgang K. Härdle by Deutsche Forschungsgemeinschaft through SFB 649 "Economic Risk".

References

[1] E. Altman, Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, *J. Finance* **23** (1968), no. 4, 589–609.

[2] E. Altman and A. Saunders, Credit risk measurement: Developments over the last 20 years, *J. Banking Finance* **21** (1998), no. 11–12, 1721–1742.

[3] R. E. Barlow, J. M. Bartholomew, J. M. Bremmer and H. D. Brunk, *Statistical Inference Under Order Restrictions*, John Wiley & Sons, New York, 1972.

[4] E. Carrizosa, A. Nogales-Gómez and D. R. Morales, Strongly agree or strongly disagree? Rating features in support vector machines, *Inform. Sci.* **329** (2016), 256–273.
[5] A. H. Chen, N. Ju, S. C. Mazumdar and A. Verma, Correlated default risks and bank regulations, *J. Money Credit Banking* **38** (2006), 375–398.
[6] C.-C. Chen and S.-T. Li, Credit rating with a monotonicity-constrained support vector machine model, *Expert Syst. Appl.* **41** (2014), no. 16, 7235–7247.
[7] W.-H. Chen and J.-Y. Shih, A study of Taiwan's issuer credit rating systems using support vector machines, *Expert Syst. Appl.* **30** (2006), no. 3, 427–435.
[8] J. de Leeuw, K. Hornik and P. Mair, Isotone optimization in R: Pool-adjacent-violators algorithm (PAVA) and active set methods, *J. Stat. Softw.* **32** (2009), no. 5, 1–24.
[9] J.-C. Duan, J. Sun and T. Wang, Multiperiod corporate default prediction – A forward intensity approach, *J. Econometrics* **170** (2012), no. 1, 191–209.
[10] D. W. Dwyer, A. E. Kocagil and R. M. Stein, Moody's KMV RiskCalc v3.1 model: Next-generation technology for predicting private firm credit risk, White Paper (2004), Moody's KMV Company.
[11] B. Engelmann, E. Hayden and D. Tasche, Measuring the discriminative power of rating systems, Discussion Paper Series 2: Banking and Financial Supervision, Deutsche Bundesbank, 2003.
[12] E. Falkenstein, A. Boral and L. Carty, Riskcalc for private companies: Moody's default model, Report 56402, Moody's Investors Service, New York, 2000.
[13] J. E. Fernandes, Corporate credit risk modeling: Quantitative rating system and probability of default estimation, preprint (2005), http://econwpa.repec.org/eps/fin/papers/0505/0505013.pdf.
[14] T. V. Gestel, J. A. K. Suykens, G. Lanckriet, A. Lambrechts, B. D. Moor and J. Vandewalle, Bayesian framework for least-squares support vector machine classifiers, Gaussian processes, and kernel Fisher discriminant analysis, *J. Neural Comput.* **14** (2002), no. 5, 1115–1147.
[15] D. Glennon and P. Nigro, Measuring the default risk of small business loans: A survival analysis approach, *J. Money Credit Banking* **37** (2005), no. 5, 923–947.
[16] S. J. Grotzinger and C. Witzgall, Projections onto simplices, *Appl. Math. Optim.* **12** (1984), no. 1, 247–270.
[17] P. Hajek and K. Michalak, Feature selection in corporate credit rating prediction, *Knowledge Based Syst.* **51** (2013), 72–84.
[18] D. T. Hamilton and R. Cantor, Measuring corporate default rates, Report 100779, Moody's Investors Service, 2006.
[19] D. Hand, Measuring classifier performance: A coherent alternative to the area under the ROC curve, *Mach. Learn.* **77** (2009), no. 1, 103–123.
[20] J. A. Hanley and B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology* **143** (1982), no. 1, 29–36.
[21] Z. Huang, H. Chen, C.-J. Hsu, W.-H. Chen and S. Wu, Credit rating analysis with support vector machines and neural networks: A market comparative study, *Decis. Support Syst.* **37** (2004), no. 4, 543–558.
[22] A. Karatzoglou, D. Meyer and K. Hornik, Support vector machines in R, *J. Stat. Softw.* **15** (2006), no. 9, 1–28.
[23] K. Kim and H. Ahn, A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach, *Comput. Oper. Res.* **39** (2012), no. 8, 1800–1811.
[24] J. T.-Y. Kwok, Moderating the outputs of support vector machine classifiers, *IEEE Trans. Neural Networks* **10** (1999), no. 5, 1018–1031.
[25] J. T.-Y. Kwok, The evidence framework applied to support vector machines, *IEEE Trans. Neural Networks* **11** (2000), no. 5, 1162–1173.
[26] Y.-C. Lee, Application of support vector machines to corporate credit rating prediction, *Expert Syst. Appl.* **33** (2007), no. 1, 67–74.
[27] H. B. Mann and D. R. Whitney, On a test of whether one of two random variables is stochastically larger than the other, *Ann. Math. Stat.* **18** (1947), no. 1, 50–60.
[28] M. J. Manning, Exploring the relationship between credit spreads and default probabilities, Working Paper 225, Bank of England, 2004.
[29] D. Martens, B. Baesens, T. van Gestel and J. Vanthienen, Comprehensible credit scoring models using rule extraction from support vector machines, Working Paper 878283, Social Science Research Network, 2006.
[30] D. Martin, Early warning of bank failure: A logit regression approach, *J. Banking Finance* **1** (1977), no. 3, 249–276.
[31] R. C. Merton, On the pricing of corporate debt: The risk structure of interest rates, *J. Finance* **29** (1974), no. 2, 449–470.
[32] J. A. Ohlson, Financial ratios and the probabilistic prediction of bankruptcy, *J. Accounting Res.* **18** (1980), no. 1, 109–131.
[33] J. C. Platt, Probabilities for SV machines, in: *Advances in Large Margin Classifiers*, The MIT Press, Cambridge (2000), 61–73.
[34] M. D. Richard and R. P. Lippmann, Neural network classifiers estimate Bayesian a posteriori probabilities, *J. Neural Comput.* **3** (1991), no. 4, 461–483.
[35] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, London, 2002.

[36] W. Sjur, S. Hol and N. van der Wijst, Capital structure, business risk, and default probability, Discussion Paper 975302, Social Science Research Network, 2007.
[37] V. M. Tikhomirov, The evolution of methods of convex optimization, *Amer. Math. Monthly* **103** (1996), no. 1, 65–71.
[38] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.
[39] L. Yu, X. Yao, S. Wang and K. K. Lai, Credit risk evaluation using a weighted least squares svm classifier with design of experiment for parameter selection, *Expert Syst. Appl.* **12** (2011), 15392–15399.

International Journal of Theoretical and Applied Finance

Sieve estimation of the minimal entropy martingale marginal density with application to pricing kernel estimation*

Denis Belomestny
Duisburg-Essen University, Faculty of Mathematics
Thea-Leymann-Str. 9 D-45127 Essen, Germany
and
National Research University Higher School of Economics
Shabolovka, 26, 119049 Moscow, Russia
denis.belomestny@uni-due.de

Wolfgang Karl Härdle
C.A.S.E-Center for Applied Statistics and Economics
Humboldt-Universität zu Berlin
D-10178 Berlin, Germany
haerdle@wiwi.hu-berlin.de

Ekaterina Krymova
Duisburg-Essen University, Faculty of Mathematics
Thea-Leymann-Str. 9 D-45127 Essen, Germany
and
IITP RAS, Moscow
ekaterina.krymova@uni-due.de

Received (Day Month Year)
Revised (Day Month Year)

We study the problem of non-parametric estimation of the risk-neutral densities from options data. The underlying statistical problem is known to be ill-posed and needs to be regularised. We propose a novel regularised empirical sieve approach for the estimation of the risk-neutral densities which relies on the notion of the minimal martingale entropy measure. The proposed approach can be used to estimate the so-called pricing kernels which play an important role in assessing the risk aversion over equity returns. The asymptotic properties of the resulting estimate are analysed and its empirical performance is illustrated.

Keywords: Pricing Kernel; Risk neutral Density; Ill-Posed Problems; Kullback-Leibler Divergence

*D.B. acknowledges the financial support from the Russian Academic Excellence Project "5-100". E.K. acknowledges the financial support from the Deutsche Forschungsgemeinschaft (DFG) through the SFB 823 "Statistical modelling of nonlinear dynamic processes" and from the RF President grant MK-9662.2016.9.

1. Introduction

A pricing kernel is defined as a ratio of the economic risk which contains the preferences of investors and statistical risk which provides information on the dynamics of the data generating process (DGP). The pricing kernel is an important link between economics and finance and it plays a pivotal role in assessing the risk aversion over equity returns. The economic risk is usually approximated using the risk neutral density q obtained from the derivative market. Thus, obtaining an accurate estimator of q is a crucial step for pricing kernel estimation. Estimating pricing kernels from option prices is, for example, discussed in Aït-Sahalia and Lo (2000), Aït-Sahalia and Duarte (2003), Jackwerth (2000). We refer to a recent work by Song and Xiu (2016) for a comprehensive list of references. In particular, Aït-Sahalia and Lo (2000) present several methods to estimate the risk neutral density q by using different nonparametric methods. Härdle et al. (2013) developed uniform confidence bands for pricing kernels. The developed theory is helpful for testing parametric specifications of pricing kernels and has a direct extension to estimating risk aversion patterns. Golubev et al. (2013) and Beare and Schmidt (2012) proposed statistical tests of pricing kernel monotonicity. Beare (2011) showed how the theory of monotone rearrangements may be used to derive an explicit solution for the cost minimizing measure preserving derivative written on some underlying asset.

In this paper, we propose to estimate the pricing kernel nonparametrically by a kind of sieve empirical minimisation with a penalty involving the ratio of the risk-neutral density estimator and the subjective density estimator. In particular, the risk-neutral density is approximated by a weighted kernel density estimate with varying unknown weights for different observations. By observing stock prices or returns that investors expect to obtain at time to maturity, the subjective density can be approximated by kernel density estimate of historical stock prices with equal weights. We represent the European call option price function by the second order integration of the risk-neutral density, so that the unknown weights are obtained through one-step penalised least squares estimation with the Kullback-Leibler divergence as the penalty function. Statistical risk provides an overview over statistical properties of the DGP and is given by the distribution p of future prices conditional on current prices also known as historical density. The historical density p can be estimated using the past of the time series of the underlying stock (S_t). Due to the large number of observations in the derivative option market, the risk neutral density q can be well estimated with large-sample asymptotic properties given.

Let $C(t; K, T)$ be the price, at time t, of a call option of strike K and maturity T on an underlying asset that trades at time t for the price (S_t). For the sake of simplicity we restrict our attention to the case of zero interest rates and dividend yields. In the absence of arbitrage between options, stocks, and bonds at each maturity one may deduce from standard arguments the existence of a risk neutral density

for the stock price at future time T, $q(S,T)$ such that

$$C(t;K,T) = \int_K^\infty (S-K)\, q(S,T)\, dS. \tag{1.1}$$

In fact we may identify the risk-neutral density from option prices at time t using the Breeden and Litzenberger (1978) formula which reads (if interest rates and the dividends are zero) as

$$q(S,T) = \left.\frac{\partial^2}{\partial K^2}\right|_{K=S} C(t;K,T). \tag{1.2}$$

Note that given a finite number of strikes K for which options are available on the market, the problem of estimating the risk neutral density q becomes ill-posed, i.e., a small perturbation in C can lead to a big change in q. It follows from (1.2) that all the densities in the family of risk-neutral densities have a constant expectation:

$$S_t = \int_0^\infty S\, q(S,T)\, dS. \tag{1.3}$$

Moreover the results of Rothschild and Stiglitz Rothschild and Stiglitz (1970, 1971, 1973) and Kellerer (1972) imply that if the family of distributions $q(S,T)$ with a constant expectation has the following *positive calendar spread* property

$$C(t;K,T_1) \le C(t;K,T_2), \quad T_1 < T_2, \tag{1.4}$$

then there exists a martingale $(S_t^M)_{t\ge 0}$ such that for all T the density of S_T^M is $q(S,T)$. By the fundamental theorem of asset pricing, choosing an arbitrage-free pricing method is basically equivalent to choosing a martingale measure $\mathbb{Q} \sim \mathbb{P}$, where $\mathbb{P}$ is the *historical* measure. Unfortunately the martingale measure is not unique in most cases and one needs to have a criteria to find $\mathbb{Q}$. This is typically done by solving an optimization problem. A widely studied family of objective functions for choosing $\mathbb{Q}$ consists of criteria which can be expressed in the form

$$J_f(\mathbb{Q}) = \mathbb{E}^{\mathbb{Q}}\left[f\left(\frac{d\mathbb{Q}}{d\mathbb{P}}\right)\right]. \tag{1.5}$$

An important example of the deviation measures described above is the *relative entropy*

$$\mathcal{E}(\mathbb{Q},\mathbb{P}) = \mathbb{E}^{\mathbb{Q}}\left[\log\frac{d\mathbb{Q}}{d\mathbb{P}}\right] = \mathbb{E}^{\mathbb{P}}\left[\frac{d\mathbb{Q}}{d\mathbb{P}}\log\frac{d\mathbb{Q}}{d\mathbb{P}}\right] \tag{1.6}$$

which corresponds to $f(x) = \log(x)$. Given a stochastic model S_t, the *minimal entropy martingale* model is defined as a martingale S_t^* such that the law $\mathbb{Q}^*$ of S^* minimizes the relative entropy with respect to $\mathbb{P} = \mathbb{P}^S$ among all martingale processes: it minimizes the relative entropy under the constraint of being a martingale. Since we have in our disposal only a family of densities $\mathbb{D} = \{q(X,t),\, t>0\}$ (only information about marginals can be retrieved from vanilla options) it is reasonable to define *minimal entropy martingale marginal* model as a martingale S_t^{*M} such that for each t the law of S_t^{*M} is given by the density $q(M,t)$ and the measure $\mathbb{Q}_M$

with marginals from $\mathbb{D}$ minimizes the relative entropy to $\mathbb{P}^S$. The minimal entropy martingale model has an information theoretic interpretation: minimizing relative entropy corresponds to choosing a martingale measure (or marginals martingale measure) by adding the least amount of information to the prior model.

To take into account the prices of derivative products traded in the market, Kallsen (2002) has introduced a notion of consistent pricing measure, that is, a measure that correctly reproduces the market-quoted prices for a given number of derivative products. Following this line and taking into account the formulas (1.2) and (1.3), we can formulate the calibration problem as one of finding the family of densities $q(S,T)$ such that for each T the density $q(\cdot,T)$ minimizes the functional

$$\mathcal{Q}(q) := \int_0^\infty |C(t;K,T) - (Aq)(K)|^2 \, d\mu(K) + \alpha \, \mathrm{KL}(p||q) \tag{1.7}$$

and the integral $\int_0^\infty q(x,T)\,dx$ is constant for all T. In (1.7) p is a prior density estimated from a historical time series which may depend on t, A is a linear operator of second order integration (see (1.1)), μ is a finite measure on $\mathbb{R}_+$ and $\mathrm{KL}(p||q)$ is the *Kullback-Leibler divergence* between p and q, i.e.,

$$\mathrm{KL}(p||q) := \int_{\mathbb{R}} p(x) \log \frac{p(x)}{q(x)} \, dx. \tag{1.8}$$

In this paper we propose an efficient way of solving the optimisation problem (1.7) and study the properties of the corresponding estimator. We then apply our estimation procedure to the well-known pricing kernel problem. From statistical point of view, the penalised least squares estimate coming from (1.7) gains in numerical stability and can retain some desirable properties of the historical density p.

The paper is organized as follows. In Section 2 and 3 we formulate the problem and construct the estimators. In Section 4 and 5 we present the main and auxiliary theoretical results. In Section 6, we introduce KL-divergence of log-normals. Section 7 provides empirical results of a simulation study. The proposed estimation procedure is illustrated by analyzing a Strike-Call price dataset in Section 8.

2. Problem formulation and main results

Let $(K_1,C_1),\dots,(K_m,C_m)$ be observed pairs of strikes and the corresponding call option prices. Let also $X_1,\dots,X_n$ be identically distributed random variables distributed with a density $p(x)$. First, we construct a kernel density estimator p_n for p as

$$p_n(x) := \frac{1}{nh} \sum_{i=1}^n \mathcal{K}\left(\frac{x-X_i}{h}\right), \tag{2.1}$$

where K is a kernel and $h = h(n) > 0$ is a bandwidth. Next we define a class of estimators for the risk-neutral density q via

$$q_n(x;W) := \frac{1}{h} \sum_{i=1}^n w_i \, \mathcal{K}\left(\frac{x-X_i}{h}\right), \tag{2.2}$$

where the weights $W = \{w_i\}_{i=1}^n$ are nonnegative and sum to one.

Remark 2.1. A motivation for the weighted kernel density estimate (2.2) comes from some well known results on penalised density estimation (see, for example, Eggermont and LaRiccia (2001)). Indeed, consider the following optimization problem for q :

$$\begin{aligned} &\text{minimize} \quad -2\int_{\mathbb{R}} \omega(x)\, q(x)\, dP_n(x) + \int_{\mathbb{R}} |q(x)|^2\, dx + h^2 R(q) \\ &\text{subject to} \quad q \text{ is a continuous density,} \end{aligned} \tag{2.3}$$

where ω stands for the Radon-Nikodym derivative dQ/dP, R is the roughness penalization term and h is the smoothing parameter. Under the choice

$$R(q) = \int_{\mathbb{R}} |q'(x)|^2\, dx, \tag{2.4}$$

the solution q of (2.3) satisfies (see, e.g., Eggermont and LaRiccia (2001)) the boundary value problem

$$\begin{aligned} -h^2 q'' + q = \omega\, dP_n(x), \quad -\infty < x < \infty \\ q(x) \to 0, \qquad |x| \to \infty \end{aligned} \tag{2.5}$$

and is given by

$$q_n(x) = \frac{1}{nh} \sum_{i=1}^n \omega(X_i) \mathcal{K}\left(\frac{x - X_i}{h}\right) \tag{2.6}$$

provided that $\frac{1}{n}\sum_{i=1}^n \omega(X_i) = 1$ and $\mathcal{K}(\cdot)$ is a two-sided exponential kernel.

Now by plugging the estimator (2.2) into (1.7) and optimising over the weights W, we get an estimate for the risk-neutral density q, which is expected to be close to some solution of (1.7). Further, for a fixed sample $X_1, \ldots, X_n$ from the distribution with density p, we approximate the solution of the minimisation problem (1.7) by

$$q_{n,m} := \operatorname{argmin}_{q \in \mathcal{C}_{n,X}} \mathcal{Q}_{n,m}(q), \tag{2.7}$$

where

$$\mathcal{C}_{n,X} := \left\{ \sum_{i=1}^n w_i \mathcal{K}_h(x - X_i), \quad \sum_{i=1}^n w_i = 1 \right\}, \quad \mathcal{K}_h(x) := \frac{1}{h}\mathcal{K}\left(\frac{x}{h}\right), \quad x \in \mathbb{R} \tag{2.8}$$

and

$$\mathcal{Q}_{n,m}(q) := \frac{1}{m}\sum_{i=1}^m |C_i - (Aq)(K_i)|^2 - \frac{\alpha}{n}\sum_{i=1}^n \log(n w_i) \tag{2.9}$$

for some $\alpha > 0$ and $(Aq)(K) := \int_K^\infty (s - K)\, q(s)\, ds$. The form of the penalty in (2.9) can be motivated by the fact that if

$$w_i = \frac{\omega(X_i)}{\sum_{j=1}^n \omega(X_j)}, \quad i = 1, \ldots, n, \quad \omega(x) = q(x)/p(x), \tag{2.10}$$

then

$$\frac{1}{n}\sum_{i=1}^{n}\log(nw_i) \to \int_{\mathbb{R}} \log\omega(x)\, dP = -\mathrm{KL}(p||q) \tag{2.11}$$

as $n \to \infty$.

Remark 2.2. The parameter α determines the degree of regularisation: larger is α, more influence has the historical data on the estimation of q. It is reasonable to let $\alpha \to 0$ as $m \to \infty$, as the amount of information in options data becomes larger.

3. Main results

We assume the following statistical model for call prices

$$Y_i = (Aq^*)(K_i) + \varepsilon_i, \quad i = 1, \dots, m, \tag{3.1}$$

with some density q^*, where ε_i are $\mathcal{N}(0, \sigma^2)$ i.i.d. random variables representing some frictions on the market (e.g. bid ask spread). The additive errors scheme (3.1) applies to European call prices in an intraday context. In fact, for statistical analysis Renault (1997) interprets the error as mispricings which could be exploited by arbitrage strategies. In the next theorem we show that the estimate $q_{n,m}$ is close to $\bar{q}_{n,m} := \mathrm{argmin}_{q\in\mathcal{C}_{n,X}} \bar{\mathcal{Q}}_{n,m}(q)$, where

$$\bar{\mathcal{Q}}_{n,m}(q) := \frac{1}{m}\sum_{i=1}^{m} |(Aq^*)(K_i) - (Aq)(K_i)|^2 - \frac{\alpha}{n}\sum_{i=1}^{n}\log(nw_i). \tag{3.2}$$

In fact (3.2) can be viewed as a discretised version of (1.7).

Theorem 3.1. *Set* $\Delta := \bar{q}_{n,m} - q_{n,m}$. *If*

$$\|Aq^*\|_\infty + \|A\mathcal{K}_h\|_\infty < \infty \tag{A1}$$

and

$$\inf_{q\in\mathcal{C}_{n,X}} \bar{\mathcal{Q}}_{n,m}(q) \le \alpha/4, \tag{A2}$$

then it holds for $U < \sqrt{m}$ *and* $h \asymp n^{-1/5}$,

$$\mathbb{P}\left(\|A\Delta\|_m^2 \ge C(U/\sqrt{m} + \alpha)\big|\, \mathbf{X}\right) \le 2ne^{-U^2/B^2} \tag{3.3}$$

for some constants $C > 0$ *and* $B^2 = \max\{8\sigma^2 \max_j \|A\mathcal{K}_{jh}\|_m^2, 8\sigma^2\|Aq^*\|_m^2, 4\sigma^4\}$.

Remark 3.1. It follows from (A2) that

$$\|A(\bar{q}_{n,m} - q^*)\|_m^2 \le \alpha/4 \tag{3.4}$$

and hence

$$\mathbb{P}\left(\|A(q_{n,m} - q^*)\|_m^2 \ge C(U/\sqrt{m} + \alpha)\big|\, \mathbf{X}\right) \le 2ne^{-U^2/B^2} \tag{3.5}$$

for some constant $C > 0$.

Remark 3.2. The assumption (A1) imposes some restrictions on the "true" option prices and on the kernel. The assumption (A2) puts an apriori bound on the approximation error, which is deterministic for $X_1, \dots, X_n$ fixed.

Proof. Denote

$$\begin{aligned} T :=& \bar{\mathcal{Q}}_{n,m}(q) - \mathcal{Q}_{n,m}(q) \\ =& \frac{1}{m} \sum_{i=1}^{m} \left[|(Aq^*)(K_i) - (Aq)(K_i)|^2 - |C_i - (Aq)(K_i)|^2 \right]. \end{aligned} \tag{3.6}$$

We have

$$T = -\frac{1}{m} \sum_{i=1}^{m} \varepsilon_i^2 - \frac{2}{m} \sum_{i=1}^{m} (Aq^*)(K_i)\, \varepsilon_i + \frac{2}{m} \sum_{i=1}^{m} (Aq)(K_i)\, \varepsilon_i = T_1 + T_2 + T_3, \tag{3.7}$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Since we consider q only from the class of convex combinations $\mathcal{C}_{n,X}$ and a linear functional of convex combinations achieves its maximum value at the vertices, we get

$$\frac{m}{2} \sup_{q \in \mathcal{C}_{n,X}} |T_3(q)| \le \sup_{q \in \mathcal{C}_{n,X}} \left| \sum_{i=1}^{m} (Aq)(K_i)\, \varepsilon_i \right| = \max_{j=1,\dots,n} \left| \sum_{i=1}^{m} (A\mathcal{K}_{jh})(K_i)\, \varepsilon_i \right|, \tag{3.8}$$

where $\mathcal{K}_{jh} = \mathcal{K}_h(x - X_j)$. Hence and due to Lemma Appendix A.1,

$$\mathbb{P}\left(\sup_{q \in \mathcal{C}_{n,X}} |T_3(q)| > U/\sqrt{m} \,\middle|\, \mathbf{X} \right) \le 2n \exp\left(-\frac{U^2}{8\sigma^2 \max_j \|A\mathcal{K}_{jh}\|_m^2} \right). \tag{3.9}$$

Similarly, Lemma Appendix A.1 implies

$$\mathbb{P}\big(|T_2| > U/\sqrt{m} \big| \mathbf{X} \big) \le 2 \exp\left(-\frac{U^2}{8\sigma^2 \|Aq^*\|_m^2} \right) \tag{3.10}$$

and

$$\mathbb{P}\big(|T_1 + \sigma^2| > U/\sqrt{m} \big| \mathbf{X} \big) \le \exp\left(-\frac{U^2}{4\sigma^4} \right) + \exp\left(-\frac{U\sqrt{m}}{3\sigma^2} \right). \tag{3.11}$$

Hence, for $U \le \sqrt{m}$, $m \to \infty$, we have

$$\mathbb{P}\left(|T| > U/\sqrt{m} \big| \mathbf{X} \right) \le 2n(1 + o(1)) e^{-U^2/B^2} \tag{3.12}$$

where $B^2 = \max\{8\sigma^2 \max_j \|A\mathcal{K}_{jh}\|_m^2, 8\sigma^2 \|Aq^*\|_m^2, 4\sigma^4\}$. So, we have proved that with probability greater than $1 - 2ne^{-U^2/B^2}$

$$\sup_{q \in \mathcal{C}_{n,X}} |\mathcal{Q}_{n,m}(q) - \bar{\mathcal{Q}}_{n,m}(q)| \le \frac{U}{\sqrt{m}} \tag{3.13}$$

Hence,

$$\begin{aligned} 0 \le \bar{\mathcal{Q}}_{n,m}(q_{n,m}) - \bar{\mathcal{Q}}_{n,m}(\bar{q}_{n,m}) &\le \bar{\mathcal{Q}}_{n,m}(q_{n,m}) - \mathcal{Q}_{n,m}(\bar{q}_{n,m}) + \frac{U}{\sqrt{m}} \\ &\le \bar{\mathcal{Q}}_{n,m}(q_{n,m}) - \mathcal{Q}_{n,m}(q_{n,m}) + \frac{U}{\sqrt{m}} \\ &\le 2\frac{U}{\sqrt{m}} \end{aligned} \tag{3.14}$$

with probability greater than $1-2ne^{-U^2/B^2}$. On the other hand,

$$0 \le \bar{\mathcal{Q}}_{n,m}(q_{n,m}) - \bar{\mathcal{Q}}_{n,m}(\bar{q}_{n,m}) = -\frac{2}{m}\sum_{i=1}^{m}(A\Delta)(K_i)(A\bar{q}_{n,m} - Aq^*)(K_i) + \frac{1}{m}\sum_{i=1}^{m}(A\Delta)^2(K_i) + \frac{\alpha}{n}\sum_{i=1}^{n}\log\frac{\bar{w}_i}{w_i}, \tag{3.15}$$

where $\bar{q}_{n,m}(x) = \sum_{i=1}^{n}\bar{w}_i\mathcal{K}_h(x-X_i)$. Due to (A2)

$$\|Aq^* - A\bar{q}_{n,m}\|_m^2 < \alpha/4. \tag{3.16}$$

If $\|A\Delta\|_m^2 \ge 4\alpha$, then

$$\left|\frac{1}{m}\sum_{i=1}^{m}(A\Delta)(K_i)(A\bar{q}_{n,m} - Aq^*)(K_i)\right| \le \|A\Delta\|_m\|A\bar{q}_{n,m} - Aq^*\|_m \le \alpha^{1/2}\|A\Delta\|_m/2 \le \|A\Delta\|_m^2/4 \tag{3.17}$$

and

$$-\frac{2}{m}\sum_{i=1}^{m}(A\Delta)(K_i)(A\bar{q}_{n,m} - Aq^*)(K_i) + \|A\Delta\|_m^2 \ge \|A\Delta\|_m^2/2. \tag{3.18}$$

Thus, either $\|A\Delta\|_m^2 \le 4\alpha$ or

$$\frac{1}{2m}\sum_{i=1}^{m}(A\Delta)^2(K_i) + \frac{\alpha}{n}\sum_{i=1}^{n}\log\frac{\bar{w}_i}{w_i} \le 2\frac{U}{\sqrt{m}} \tag{3.19}$$

which is equivalent to

$$\frac{1}{2m}\sum_{i=1}^{m}(A\Delta)^2(K_i) \le 2\frac{U}{\sqrt{m}} + \alpha/4 \tag{3.20}$$

because of (A2). □

Let us now assess the error of approximating $\mathrm{KL}(p||q)$ by $-\frac{1}{n}\sum_{i=1}^{n}\log(nw_i)$. First we prove the following result.

Lemma 3.1. *Suppose that $\int[|p''(x)|^2/p(x)]dx < \infty$ and $\mathcal{K}$ is two-sided exponential kernel, then it holds*

$$\frac{1}{n}\sum_{i=1}^{n}\log\frac{q_n(X_i;W)}{p_n(X_i)} - \frac{1}{n}\sum_{i=1}^{n}\log(nw_i) = O_P(h). \tag{3.21}$$

Proof. Consider weights of the form $w_i = \frac{v_i}{\sum_j v_j}$, $i = 1,\dots,n$, then the optimization problem becomes

$$\min_{v_i}\frac{1}{m}\left\|Y - \frac{1}{\sum_i v_i}Qv\right\|^2 - \frac{\alpha}{n}\sum_j\log\left(\frac{v_j}{\sum_i v_i}\right) \tag{3.22}$$

with $q_{il} = k(K_l, X_i)$, $i = 1, \ldots, n$, $l = 1, \ldots, m$, where

$$k(x, X_i) = \frac{1}{h} \int_x^\infty (s - x) \mathcal{K}\left(\frac{s - X_i}{h}\right) ds. \tag{3.23}$$

We need to solve

$$\frac{2}{m} \sum_l \left[Y_l - \frac{1}{\sum_r v_r} \sum_j q_{jl} v_j \right] \left[-\frac{q_{kl}}{\sum_r v_r} + \frac{\sum_j q_{jl} v_j}{(\sum_r v_r)^2} \right] - \frac{\alpha}{n v_k} + \frac{\alpha}{\sum_r v_r} = 0. \tag{3.24}$$

Multiplying both sides by $\sum_j v_j$, we get

$$\frac{\sum_j v_j}{n v_k} = 1 + \frac{2}{m\alpha} \sum_l \left[Y_l - \frac{1}{\sum_r v_r} \sum_j q_{jl} v_j \right] \left[-q_{kl} + \frac{1}{\sum_r v_r} \sum_j q_{jl} v_j \right]. \tag{3.25}$$

Then for large $m\alpha$

$$\begin{aligned} \frac{w_k - w_i}{w_i} &= \frac{\frac{2}{m\alpha} \sum_l \left[Y_l - \frac{1}{\sum_r v_r} \sum_j q_{jl} v_j \right] [q_{il} - q_{kl}]}{1 + \frac{2}{m\alpha} \sum_l \left[Y_l - \frac{1}{\sum_r v_r} \sum_j q_{jl} v_j \right] \left[-q_{kl} + \frac{1}{\sum_r v_r} \sum_j q_{jl} v_j \right]} \\ &\approx \frac{2}{m\alpha} \sum_l \left[Y_l - \frac{1}{\sum_r v_r} \sum_j q_{jl} v_j \right] [q_{il} - q_{kl}] \end{aligned} \tag{3.26}$$

and we have to estimate

$$\sum_k \left(\frac{w_k}{w_i} - 1 \right) \mathcal{K}_{kh}(X_i) = \frac{2}{m\alpha} \sum_l \left[Y_l - \frac{1}{\sum_r v_r} \sum_j q_{jl} v_j \right] \sum_k [q_{il} - q_{kl}] \mathcal{K}_{kh}(X_i). \tag{3.27}$$

Consider

$$\begin{aligned} \sum_k (q_{il} - q_{kl}) \mathcal{K}\left(\frac{X_k - X_i}{h}\right) &= \frac{1}{h} \sum_k \mathcal{K}\left(\frac{X_k - X_i}{h}\right) \\ &\int_{K_l}^\infty (s - K_l) \left[\mathcal{K}\left(\frac{s - X_i}{h}\right) - \mathcal{K}\left(\frac{s - X_k}{h}\right) \right] ds. \end{aligned} \tag{3.28}$$

Now it is rather straightforward to show that under general assumptions on the kernel $\mathcal{K}$,

$$\frac{\sum_k (q_{il} - q_{kl}) \mathcal{K}\left(\frac{X_k - X_i}{h}\right)}{\sum_j \mathcal{K}\left(\frac{X_j - X_i}{h}\right)} = O(h). \tag{3.29}$$

Denote now

$$T = \frac{1}{n} \sum_{i=1}^n \log p(X_i) - \frac{1}{n} \sum_{i=1}^n \log p_n(X_i). \tag{3.30}$$

We have

$$T = I_1 + I_2, \tag{3.31}$$

where

$$
\begin{aligned}
I_1 &= -\int \log\left(\frac{\int \mathcal{K}_h(x-y)\,d[P_n(y)-P(y)]}{\int \mathcal{K}_h(x-z)dP(z)}+1\right) dP_n(x)\\
&= \int [P_n(x)-P(x)] d_x \log\left(-\frac{\int [P_n(y)-P(y)]\mathcal{K}_h'(x-y)\,dy}{\int \mathcal{K}_h(x-z)dP(z)}+1\right)\\
&\quad -\int \log\left(\frac{\int \mathcal{K}_h(x-y)\,d[P_n(y)-P(y)]}{\int \mathcal{K}_h(x-z)dP(z)}+1\right) dP(x) \qquad (3.32)\\
&= -\int \xi_n(x)\, d_x\left(\frac{\int \xi_n(y)\mathcal{K}_h'(x-y)\,dy}{\int \mathcal{K}_h(x-z)dP(z)}\right)\\
&\quad -\int \log\left(\frac{\int \mathcal{K}_h(x-y)\,d\xi_n(y)}{\int \mathcal{K}_h(x-z)dP(z)}+1\right) dP(x) + O_{\mathbb{P}}((hn)^{-1}),
\end{aligned}
$$

where $\xi_n(y) := \sqrt{n}[P_n(y)-P(y)]$. Using integration by parts, (AX) and the formula

$$\operatorname{cov}(\xi(t),\xi(u)) = \min\{P(u),P(t)\} - P(u)P(t), \qquad (3.33)$$

we get

$$I_1 = O_{\mathbb{P}}\left(\frac{1}{nh}\right). \qquad (3.34)$$

Consider the second term I_2. It holds

$$
\begin{aligned}
I_2 &= -\frac{1}{n}\sum_{i=1}^{n}\left[\log\left(\int \mathcal{K}_h(X_i-y)p(y)\,dy\right) - \log p(X_i)\right]\\
&= \int \log\left(\frac{p(x)}{\int \mathcal{K}_h(x-y)p(y)dy}\right) d(P_n(x)-P(x)) \qquad (3.35)\\
&\quad + \int \log\left(\frac{p(x)}{\int \mathcal{K}_h(x-y)p(y)dy}\right) dP(x) = I_{21}+I_{22}.
\end{aligned}
$$

Note that I_{21} is a sum of n independent zero-mean random variables. Using the inequality

$$|\log(t)| \le 2|\log\sqrt{t}| \le 2\frac{|t-1|}{\sqrt{t}}, \quad t>0, \qquad (3.36)$$

we get with $t = p(x)/\int \mathcal{K}_h(x-y)p(y)dy$,

$$
\begin{aligned}
\operatorname{Var}(I_{21}) \le \mathbf{E}\left[I_{21}^2\right] &\le \frac{1}{n}\int p(x)\left[\log\frac{p(x)}{\int \mathcal{K}_h(x-y)p(y)dy}\right]^2 dx\\
&\le \frac{4}{n}\int \frac{\left[\int \mathcal{K}_h(x-y)p(y)dy - p(x)\right]^2}{\int \mathcal{K}_h(x-z)p(z)\,dz}\,dx.
\end{aligned} \qquad (3.37)
$$

Using the inequality (see, e.g. Tsybakov (2008))

$$\int h(x)\log\left(\frac{h(x)}{g(x)}\right) dx \le \int \frac{(g(x)-h(x))^2}{g(x)}\,dx \qquad (3.38)$$

and the properties of the two-sided exponential kernel (Eggermont and LaRiccia (2001) p. 160), we derive

$$\begin{aligned} |I_{22}| \leq & \int \frac{\left[\int \mathcal{K}_h(x-y)p(y)dy - p(x)\right]^2}{\int \mathcal{K}_h(x-y)p(y)dy}\, dx \\ \leq & h^4 \int \frac{[\int \mathcal{K}_h''(x-y)p(y)dy]^2}{\int \mathcal{K}_h(x-z)p(z)dz}\, dx \leq h^4 \int \mathcal{K}_h(x-y)\frac{|p''(y)|^2}{p(y)}\, dx\, dy. \end{aligned} \tag{3.39}$$

Then under the condition $\int[|p''(x)|^2/p(x)]dx < \infty$ (see Eggermont and LaRiccia (2001) p. 162), we derive $I_{21} = O_{\mathbb{P}}\left(\frac{h^2}{\sqrt{n}}\right)$, whereas I_{22} is of order $O(h^4)$. Thus at least $T_{22} = O_{\mathbb{P}}(n^{-4/5})$. □

Next let us consider now for any $q \in \mathcal{C}_{n,X}$, the following functional

$$T(q) := \frac{1}{n}\sum_{i=1}^{n} \log \frac{q(X_i)}{p(X_i)} - \mathrm{KL}(p||q). \tag{3.40}$$

Clearly

$$\begin{aligned} \sup_{q\in\mathcal{C}_n} |T(q)| &= \sup_{q\in\mathcal{C}_n} \left| \frac{1}{n}\sum_{i=1}^{n} \log \frac{q(X_i)}{p(X_i)} - \mathrm{KL}(q||p) \right| \\ &\leq \sup_{q\in\tilde{\mathcal{C}}_n} \left| \frac{1}{n}\sum_{i=1}^{n} \log \frac{q(X_i)}{p(X_i)} - \mathrm{KL}(q||p) \right|, \end{aligned} \tag{3.41}$$

where

$$\tilde{\mathcal{C}}_n := \mathrm{conv}_n(\mathcal{H}) = \left\{ \sum_{i=1}^{n} w_i K_h(x-a_i), \quad \sum_{i=1}^{n} w_i = 1, \quad a \in \mathbb{R}^n \right\}. \tag{3.42}$$

and

$$\mathcal{H} := \{K_h(x-a), \quad a \in \mathbb{R}\}. \tag{3.43}$$

The following lemma holds (see Rakhlin et al (2005)).

Lemma 3.2. *If density p is such that $0 < a \leq p(x) \leq b$ for all x, then with probability at least $1 - e^{-t}$*

$$\sup_{q\in\tilde{\mathcal{C}}_n} \left| \frac{1}{n}\sum_{i=1}^{n} \log \frac{q(X_i)}{p(X_i)} - \mathrm{KL}(q||p) \right| \leq \mathbb{E}_X \left[\frac{c_1}{\sqrt{n}} \int_0^b \log^{1/2} \mathcal{D}(\mathcal{H}, \varepsilon, d_n)\, d\varepsilon \right] + c_2\sqrt{\frac{t}{n}}, \tag{3.44}$$

where c_1 and c_2 are constants that depend on a and b, $\mathcal{D}(\mathcal{H}, \varepsilon, d_n)$ is the covering number of $\mathcal{H}$ at scale ε with respect to empirical distance d_n for any $\phi_1, \phi_2 \in \mathcal{H}$

$$d_n^2(\phi_1, \phi_2) = \frac{1}{n}\sum_{i=1}^{n} (\phi_1(X_i) - \phi_2(X_i)). \tag{3.45}$$

4. Simulation study

In this section, we use a simulated example to illustrate the proposed nonparametric estimation procedure. The price of an European option at time 0 equals its discounted expected terminal value $\max(S_T - K, 0)$ with expectation taken with respect to the equivalent martingale measure Q :

$$C(K,T) = \mathbb{E}\left[\max(S_T - K, 0)e^{-rT}\right] = e^{-rT}\int_K^\infty (s-K)\, q(s,T)\, dS, \tag{4.1}$$

where $q(S,T)$ is the density of Q at time T. Black & Scholes model assumes that the asset price S_T is lognormally distributed

$$q(S,T;\alpha,\rho) = \text{lognorm}(x,\alpha,\rho) = \frac{1}{\sqrt{2\pi}\rho S} e^{-\frac{(\log S-\alpha)^2}{2\rho^2}} \tag{4.2}$$

with

$$\alpha = \log(S_0) + \left(r - \frac{\sigma_0^2}{2}\right)T, \tag{4.3}$$

$$\rho = \sqrt{\sigma_0^2 T}. \tag{4.4}$$

It is well-known, that the classical Black-Scholes model is not able to represent several important phenomena (see, e.g. Jackwerth (1996), Melick (1997)) observed in options data. In Melick (1997), Neumann (1998) a mixture of two lognormal distributions was shown to be more appropriate to model option prices. We adopt this model for our simulation study.

First, we generate an equidistant strike grid $(K_i, 1 \le i \le m)^{\mathrm{T}}$ in $[0,16]$ with $m = 200$. The call prices $C(K_i)$, $i = 1,\dots,m$, are obtained using the second order integration of q which is modelled as a mixture of two lognormal densities:

$$q\,(x,\mu,\nu\beta,\sigma) = (1-\beta)\,\text{lognorm}(x,\log\mu,\sigma) + \beta\,\text{lognormal}(x,\log\nu,\sigma) \tag{4.5}$$

with $\mu = 2$, $\nu = 7$ and $\sigma = 0.5$. Suppose $r = 0$, thus $S_{0,\mu} = \mu\exp\left(\frac{\sigma^2}{2}\right)$ and $S_{0,\nu} = \nu\exp\left(\frac{\sigma^2}{2}\right)$ (see (4.2)). Then with $T = 1$ we have

$$\begin{aligned} C\,(K) &= \int_K^\infty \int_u^\infty q\,(s)\; ds\, du \\ &= \int_K^\infty \left[1 - (1-\beta)\Phi\left\{(\log u - \log\mu)/\sigma\right\} - \beta\Phi\left\{(\log u - \log\nu)/\sigma\right\}\right] du. \end{aligned} \tag{4.6}$$

The market call prices $(Y_i, 1 \le i \le m)^{\mathrm{T}}$ are generated by adding a normal white noise to $(C(K_i),\, 1 \le i \le m)$, i.e.,

$$Y_i = C\,(K_i) + \vartheta\varepsilon_i, \quad i = 1,\dots,m, \tag{4.7}$$

where $\varepsilon_i \sim N\,(0,1)$ and $\vartheta = 0.05$. Furthermore, the historical density p is modelled as a shifted version of q, i.e. $p\,(x,\mu,\sigma,\tau) = q\,(x-\tau,\mu,\sigma)$, where τ is a shift. This transformation can be viewed as a measure change (from q to p). Next we generate

a sample $(X_t, 1 \le t \le n)^{\mathrm{T}}$ from p with $n = 1000$. Then the estimated weights $\widehat{W} = \{\widehat{\omega}_t\}_{t=1}^n$ are obtained by minimizing

$$\widetilde{Q}_{n,m}(q) = \frac{1}{m}\sum_{i=1}^{m}\{Y_i - (Aq_n)(K_i)\}^2 - \frac{\alpha}{n}\sum_{t=1}^{n}\log(n\omega_t), \tag{4.8}$$

subject to $\sum_{t=1}^n \omega_t = 1$, where $q_n(z;W) = \sum_{t=1}^n \omega_t \phi_h(z - X_t)$, $\phi_h(x) = (1/h)\phi(x/h)$ and ϕ is standard normal density. Then $Aq_n(z;W) = \frac{1}{h}\sum_{t=1}^n w_t \int_z^\infty \left(\int_s^\infty \mathcal{K}\left(\frac{u-X_t}{h}\right) du\right) ds$.

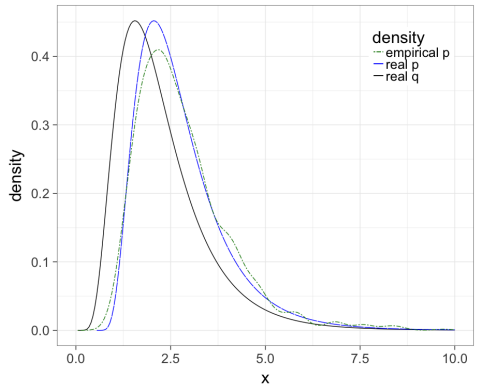


Fig. 1. Underlying densities q, p and the kernel density estimate p_n for the lognormal mixture model (4.5) with $\beta = 1$.

We compare the solution $q_n(x;\widehat{w})$ of the optimisation problem (4.8) with a local polynomial estimate of the second derivative of C (see (1.2)) (*locpoly* function from CRAN "KernSmooth" library with a Gaussian kernel). First let us consider the case $\beta = 1$, so that p and q are lognormal. For $\tau = 0.5$ the densities q, p and the kernel density estimate p_n (based on the sample $X_1, \ldots, X_n$) are shown in Figure 1. The solution of the optimisation problem (4.8) gives an estimate depicted in Figures 2 and 3.

Take now $\beta = 0.7$, i.e. consider the case of the genuine lognormal mixtures. For $\tau = 0.5$, the densities q, p and empirical density estimate p_n are shown in Figure 4, the solution of the corresponding optimisation problem gives an estimate depicted in Figure 5 for $\alpha = 0$ (no regularization) and for $\alpha = 0.1$ in Figure 6.

The integrated RMSE of the local polynomial estimator and the estimator obtained as a solution to (4.8) with $\beta = 0.7$ are shown in Figure 7 for $\alpha \in \{0, 0.1, 1\}$.

Empirical results show that our method (solution to problem (4.8)) is likely to be more suitable for the estimation of the risk-neutral density then local polynomial method for mixture-type models.

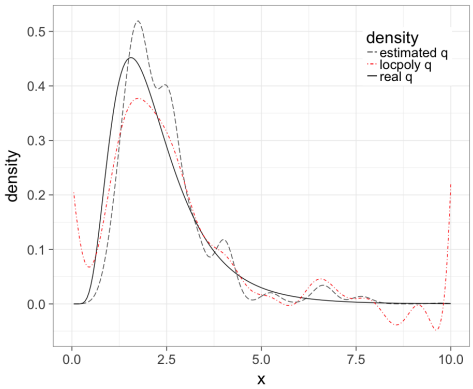


Fig. 2. Estimates of the risk-neutral density q in a log-normal model using the penalised sieve approach without penalisation (dashed line) and local polynomial approach (dashed-dotted line).

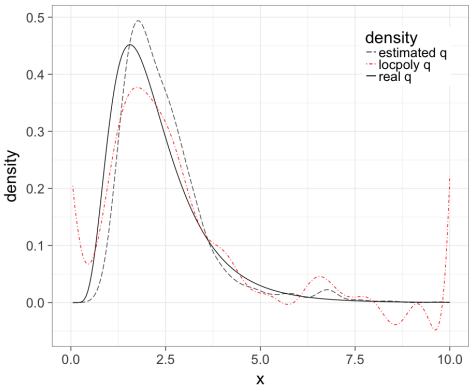


Fig. 3. Estimates of the risk-neutral density q in a log-normal model using the penalised sieve approach with $\alpha = 0.1$ (dashed line) and local polynomial approach (dashed-dotted line).

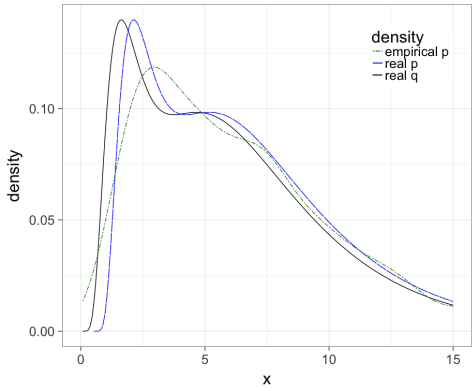


Fig. 4. Underlying densities q, p and the kernel density estimate p_n for the lognormal mixture model (4.5) with $\beta = 0.7$.

5. Real data analysis

We use the Strike-Call DAX dataset on November 16, 2011 to estimate model (4.7). There are $m = 1621$ observations on this day. Let $(Z_i, Y_i, i = 1, \ldots, 1621)$ denote the strike and call prices. Figure 8 shows the scatter plot of the call prices against the strike prices. Clearly, the option call price has a monotone decreasing pattern. We use the realisations of the historical stock price (DAX index) $(X_t, t = 1, \ldots, n)$ from March 12, 2009 to November 16, 2011, so that $n = 500$. The historical density function p is estimated by $p_n(x) = n^{-1} \sum_{t=1}^{n} \mathcal{K}_h(x - X_t)$. The risk-neutral density function estimate is then defined as

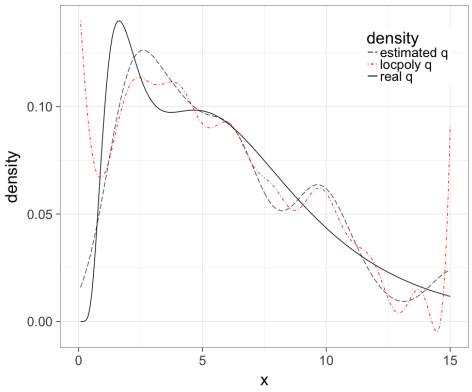


Fig. 5. Estimates of the risk-neutral density q in the lognormal mixture model (4.5) with $\beta = 0.7$ using the penalised sieve approach without penalisation (dashed line) and local polynomial approach (dashed-dotted line). The historical density is obtained from q by shifting it by $\tau = 0.5$

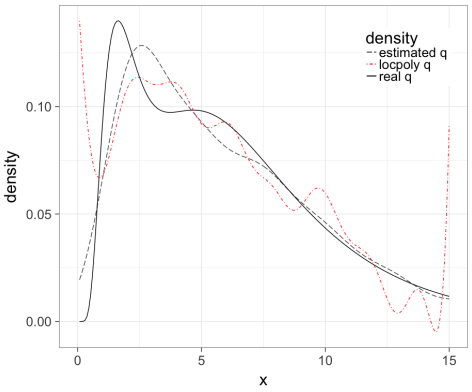


Fig. 6. Estimates of the risk-neutral density q in in the lognormal mixture model (4.5) with $\beta = 0.7$ using the penalised sieve approach with $\alpha = 0.1$ penalisation (dashed line) and local polynomial approach (dashed-dotted line). The historical density is obtained from q by shifting it by $\tau = 0.5$

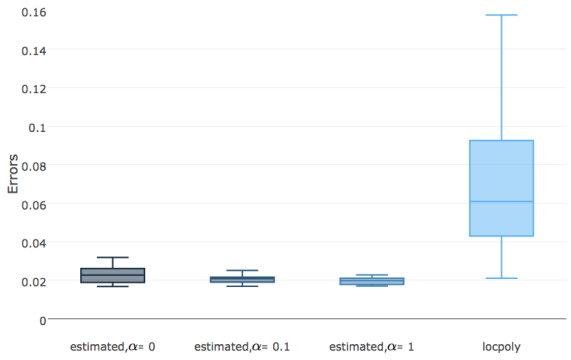


Fig. 7. The RMSE of the penalised sieve approach with $\alpha \in \{0, 0.1, 1\}$ and the local polynomial method for the case of the mixture model (4.5) with $\beta = 0.7$. The historical density is obtained from q by shifting it by $\tau = 0.5$

$q_n(z) = \sum_{t=1}^{n} \widehat{w}(X_t)\mathcal{K}_h(z - X_t)$, where the weights $\widehat{W} = \{\widehat{w}(X_t)\}_{t=1}^{n}$ are obtained by minimising (4.8). Define the moneyness at time t as $M_t = X_t/Z$. Figures 9 and 10 show the plots of $q_n(x; \widehat{W})$ (dashed line) and $p_n(x)$ (solid line) against moneyness for the cases of $\alpha = 0$ and $\alpha = 1000$. For the large value of α the one can see that the estimated risk-neutral density q is closer to the empirical estimate p.

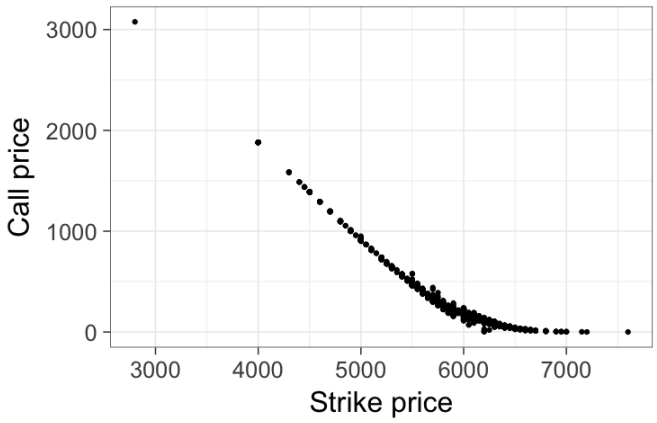


Fig. 8. Call option prices against strike prices from DAX dataset

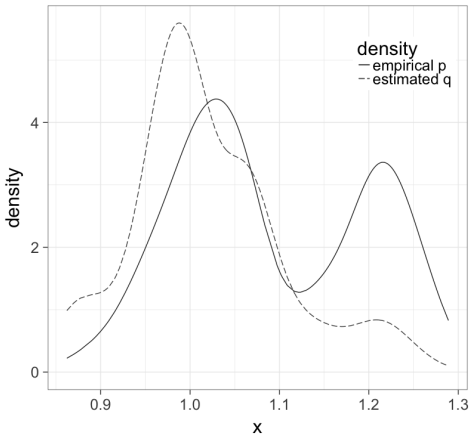


Fig. 9. The estimated historical density p_n and the risk-neutral density q obtained via penalised sieve approach with $\alpha = 0$ (no penalisation).

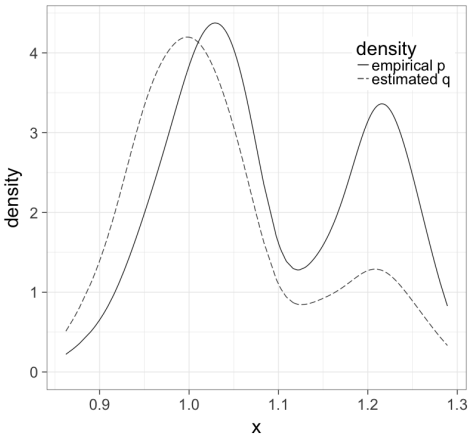


Fig. 10. The estimated historical density p_n and the risk-neutral density q obtained via penalised sieve approach with $\alpha = 1000$.

Appendix A. Appendix

Lemma Appendix A.1. *Let* $A = \{a_{ij}, \quad i,j = 1,\dots,N\}$ *be a* $N \times N$ *matrix. Denote the values* S_A *and* λ_A *by*

$$S_A^2 = 2\,\mathrm{tr}(A^\top A)^2, \quad \lambda_A = \|AA^\top\|_\infty. \tag{A.1}$$

If $\varepsilon_1, \ldots, \varepsilon_N$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables and $b = (b_1, \ldots, b_N)^\top$ is a deterministic vector then

$$\mathbb{P}(2|b^\top A\varepsilon| > z\sigma\|b\|(2\lambda_A)^{1/2}) \leq e^{-z^2/2} \tag{A.2}$$

and

$$\mathbb{P}(|\varepsilon^\top A^\top A\varepsilon - \mathrm{tr}(A^\top A)| > zS_A) \leq e^{-z^2/4} + e^{-zS_A/6\lambda_A} \tag{A.3}$$

References

Aït-Sahalia, Y. and Lo, A. W. (2000). Nonparametric risk management and implied risk aversion. *Journal of Econometrics* 94, 9-51.

Aït-Sahalia, Y., and Duarte, J. (2003). Nonparametric option pricing under shape restrictions. *Journal of Econometrics*, 116(1), 9-47.

Beare, B. (2011). Measure preserving derivatives and the pricing kernel puzzle. *Journal of Mathematical Economics* 47, 689-697.

Beare, B. and Schmidt, L. (2012). An empirical test of pricing kernel monotonicity. *Manuscript.*

Black, F. and Scholes M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 637-659.

Tsybakov, A.B. (2008) Introduction to Nonparametric Estimation. Springer Series in Statistics. Springer New York.

Eggermont, P. P. B. and LaRiccia, V. N. (2001). Maximum penalized likelihood estimation. Vol. I. Density estimation. Springer Series in Statistics. Springer-Verlag, New York.

Golubev, Y., Härdle, W., and Timonfeev, R. (2013) Testing monotonicity of pricing kernels. *Discussion paper.*

Härdle, W., Okhrin, Y., and Wang, W. (2013). Uniform confidence bands for pricing kernels. *Journal of Financial Economics* Revised and Submitted.

Jackwerth, J. C. (2000). Recovering risk aversion from option prices and realized returns. *Review of Financial Studies*, 13(2), 433-451.

Jackwerth, C., Rubinstein, M. (1996). Recovering Probability Distributions from Option. *Prices Journal of Finance*, 1611-1631.

Melick, W. R., Thomas, C. P. (1997). Recovering an Asset s Implied PDF from Option Prices An Application to Crude Oil during the Gulf Crisis. *Journal of Financial and Quantitative Analysis*, 9-115.

Neumann, M. (1998) Option Pricing under the Mixture of Distributions Hypothesis, Working paper 2008, Universitaet Karlsruhe.

Rakhlin, A., Panchenko, D., & Mukherjee, S. (2005). Risk bounds for mixture density estimation. *ESAIM: Probability and Statistics*, 9, 220-229.

Kallsen, J. (2002). Utility-based derivative pricing in incomplete markets. *In Mathematical Finance? Bachelier Congress 2000* 313-338.

Kellerer, H. G. (1972). Markov-komposition und eine anwendung auf martingale. *Mathematische Annalen* 198, 99-122.

Renault, E. (1997). Econometric models of option pricing errors. *Econometric Society Monographs* 28, 223-278.

Rothschild, M. and Stiglitz, J. E. (1970). Increasing risk: I. a definition. *Journal of Economic theory* 2, 225-243.

Rothschild, M. and Stiglitz, J. E. (1971). Increasing risk ii: Its economic consequences. *Journal of Economic theory* 3, 66-84.

Rothschild, M. and Stiglitz, J. E. (1973). Some further results on the measurement of inequality. *Journal of Economic theory* 6, 188-204.

Song, Z., and Xiu, D. (2016). A tale of two option markets: Pricing kernels and volatility risk. Journal of Econometrics, 190(1), 176-196.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory.* Springer, New York.

Vapnik, V. (1998). *Statistical Learning Theory.* Wiley, New York.

Computational Statistics and Data Analysis 121 (2018) 1–19

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Multivariate factorizable expectile regression with application to fMRI data☆

Shih-Kang Chao [a], Wolfgang K. Härdle [b,c], Chen Huang [d,*]

[a] Department of Statistics, Purdue University, 250 N University St., West Lafayette, IN 47907-2066, USA
[b] Ladislaus von Bortkiewicz Chair of Statistics, C.A.S.E. - Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany
[c] Sim Kee Boon Institute for Financial Economics, Singapore Management University, 50 Stamford Road, Singapore 178899, Singapore
[d] Faculty of Mathematics and Statistics, University of St. Gallen, Bodanstrasse 6, 9000 St. Gallen, Switzerland

ARTICLE INFO

Article history:
Received 10 March 2017
Received in revised form 28 October 2017
Accepted 3 December 2017
Available online 12 December 2017

Keywords:
Multivariate regression
Factor analysis
Expectile regression
Functional magnetic resonance imaging
Risk preference

ABSTRACT

A multivariate expectile regression model is proposed to analyze the tail events of large cross-sectional and spatial data, where the tail events are linked by a latent factor structure. The computational advantage of the method is demonstrated, and the estimation risk is analyzed for every *fixed number of iteration* and *fixed sample size*, when the latent factors are either exactly or approximately sparse. The proposed method is applied on the functional magnetic resonance imaging (fMRI) data taken during an experiment of investment decisions making. It is shown that the negative extreme blood oxygenation level dependent (BOLD) responses may be relevant to the risk preferences.

1. Introduction

Analyzing cross-sectional or spatial data is of critical interest in many scientific fields. Particularly, the interests in these fields are mostly in the *tail events*, which are the extreme events that occur with very small (or very large) probability. For example, in finance, Value-at-Risk (VaR) defined by the 1% quantile of the distribution of investment portfolio is widely used for measuring the market risk. In climatology, one of the major interests is the prediction of extreme precipitation defined by the tail quantile with level very close to 1. The estimation or prediction of tail events is often complicated by high dimensionality, which is common in many modern applications. However, the latent factors that influence all the cross-sections or spatial points may be sparse.

Multivariate regression (Izenman, 1975; Reinsel and Velu, 1998) is a classical tool for analyzing the cross-sectional or spatial data, and the penalization methods with matrix nuclear norm (Yuan et al., 2007; Negahban and Wainwright, 2011; Negahban et al., 2012) is applied to handle high dimensionality. However, the literature in multivariate regression is mostly silent about the estimation and prediction of tail events. On the other hand, quantile regression proposed by Koenker and Bassett (1978) is a well-known method for estimating the conditional quantiles, which is done through optimizing a non-differentiable loss function. Koenker and Portnoy (1990) generalize the quantile regression to a multivariate regression framework, but it cannot be applied to modern high dimensional data. To deal with high dimensional data with certain sparsity structure, it seems necessary to use some penalization methods, but the non-differentiable loss function of quantile regression is less convenient when being optimized with a penalty that is also non-differentiable, such as the nuclear norm.

☆ The codes to implement the algorithms are publicly accessible via the website Qwww.quantlet.de.

* Corresponding author.
E-mail address: chen.huang@unisg.ch (C. Huang).

https://doi.org/10.1016/j.csda.2017.12.001
0167-9473/

In this paper, we propose to estimate the tail events of a factorizable multivariate model using expectile regression (see (2.3) in Section 2 for the specific form of the model). Expectiles illustrate the tail events, and are closely related to quantiles (see, e.g. Section 2 of Rossi and Harvey, 2009). The expectile regression is proposed by Newey and Powell (1987) and is done through optimizing a smooth loss function. The smooth loss function of expectile regression yields computational advantages when being combined with a non-differentiable penalty, which will be shown in the algorithmic convergence analysis in Section 2.2. Furthermore, our method can be easily and efficiently implemented with the fast iterative shrinkage-thresholding algorithm of Beck and Teboulle (2009).

In addition to the convergence analysis, we *jointly* analyze the algorithmic and stochastic risk of our iterative estimator in Theorem 2.3, which characterizes the estimation error for each *fixed sample size* and *fixed number of iteration*. In particular, the theorem shows that our estimator is consistent as long as $\max\{p,\ m\} \ll n$ while $p,\ m \to \infty$, where p is the dimension of the covariates, m is the number of cross-sections or spatial points, and n is the sample size obtained in each cross-section or spatial point. The theorem is established under the weak assumption that the number of latent factors jointly influencing all the cross-sections or spatial points are approximately sparse.

Much interest has concentrated on using the functional magnetic resonance imaging (fMRI) data to understand the risk perception of humans (Heekeren et al., 2008). While the positive blood oxygenation level dependent (BOLD) signals are the focus in most studies, an increasing number of researchers are intrigued by the observed negative BOLD signals and their implications. Many hypotheses on the causes and implications of the negative BOLD are proposed, but they are still highly debatable (Mullinger et al., 2014).

We apply our method on the BOLD signals measured on the human subjects during an experiment on investment decisions making, and shed light on how the negative BOLD responses may be relevant in the decision making process. Using the same data, Majer et al. (2016) retrieve factor loadings from a dynamic factor model of the BOLD signals, and apply these loadings on explaining the subjects' risk attitude. However, their analysis only focus on the mean, and neglect the tail information of the BOLD signals. We apply our method on the BOLD responses obtained from 19 subjects, and estimate the factors and loadings at both high and low extreme expectile levels. We find that the factor loadings from the negative tail of the BOLD signals could not only well explain the revealed risk preference of the subjects in terms of R^2, but also *predict* the revealed risk preference. The prediction performance of the negative extreme BOLD is generally similar to that of the positive extreme BOLD, but sometimes they can be more accurate. Nonetheless, we note that our results do not yield any conclusions on the source of the negative BOLD responses.

The rest of the paper is arranged as follows. Section 2 introduces the model setting, estimation method and theoretical properties of the estimator. Simulation studies of our method are shown in Section 3. Section 4 illustrates the empirical application with the fMRI data. Section 5 concludes this paper. Proofs and auxiliary results are provided in Appendices.

2. Method

2.1. Model

We start with defining some notations. Denote a matrix $\mathbf{S} = (s_{lj}) = [\mathbf{S}_{\cdot 1} \ldots \mathbf{S}_{\cdot m}] \in \mathbb{R}^{p\times m}$, where $\mathbf{S}_{\cdot j} \in \mathbb{R}^p$ are the column vectors. Let $\|\mathbf{S}\|_{\mathrm{F}}$, $\|\mathbf{S}\|_*$ and $\|\mathbf{S}\|$ be the matrix Frobenius, nuclear and spectral norm. Denote $\sigma_{\min}(\mathbf{S})$ and $\sigma_{\max}(\mathbf{S})$ the smallest and largest singular values. For a vector $\mathbf{v} \in \mathbb{R}^p$, $\|\mathbf{v}\|_2$ is the Euclidean norm. Define $\langle\langle \mathbf{A}, \mathbf{B}\rangle\rangle \stackrel{\text{def}}{=} \mathrm{tr}(\mathbf{A}^\top \mathbf{B})$. $\mathbf{I}_m$ is the identity matrix with dimension m.

Let $\{(\boldsymbol{X}_i, Y_{i1}, \ldots, Y_{im})\}_{1\le i\le n}$ be the samples with $Y_{ij} \in \mathbb{R}$ and $\boldsymbol{X}_i \in \mathbb{R}^p$. Specifically, Y_{ij} represents the value observed from the response j at the time point i, and $\{\boldsymbol{X}_i\}_{i=1}^n$ are the covariates. For simplicity, we assume that the samples are i.i.d. over i. For $\tau \in (0, 1)$, the conditional expectile $e_j(\tau|\boldsymbol{X}_i)$ of Y_{ij} given $\boldsymbol{X}_i$ is defined by

$$e_j(\tau|\boldsymbol{X}_i) = \boldsymbol{X}_i^\top \boldsymbol{\gamma}_j(\tau), \tag{2.1}$$

where

$$\boldsymbol{\gamma}_j(\tau) \stackrel{\text{def}}{=} \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^p} \mathsf{E}\big[\rho_\tau(Y_{ij} - \boldsymbol{X}^\top \boldsymbol{\gamma})\big], \tag{2.2}$$

and $\rho_\tau(u) \stackrel{\text{def}}{=} |\tau - \mathbf{1}\{u < 0\}|\, |u|^2$. Define the coefficient matrix

$$\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_\tau \stackrel{\text{def}}{=} [\boldsymbol{\gamma}_1(\tau) \ldots \boldsymbol{\gamma}_m(\tau)].$$

We assume that the expectiles $e_1(\tau|\boldsymbol{X}_i), \ldots, e_m(\tau|\boldsymbol{X}_i)$ are related through a factor model:

$$e_j(\tau|\boldsymbol{X}_i) = \sum_{k=1}^{r} \psi_{j,k}(\tau) f_k^\tau(\boldsymbol{X}_i), \tag{2.3}$$

where $f_k^\tau(\boldsymbol{X}_i)$ is the kth factor, r is the number of factors, and $\psi_{j,k}(\tau)$ are the factor loadings. Furthermore, factors are constructed by linear combinations of covariates $\boldsymbol{X}_i$:

$$f_k^\tau(\boldsymbol{X}_i) = \boldsymbol{\varphi}_k(\tau)^\top \boldsymbol{X}_i \tag{2.4}$$

where $\boldsymbol{\varphi}_k(\tau) = (\varphi_{k,1}(\tau), \ldots, \varphi_{k,p}(\tau))^\top$. By substituting (2.4) into (2.3), it can be seen that the factor structure yields the reparametrization $\boldsymbol{\Gamma}^\top = \boldsymbol{\Psi}_\tau \boldsymbol{\Phi}_\tau$, where the matrix $\boldsymbol{\Psi}_\tau = (\psi_{j,k}(\tau))_{j\le m, k\le r}$ and $\boldsymbol{\Phi}_\tau = (\varphi_{k,l}(\tau))_{k\le r, l\le p}$. Unfortunately, the matrix factorization is in general not unique, so the factors and loadings may not be identifiable from $\boldsymbol{\Gamma}$. We alleviate the identifiability issue by imposing the normalization restrictions as Eq. (2.14) on page 28 of Reinsel and Velu (1998):

$$\boldsymbol{\Psi}_\tau^\top \boldsymbol{\Psi}_\tau = \mathbf{I}_m, \quad \boldsymbol{\Phi}_\tau \boldsymbol{\Phi}_\tau^\top = \operatorname{diag}(\sigma_1(\boldsymbol{\Gamma}), \ldots, \sigma_{p\wedge m}(\boldsymbol{\Gamma})). \tag{2.5}$$

The restrictions (2.5) make the factors and loadings associated with the nonzero singular values of $\boldsymbol{\Gamma}$ identifiable up to sign, if the nonzero singular values are distinct. When there exist repeated singular values, $\boldsymbol{\Psi}_\tau$ and $\boldsymbol{\Phi}_\tau$ cannot be uniquely identified; see Remark 2.1. Given the singular value decomposition $\boldsymbol{\Gamma} = \mathbf{UDV}^\top$, we have $\boldsymbol{\Psi}_\tau = \mathbf{V}$ and $\boldsymbol{\Phi}_\tau = \mathbf{D}^\top \mathbf{U}^\top$. Suppose an estimator $\widehat{\boldsymbol{\Gamma}}$ is available, we can estimate the kth factor by $\widehat{f}_k^\tau(\boldsymbol{X}_i) = \boldsymbol{X}_i^\top \widehat{\boldsymbol{\varphi}}_k(\tau) = \widehat{\sigma}_k \boldsymbol{X}_i^\top \widehat{\mathbf{U}}_{\cdot k}$ and the factor loadings for the jth response by $\widehat{\boldsymbol{\psi}}_j(\tau) = \widehat{\mathbf{V}}_{j\cdot}$, where $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ are unitary matrices obtained from the singular value decomposition $\widehat{\boldsymbol{\Gamma}} = \widehat{\mathbf{U}}\widehat{\mathbf{D}}\widehat{\mathbf{V}}^\top$.

Remark 2.1 (*Identifiability and Free Parameters*)**.** If there exist repeated singular values, then the singular vectors associated with these repeated singular values are not unique, and the factors and loadings are not uniquely identifiable. In particular, suppose the multiplicity of lth singular value $\mu_l > 1$, the number of free parameters for factor loadings (eigenvectors of the right singular spaces) is $\mu_l^2 - \binom{\mu_l}{2} - \mu_l = \mu_l(\mu_l - 1)/2$, where "$\mu_l^2$" is the total number of coefficients that determine the factor loadings associated with the lth singular value, "$-\binom{\mu_l}{2}$" is from the orthogonality constraints and "$-\mu_l$" is from the normalization constraints. Since the sign in the matrix factorization cannot be determined, the sign of the loadings and factors are not identifiable. In our empirical analysis in Section 4, we only use the absolute value of the loadings.

The factor model (2.3) implies that $\boldsymbol{\Gamma}$ is of rank r, and the model (2.1) corresponds to a multivariate linear regression model. For the standard regression with square loss, Reinsel and Velu (1998) propose to estimate $\boldsymbol{\Gamma}$ with the reduced-rank regression under the knowledge of r. However, r is usually unknown in practice. Yuan et al. (2007) propose to perform the multivariate regression with the nuclear norm penalty, which does not require the knowledge of r. The latter inspired the use of the nuclear norm penalty in the next section. However, Yuan et al. (2007) do not provide an algorithm that can scale up to large dimensions.

2.2. Algorithm

To estimate our model under the factor model (2.3), we combine an asymmetric loss with the nuclear norm penalty. To be more specific, we estimate $\boldsymbol{\Gamma}$ (defined in Section 2.1) by solving:

$$\widehat{\boldsymbol{\Gamma}}_{\tau,\lambda} \overset{\text{def}}{=} \arg\min_{\boldsymbol{\Gamma}\in\mathbb{R}^{p\times m}} F(\boldsymbol{\Gamma}), \tag{2.6}$$

$$F(\boldsymbol{\Gamma}) \overset{\text{def}}{=} (mn)^{-1} \sum_{i=1}^{n}\sum_{j=1}^{m} \rho_\tau(Y_{ij} - \boldsymbol{X}_i^\top \boldsymbol{\Gamma}_{\cdot j}) + \lambda \|\boldsymbol{\Gamma}\|_*, \tag{2.7}$$

where λ is a tuning parameter, $\boldsymbol{\Gamma}_{\cdot j}$ is the jth column of $\boldsymbol{\Gamma}$. The second term $\|\boldsymbol{\Gamma}\|_* = \sum_{l=1}^{\min(p,\,m)} \sigma_l(\boldsymbol{\Gamma})$, where the singular values $\sigma_1(\boldsymbol{\Gamma}) \ge \sigma_2(\boldsymbol{\Gamma}) \ge \cdots \ge \sigma_{\min(p,\,m)}(\boldsymbol{\Gamma})$. We note that (2.7) is a convex optimization problem. The number of factors r in (2.3) does not need to be specified. To simplify the notation, we denote $\widehat{\boldsymbol{\Gamma}}$ for $\widehat{\boldsymbol{\Gamma}}_{\tau,\lambda}$ hereinafter.

To solve the optimization problem (2.7), we apply the fast iterative shrinkage-thresholding algorithm (FISTA) of Beck and Teboulle (2009). FISTA solves the optimization problems of the form:

$$\min_{\boldsymbol{\Gamma}} \{g(\boldsymbol{\Gamma}) + h(\boldsymbol{\Gamma})\}, \tag{2.8}$$

where g is a smooth convex function with Lipschitz continuous gradient ∇g,

$$\|\nabla g(\boldsymbol{\Gamma}_1) - \nabla g(\boldsymbol{\Gamma}_2)\|_{\mathrm{F}} \le L_{\nabla g} \|\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2\|_{\mathrm{F}}, \forall \boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2 \in \mathbb{R}^{p\times m}, \tag{2.9}$$

where $L_{\nabla g}$ is the Lipschitz constant of ∇g and h is a continuous convex (possibly non-smooth) function (Ji and Ye, 2009).

In view of (2.7), this corresponds to

$$g(\boldsymbol{\Gamma}) \overset{\text{def}}{=} (mn)^{-1} \sum_{i=1}^{n}\sum_{j=1}^{m} \rho_\tau(Y_{ij} - \boldsymbol{X}_i^\top \boldsymbol{\Gamma}_{\cdot j}), \tag{2.10}$$

$$h(\boldsymbol{\Gamma}) \overset{\text{def}}{=} \lambda \|\boldsymbol{\Gamma}\|_*. \tag{2.11}$$

The Lipschitz constant of ∇g is $L_{\nabla g} = 2(mn)^{-1} \max(\tau, 1-\tau)\|\mathbf{X}\|_{\mathrm{F}}^2$; see Appendix A.1.

Algorithm 1 is an application of FISTA, with g and h chosen as (2.10) and (2.11).

The subroutine $\mathtt{SVT}_{\lambda,g}$ in Algorithm 1 is the singular value thresholding operator given by $\mathtt{SVT}_{\lambda,g}(\mathbf{S}) \overset{\text{def}}{=} \mathbf{U}_\mathbf{S}(\mathbf{D}_\mathbf{S} - (\lambda/L_{\nabla g})\mathbf{I}_{p\times m})_+ \mathbf{V}_\mathbf{S}^\top$, where SVD implies $\mathbf{S} = \mathbf{U}_\mathbf{S}\mathbf{D}_\mathbf{S}\mathbf{V}_\mathbf{S}^\top$, $\mathbf{I}_{p\times m}$ is a rectangular identity matrix with main diagonal elements equal to 1, and $(\mathbf{S})_+ = (\max\{0, s_{ij}\})$.

Algorithm 1: FISTA for expectile regression with nuclear norm penalty.

Input: $\{\boldsymbol{Y}_i\}_{i=1}^n$, $\{\boldsymbol{X}_i\}_{i=1}^n$, λ
Output: $\widehat{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}_T$
Initialization: $\boldsymbol{\Gamma}_0 = 0$, $\boldsymbol{\Omega}_1 = 0$, step size $\delta_1 = 1$;
for $t = 1, 2, \ldots, T$ **do**
  $\boldsymbol{\Gamma}_t = \mathrm{SVT}_{\lambda, g}\big(\boldsymbol{\Omega}_t - L_{\nabla g}^{-1} \nabla g(\boldsymbol{\Omega}_t)\big)$;
  $\delta_{t+1} = \frac{1+\sqrt{1+4\delta_t^2}}{2}$;
  $\boldsymbol{\Omega}_{t+1} = \boldsymbol{\Gamma}_t + \frac{\delta_t - 1}{\delta_{t+1}}(\boldsymbol{\Gamma}_t - \boldsymbol{\Gamma}_{t-1})$;
end

Remark 2.2 (*Initialization and the Stopping Rule*)**.** We suggest to initialize the algorithm with $\boldsymbol{\Gamma}_0 = 0$ in Algorithm 1 , but because the optimization problem is convex, this can be replaced by any matrix. Of course, the algorithm converges faster if we initialize it with a matrix that is close to the minimizer. We suggest to stop the algorithm at iteration T satisfying $|F(\boldsymbol{\Gamma}_{T+1}) - F(\boldsymbol{\Gamma}_T)| \le \epsilon$, for some small $\epsilon > 0$. In the simulation and empirical analysis of this paper, $\epsilon = 10^{-6}$.

The convergence of Algorithm 1 in terms of the loss function is guaranteed by the following theorem.

Theorem 2.1 (*Bounds for the Loss Difference and Convergence Rate in Algorithm 1*)**.** *Let* $\{\boldsymbol{\Gamma}_t\}_{t=0}^T$ *be the sequence obtained by the iteration of Algorithm 1 . Then*

$$|F(\boldsymbol{\Gamma}_t) - F(\widehat{\boldsymbol{\Gamma}})| \le \frac{4(mn)^{-1} \max(\tau, 1-\tau) \|\mathbf{X}\|_{\mathrm{F}}^2 \|\boldsymbol{\Gamma}_0 - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}^2}{(t+1)^2}. \tag{2.12}$$

In particular, if for $\epsilon > 0$,

$$t \ge \frac{2\sqrt{\max(\tau, 1-\tau)} \|\mathbf{X}\|_{\mathrm{F}} \|\boldsymbol{\Gamma}_0 - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}}{\sqrt{mn\epsilon}} - 1, \tag{2.13}$$

then $|F(\boldsymbol{\Gamma}_t) - F(\widehat{\boldsymbol{\Gamma}})| \le \epsilon$.

The bound (2.12) comes from a careful calculation of the Lipschitz constant of the gradient of g. The proof of Theorem 2.1 can be found in Appendix A.1.

Theorem 2.1 shows that to get an ϵ-accurate solution, it requires $1/\sqrt{\epsilon}$ steps when holding other parameters fixed. This is smaller than $1/\epsilon$ steps given by quantile regression and $1/\epsilon^2$ by the general subgradient methods, see Theorem 2.3 and Remark 2.4 in Chao et al. (2016). In view of (2.13), when τ is approaching 0 or 1, the number of iterations that is required to achieve an ϵ-accurate solution would increase.

Furthermore, utilizing the strong convexity of g, we can obtain a bound for $\|\boldsymbol{\Gamma}_t - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}^2$. For this purpose, additional assumptions on the design $\boldsymbol{X}$ are required.

(A1) Suppose $\mathrm{E}\boldsymbol{X}_i = 0$, $\mathrm{E}\boldsymbol{X}_i\boldsymbol{X}_i^\top = \boldsymbol{\Sigma}$ with $\sigma_{\min}(\boldsymbol{\Sigma}) > C_1$ and $\sigma_{\max}(\boldsymbol{\Sigma}) < C_2$ for some constants $C_1, C_2 > 0$ uniformly in p. For some sequence $0 < a_n < 1$, constants $c_1, c_2 > 0$,

$$\mathrm{P}\left[\sigma_{\min}\left(\frac{\mathbf{X}^\top\mathbf{X}}{n}\right) \ge c_1\sigma_{\min}(\boldsymbol{\Sigma}), \sigma_{\max}\left(\frac{\mathbf{X}^\top\mathbf{X}}{n}\right) \le c_2\sigma_{\max}(\boldsymbol{\Sigma})\right] \ge 1 - a_n. \tag{2.14}$$

Assumption (A1) holds for Gaussian design $\boldsymbol{X}$ with $c_1 = 1/9$, $c_2 = 9$ and $a_n = 4\exp(-n/2)$. See Wainwright (2009). It can be shown that (A1) holds for the sub-gaussian designs; see Vershynin (2012a) for details.

The following theorem characterizes the convergence in the Frobenius norm.

Theorem 2.2. *Given (A1), the sequence* $\boldsymbol{\Gamma}_t$ *obtained from Algorithm 1 satisfy*

$$\|\boldsymbol{\Gamma}_t - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}^2 \le \frac{36}{n(t+1)^2} \frac{\max(\tau, 1-\tau)}{\min(\tau, 1-\tau)} \frac{\|\mathbf{X}\|_{\mathrm{F}}^2}{\sigma_{\min}(\boldsymbol{\Sigma})} \|\boldsymbol{\Gamma}_0 - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}^2, \tag{2.15}$$

with probability greater than $1 - a_n$. *In particular, if for* $\epsilon > 0$,

$$t \ge 6\sqrt{\frac{\max(\tau, 1-\tau)}{\min(\tau, 1-\tau)}} \frac{\|\mathbf{X}\|_{\mathrm{F}} \|\boldsymbol{\Gamma}_0 - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}}{\sqrt{n\sigma_{\min}(\boldsymbol{\Sigma})\epsilon}} - 1, \tag{2.16}$$

then $\|\boldsymbol{\Gamma}_t - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}^2 \le \epsilon$ *holds with probability greater than* $1 - a_n$.

The proof of Theorem 2.2 is in Appendix A.2. We discuss the estimation of the number of factors in the following remark.

Remark 2.3 (*Estimation of the Number of Factors*). The number of factors r defined in Section 2.1 can be estimated by $\text{rank}(\mathbf{\Gamma}_T)$, which is the estimator generated by Algorithm 1 . If the number of factors is exactly sparse, $\text{rank}(\mathbf{\Gamma}_T)$ is usually a good estimator; see the simulation study in Section 3.

2.3. Oracle inequalities

In this section, we derive the bounds for the difference between the sequence $\mathbf{\Gamma}_t$ generated by Algorithm 1 and the true matrix $\mathbf{\Gamma}$. These results heavily rely on the strong convexity of ρ_τ.

We make the following assumptions.

(A2) There exists $C > 0$ such that for $u_{ij} \overset{\text{def}}{=} Y_{ij} - \boldsymbol{X}_i^\top \mathbf{\Gamma}_{\cdot j}$, $\mathrm{P}(|u_{ij}| > s) \le \exp(1 - s^2/C^2)$, $\forall s \ge 0$) with sub-gaussian norm $\|u_{ij}\|_{\psi_2} \overset{\text{def}}{=} \sup_{p\ge 1} p^{-1/2}(\mathsf{E}|u_{ij}|^p)^{1/p}$, and let $K_u \overset{\text{def}}{=} \max_{1\le j\le m} \|u_{ij}\|_{\psi_2}$.

(A3) Conditional on $\boldsymbol{X}_i$, Y_{ij} are independent over j.

(A2) regulates the tails of Y_{ij}. (A3) is required for obtaining the bounds on the tail probabilities of the estimation error. In Theorem 2.3, we state a non-asymptotic bound for $\|\mathbf{\Gamma}_t - \mathbf{\Gamma}\|_{\mathrm{F}}$ in the general situation that the number of factors can be increasing with n.

Theorem 2.3 (*Approximately Sparse Factors*). *Under (A1)– (A3), $\lambda = 2\ \mathrm{cm}^{-1} \max(\tau, 1-\tau) K_u \sqrt{\|\mathbf{\Sigma}\|}\sqrt{\frac{p+m}{n}}$ for some absolute constant $c > 0$. Then for any $q \in \{1, \dots, p \wedge m\}$, the sequence $\mathbf{\Gamma}_t$ obtained by Algorithm 1 satisfy*

$$\|\mathbf{\Gamma}_t - \mathbf{\Gamma}\|_{\mathrm{F}}^2 \le c''\Big(\frac{R_t}{n} + 1\Big)\sqrt{\frac{p+m}{n}}\zeta_\tau \left\{\sqrt{\frac{p+m}{n}}\zeta_\tau q + \sum_{j=q+1}^{p\wedge m} \sigma_j(\mathbf{\Gamma})\right\} + \frac{c'' R_t}{n}\|\mathbf{\Gamma}_0 - \mathbf{\Gamma}\|_{\mathrm{F}}^2, \tag{2.17}$$

with probability greater than $1 - 3\cdot 8^{-(p+m)} - a_n$, where $c'' > 0$ is an absolute constant, $R_t \overset{\text{def}}{=} \frac{1}{(t+1)^2}\frac{\max(\tau,1-\tau)}{\min(\tau,1-\tau)}\frac{\|\mathbf{X}\|_{\mathrm{F}}^2}{\sigma_{\min}(\mathbf{\Sigma})}$ and $\zeta_\tau \overset{\text{def}}{=} \frac{\max(\tau,1-\tau)}{\min(\tau,1-\tau)}\frac{\sqrt{\|\mathbf{\Sigma}\|}}{\sigma_{\min}(\mathbf{\Sigma})}K_u$.

Please see Appendix B for a proof of Theorem 2.3. Note that (2.17) holds for any $q \in \{1, \dots, p \wedge m\}$. The optimal bound is obtained by selecting q that balances $\sqrt{\frac{p+m}{n}}\zeta_\tau q$ and $\sum_{j=q+1}^{p\wedge m}\sigma_j(\mathbf{\Gamma})$. For a fixed number of iterations t in Algorithm 1 and τ, a sufficient condition for (2.17) tending to zero is that the number of factors r is *approximately sparse* ($\mathbf{\Gamma}$ is *approximately low rank*): there exists an increasing sequence $q = q_n \in \mathbb{N}$ such that

$$\lim_{n\to\infty}\frac{p+m}{n}\zeta_\tau^2 q = 0 \quad \text{and} \quad \lim_{n\to\infty}\left\{\sum_{j=q+1}^{p\wedge m}\sqrt{\frac{p+m}{n}}\zeta_\tau\sigma_j(\mathbf{\Gamma})\right\} = 0, \tag{2.18}$$

where p and m can be growing sequences in n. The quantity R_t characterizes how the computational cost influences the error bound. We can increase the number of iterations in Algorithm 1 to shrink R_t, but this also increases the computational cost. Similar to Theorems 2.1 and B.1, when τ is approaching to the boundaries of (0, 1), the bound in (2.17) will increase. Furthermore, heavier tails for Y_{ij} make higher K_u, and lead to higher error bounds.

If the number of factors is fixed and is not increasing with n ($\text{rank}(\mathbf{\Gamma})$ is fixed), then (2.17) is minimized by selecting $q = \text{rank}(\mathbf{\Gamma})$ and $\sum_{j=q+1}^{p\wedge m}\sqrt{\frac{p+m}{n}}\zeta_\tau\sigma_j(\mathbf{\Gamma}) = 0$ in (2.17). Hence, we have the following corollary.

Corollary 2.1 (*Exactly Sparse Factors*). *Under the conditions of Theorem 2.3,*

$$\|\mathbf{\Gamma}_t - \mathbf{\Gamma}\|_{\mathrm{F}}^2 \le c''\Big(\frac{R_t}{n} + 1\Big)\frac{p+m}{n}\zeta_\tau^2\,\text{rank}(\mathbf{\Gamma}) + \frac{c'' R_t}{n}\|\mathbf{\Gamma}_0 - \mathbf{\Gamma}\|_{\mathrm{F}}^2, \tag{2.19}$$

with probability greater than $1 - 3\cdot 8^{-(p+m)} - a_n$, where $c'' > 0$ is an absolute constant, $R_t = \frac{1}{(t+1)^2}\frac{\max(\tau,1-\tau)}{\min(\tau,1-\tau)}\frac{\|\mathbf{X}\|_{\mathrm{F}}^2}{\sigma_{\min}(\mathbf{\Sigma})}$ and $\zeta_\tau = \frac{\max(\tau,1-\tau)}{\min(\tau,1-\tau)}\frac{\sqrt{\|\mathbf{\Sigma}\|}}{\sigma_{\min}(\mathbf{\Sigma})}K_u$.

Remark 2.4. As explained in Section 2.1, we estimate $\mathbf{V}_{\cdot k,t}$ (the loadings corresponding to the kth factor for all responses) in the SVD $\mathbf{\Gamma}_t = \mathbf{U}_t\mathbf{D}_t\mathbf{V}_t^\top$. By Theorem 3.10 of Chao et al. (2016), we have:

$$1 - |\mathbf{V}_{\cdot k}^\top \mathbf{V}_{\cdot k,t}| \le \frac{2(2\|\mathbf{\Gamma}\| + \|\mathbf{\Gamma}_t - \mathbf{\Gamma}\|_{\mathrm{F}})\|\mathbf{\Gamma}_t - \mathbf{\Gamma}\|_{\mathrm{F}}}{\min\{\sigma_{k-1}^2(\mathbf{\Gamma}) - \sigma_k^2(\mathbf{\Gamma}), \sigma_k^2(\mathbf{\Gamma}) - \sigma_{k+1}^2(\mathbf{\Gamma})\}}, \tag{2.20}$$

where $\mathbf{V}_{\cdot k}$ are the true loadings. Theorem 2.3 (or Corollary 2.1) can be used with (2.20) to get an explicit bound.

Table 3.1
The averaged estimated number of factors $\widehat{r}$ over simulation repetitions with respect to τ and c. Values in the parentheses are the standard deviations over the simulation repetitions.

τ	0.05	0.3	0.5	0.7	0.95
	$r = 10$				
$c = 1.3$	10.95 (0.22)	11.00 (0.00)	10.00 (0.00)	11.00 (0.00)	10.94 (0.23)
$c = 1.5$	10.70 (0.47)	11.00 (0.00)	10.00 (0.00)	11.00 (0.00)	10.71 (0.46)
$c = 1.7$	10.19 (0.61)	11.00 (0.00)	10.00 (0.00)	11.00 (0.00)	10.20 (0.60)
	$r = 5$				
$c = 1.3$	6.00 (0.00)	6.00 (0.00)	5.00 (0.04)	6.00 (0.00)	6.00 (0.00)
$c = 1.5$	6.00 (0.00)	6.00 (0.00)	5.00 (0.00)	6.00 (0.00)	6.00 (0.00)
$c = 1.7$	6.00 (0.00)	6.00 (0.06)	5.00 (0.00)	6.00 (0.04)	6.00 (0.00)
	$r = 2$				
$c = 1.3$	3.00 (0.00)	3.00 (0.00)	2.03 (0.18)	3.00 (0.00)	3.00 (0.00)
$c = 1.5$	3.00 (0.00)	2.99 (0.12)	2.00 (0.00)	2.99 (0.09)	3.00 (0.00)
$c = 1.7$	3.00 (0.00)	2.72 (0.45)	2.00 (0.00)	2.78 (0.41)	3.00 (0.00)

3. Simulation study

In this section, we apply our method on the simulated data to evaluate the estimation performance on the factors and loadings, as the number of factors varies.

Set $n = m = p = 100$. For $i = 1, \ldots, n$, $j = 1, \ldots, m$, let $\boldsymbol{X}_i \sim \mathcal{N}(0, \Sigma_{p\times p})$ with $\Sigma_{jk} = 0.5^{|j-k|}$ and $\boldsymbol{\varepsilon}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_{m\times m})$, the response variables are generated by

$$Y_{ij} = \boldsymbol{X}_i^\top \boldsymbol{\Gamma}_{\cdot j} + \varepsilon_{ij} = \sum_{k=1}^{r} \psi_{jk} f_k(\boldsymbol{X}_i) + \varepsilon_{ij} = \sum_{k=1}^{r} \mathbf{V}_{jk} \mathbf{D}_{kk} \boldsymbol{X}_i^\top \mathbf{U}_{\cdot k} + \varepsilon_{ij}, \tag{3.1}$$

where $r = \operatorname{rank}(\boldsymbol{\Gamma})$. We will set $r = 2, 5, 10$, and the nonzero diagonal components of $\mathbf{D}$ are (19.01, 18.74, 18.65, 18.22, 17.80, 17.50, 17.21, 17.02, 16.57, 16.49). The columns of $\mathbf{V}$ and $\mathbf{U}$ are the orthonormal singular vectors of a matrix with components chosen from $\mathcal{N}(0, 1)$. We repeat the data generation 500 times.

We apply Algorithm 1 with $\boldsymbol{Y}$ and $\widetilde{\boldsymbol{X}} = (\mathbf{l}_n, \widetilde{\boldsymbol{X}})$, where $\mathbf{l}_n = (1, \ldots, 1)$ is the intercept. The tuning parameter λ is selected according to Lemma B.1, i.e., $\lambda = 2\ \mathrm{cm}^{-1} \max(\tau, 1-\tau) K_u \sqrt{\|\boldsymbol{\Sigma}\|} \sqrt{\frac{p+m}{n}}$. We stop the algorithm as described in Remark 2.2. Denote the resulting estimator $\widetilde{\boldsymbol{\Gamma}}_1$, and obtain $\widetilde{\boldsymbol{\Gamma}}$ by removing the first row (the intercept) of $\widetilde{\boldsymbol{\Gamma}}_1$.

Table 3.1 reports the results for the estimated number of factors $\widehat{r}$, which is the number of nonzero singular values of $\widetilde{\boldsymbol{\Gamma}}$ that are greater than 10^{-10}. That is, the singular values smaller than 10^{-10} are treated as zero. We try several values of c in the formula for λ because we do not know its exact value. The true number of factors are generally well recovered by our algorithm, except for the expectiles that deviate more from $\tau = 0.5$. Furthermore, the estimated number of factors is robust to the model randomness as the standard deviations are very small. The results are similar for different values of c, so we fix $c = 1.3$ for all the later analysis.

The Frobenius error $\|\widetilde{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}\|_{\mathrm{F}}$ is shown in Fig. 3.1. The results are symmetric in τ around $\tau = 0.5$, and the estimation errors tend to be larger for the tail τ. In the models where r is larger, the Frobenius error is also larger. Our findings in the simulation studies are consistent with the roles of τ and r in the error bound in Corollary 2.1.

We measure the estimation performance of the factors and loadings by

$$\begin{aligned} &\|\boldsymbol{\Delta}^{\mathrm{fac}}_{k\cdot}\|_2 / \mathbf{D}_{kk}, \text{ where } \boldsymbol{\Delta}^{\mathrm{fac}} \overset{\text{def}}{=} |\widetilde{\mathbf{D}}\widetilde{\mathbf{U}}^\top| - |\mathbf{D}\mathbf{U}^\top|, \\ &\|\boldsymbol{\Delta}^{\mathrm{load}}_{\cdot k}\|_2, \text{ where } \boldsymbol{\Delta}^{\mathrm{load}} \overset{\text{def}}{=} |\widetilde{\mathbf{V}}| - |\mathbf{V}|, \\ \text{and} \quad &1 - |\mathbf{V}_{\cdot k}^\top \widetilde{\mathbf{V}}_{\cdot k}|, \end{aligned} \tag{3.2}$$

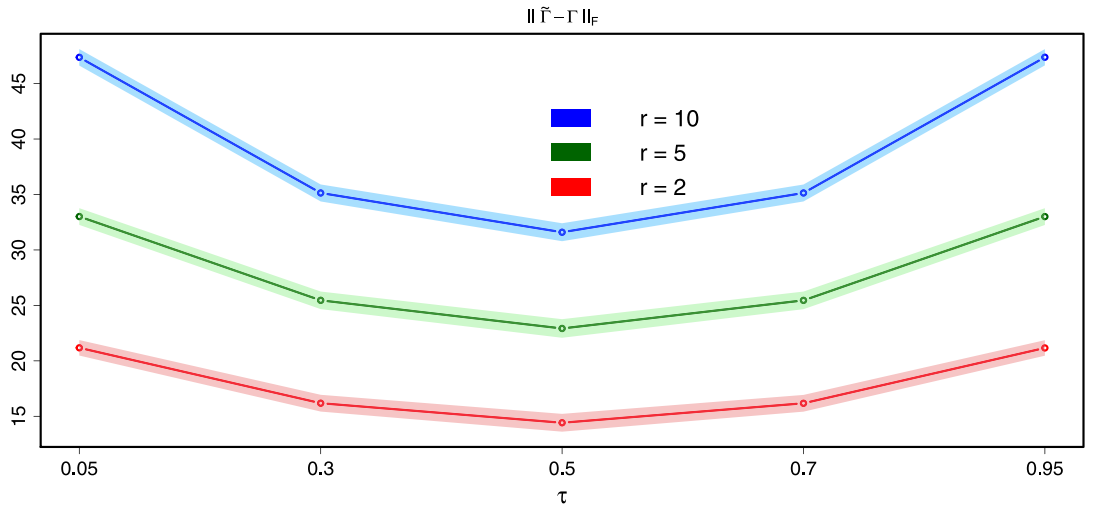


Fig. 3.1. The averaged estimation error $\|\widetilde{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_F$ ($c = 1.3$ in λ). The solid lines represent the averaged Frobenius errors over simulation repetitions, and the bands describe the standard deviations over the simulation repetitions.

for $k = 1, \ldots, r$, where $\widetilde{\mathbf{V}}$, $\widetilde{\mathbf{D}}$ and $\widetilde{\mathbf{U}}$ are based on the SVD $\widetilde{\mathbf{\Gamma}} = \widetilde{\mathbf{U}}\widetilde{\mathbf{D}}\widetilde{\mathbf{V}}^\top$, and the absolute value is taken componentwisely to the matrix. We do not include the covariate $\boldsymbol{X}_i$ in the measure for the estimation error of the factors because all factors share the same $\boldsymbol{X}_i$. We choose two measures for the estimation performance of the loadings. The $\|\mathbf{\Delta}^{\text{load}}_{\cdot k}\|_2$ measures the performance on the recovery of the absolute values of the loadings, which will be relevant in the empirical analysis in Section 4. On the other hand, $1 - |\mathbf{V}_{\cdot k}^\top \widetilde{\mathbf{V}}_{\cdot k}|$ corresponds to the theory that we stated in (2.20), which can be regarded as another measure for the recovery performance. We have also performed the analysis for $\tau = 0.05$ and 0.3, but we do not include their results in the paper because they are similar to $\tau = 0.95$ and 0.7. The results are presented in Fig. 3.2. Some general patterns are observed for the three panels. Smaller r gives smaller estimation error, but the associated standard deviation is larger. When τ deviates from 0.5, the error is larger, and this effect is particularly for $\|\mathbf{\Delta}^{\text{fac}}_{k\cdot}\|_2/\mathbf{D}_{kk}$. $\|\mathbf{\Delta}^{\text{load}}_{\cdot k}\|_2$ shows similar pattern to $1 - |\mathbf{V}_{\cdot k}^\top \widetilde{\mathbf{V}}_{\cdot k}|$, but the variance for $1 - |\mathbf{V}_{\cdot k}^\top \widetilde{\mathbf{V}}_{\cdot k}|$ is overall larger.

4. Empirical analysis: predicting risk attitude with fMRI Data

In this section, we apply our method on the fMRI data to predict the risk attitude on the investment decisions making. To understand how human brain responds to reward and risk is an important research topic in neuropsychology, financial economics and neuroeconomics (Heekeren et al., 2008; Camerer, 2007; Schultz, 2015). Previous research mainly focuses on the identification of the region of interest (ROI) with Blood Oxygenation Level Dependent (BOLD) signals (see Schultz, 2015 and the references therein). However, only a few research uses fMRI on predicting the risk attitude of subjects. Helfinstein et al. (2014) train support vector machines with the BOLD signals recorded in a Ballon Analog Risk Task (BART) on several combinations of brain regions, and this classifier can predict subjects' next choice with over 70% accuracy; van Bömmel et al. (2014) and Majer et al. (2016) retrieve factor loadings from a dynamic factor model on BOLD and apply these loadings on predicting subjects' risk attitude.

We focus on predicting the risk attitude of the subjects using the BOLD signals, but we differ from the previous studies in that we separately analyze the *positive* and *negative* BOLD signals observed in the cortical regions. The positive BOLD signals are known to be closely associated with increased neuronal activities, but the interpretation of large *negative* BOLD responses (NBR) is still controversial. Mullinger et al. (2014) argue that the best explanation for NBR at the cortical layer might be a decrease in cerebral blood flow (CBF) with a lesser reduction in the neuronal activities, which is measured by the cerebral metabolic rate of oxygen consumption ($CMRO_2$). This explanation is proven to be an important complement of the more classical "blood/vascular stealing" hypothesis (see p. 263 of Mullinger et al., 2014). However, Mullinger et al. (2014) also argue that there may exist deeper neuronal reasons for NBR than simply the inversion of the neurovascular coupling mechanism of the positive BOLD responses. Following the interpretation of NBR of Mullinger et al. (2014), we suspect that NBR also contain valuable information for predicting the risk attitude. Using our expectile based approach, we study whether the positive and negative extreme BOLD responses are relevant to the risk attitude.

4.1. Data

Our data come from a rapid event-related design experiment on investment decisions making, and this data set is firstly analyzed in Majer et al. (2016). The experiment was done as follows: 19 subjects were requested to make choices in 256

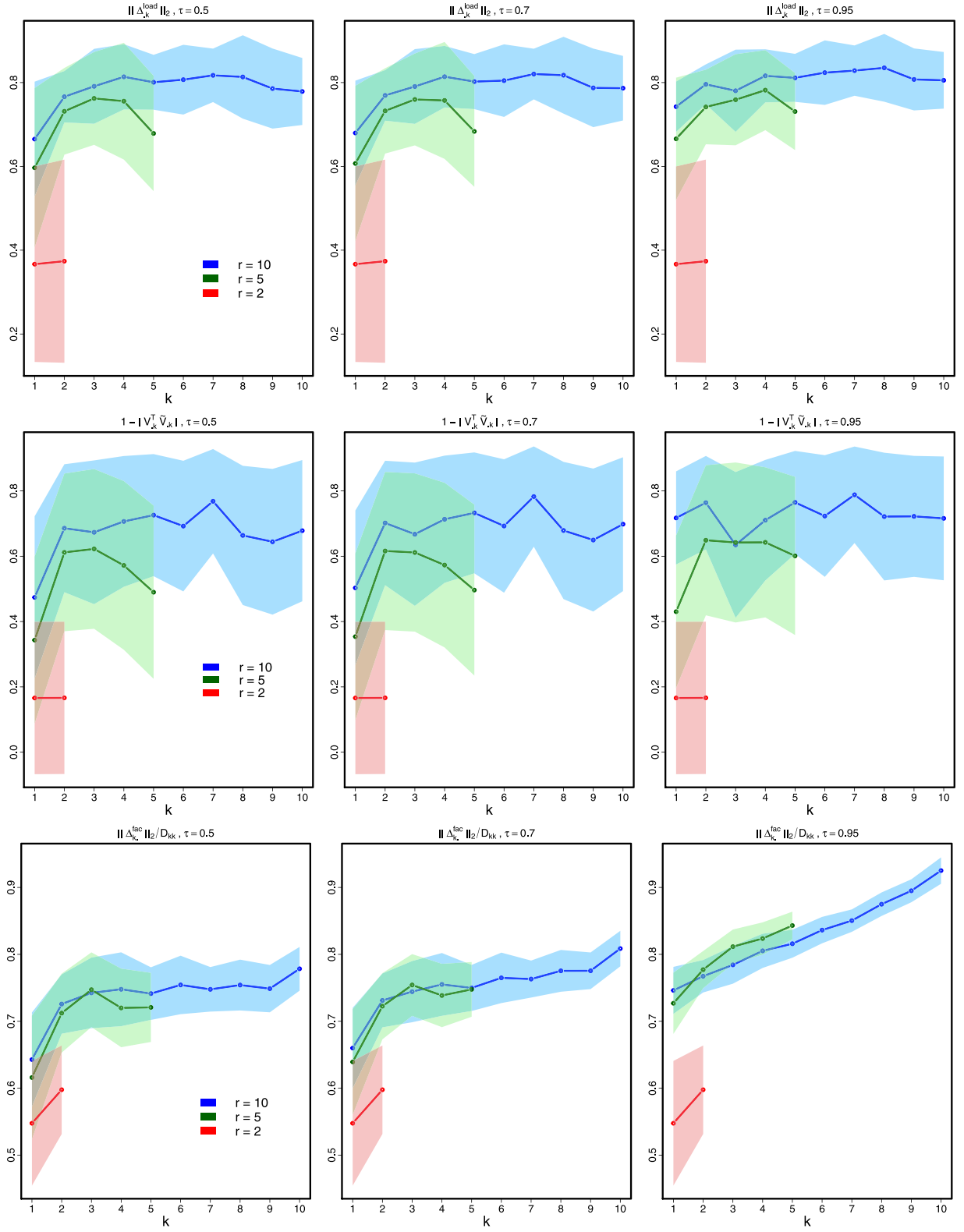


Fig. 3.2. The estimation errors $\|\boldsymbol{\Delta}^{\text{load}}_{.k}\|_2$, $1-|\mathbf{V}^{\top}_{.k}\widetilde{\mathbf{V}}_{.k}|$ for the loadings and $\|\boldsymbol{\Delta}^{\text{fac}}_{k.}\|_2/\mathbf{D}_{kk}$ for the factors, defined in (3.2). The solid lines represent the averaged errors, and the bands describe the standard deviations over simulation repetitions; $c = 1.3$ in λ.

Table 4.1
The goodness of fit R^2, Spearman's and Kendall's rank correlations for the in-sample fitting and out-of-sample prediction by (M1) or (M2) with/without constrains, under different τ, ω levels.

		Constrained model (only 1st factor)						Unconstrained model (2 factors)					
		Whole series (M1)			Task-wise (M2)			Whole series (M1)			Task-wise (M2)		
τ		0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
		In-sample fitting											
$\omega = 0.1$	R^2	0.084	0.158	0.101	0.412	0.412	**0.413**	0.312	0.263	0.226	**0.455**	0.454	0.454
	Spearman's rank corr	0.149	0.377	0.328	0.595	0.595	**0.604**	0.532	0.526	0.396	**0.618**	**0.618**	**0.618**
	Kendall's rank corr	0.076	0.263	0.228	0.462	0.462	**0.474**	0.333	0.357	0.275	**0.474**	**0.474**	**0.474**
$\omega = 0.5$	R^2	0.070	0.043	0.030	0.134	0.136	0.135	0.307	0.260	0.352	**0.445**	0.440	0.441
	Spearman's rank corr	0.177	0.140	0.226	0.335	0.316	0.326	0.547	0.528	0.596	0.533	0.544	0.544
	Kendall's rank corr	0.135	0.088	0.135	0.205	0.193	0.205	0.427	0.333	**0.415**	0.368	0.380	0.380
$\omega = 0.9$	R^2	0.199	0.238	0.148	0.206	0.205	0.205	0.393	0.367	0.229	**0.487**	**0.496**	**0.500**
	Spearman's rank corr	0.435	0.540	0.181	0.412	0.412	0.412	0.588	**0.628**	0.582	0.596	**0.637**	**0.637**
	Kendall's rank corr	0.333	0.391	0.135	0.298	0.298	0.298	**0.439**	**0.439**	**0.439**	0.462	**0.497**	**0.497**
		Out-of-sample predicting											
$\omega = 0.1$	Spearman's rank corr	−0.453	−0.181	−0.321	**0.454**	0.451	0.440	−0.079	−0.133	0.072	0.298	0.298	0.298
	Kendall's rank corr	−0.322	−0.111	−0.240	**0.357**	0.345	0.345	−0.076	−0.088	0.041	0.216	0.216	0.216
$\omega = 0.5$	Spearman's rank corr	−0.444	−0.700	−0.658	−0.119	−0.119	−0.119	−0.035	−0.196	0.247	0.205	0.204	0.212
	Kendall's rank corr	−0.275	−0.509	−0.450	−0.064	−0.064	−0.064	−0.006	−0.146	0.135	0.123	0.111	0.123
$\omega = 0.9$	Spearman's rank corr	−0.207	0.204	−0.493	0.023	0.023	0.023	0.161	0.072	−0.447	0.293	0.307	0.307
	Kendall's rank corr	−0.170	0.135	−0.345	0.006	0.006	0.006	0.076	0.041	−0.298	0.205	0.216	0.216

investment decision tasks and each task lasts 7 s. The fMRI was taken every two seconds (temporal resolution = 2 s), and this resulted in 1400 images for each subject. We have also acquired the answer for each task from each subject. Before applying our method, it is necessary to identify the region of interest (ROI), because the BOLD responses in non-ROIs are generated by noise (under the generalized linear model; see Section 6.2.1 of Lindquist (2008)) and do not have a sparse factor structure. For our data, Majer et al. (2016) identify three brain regions Anterior insula (left and right aINS) and dorsomedial prefrontal cortex (DMPFC) as the active regions related to investment decisions via spectral clustering method. We will only focus on the BOLD responses of the voxels in these three regions.

We integrate the information of each region (left and right aINS and DMPFC) spatially by taking the *quantiles* of the BOLD responses over all voxels in these regions. At each fMRI scan i of the sth subject, we take the quantiles with levels $\omega \in \{0.1, 0.5, 0.9\}$ of the BOLD responses over all voxels in the regions $b = 1$ (aINS_L), $b = 2$ (aINS_R) and $b = 3$ (DMPFC) to construct a single time series $v_i(s, b, \omega)$, where $i = 1, \ldots, N = 1400$. Fig. 4.1 gives an illustration of the BOLD time series of each cluster. For each cluster, the series of 19 subjects at ω are averaged (the solid lines) and the bands show the dispersion of the 19 time series. We observe that the series for $\omega = 0.9$ is positive, which summarizes the information of the positive BOLD responses, while the series for $\omega = 0.1$ is negative, which corresponds to the negative BOLD responses. The series for $\omega = 0.5$ is stationary and varying around the origin. From Fig. 4.1, we observe that the series with each different ω shows different volatility, and this may imply that the series with different ω contains different information. We will show in Table 4.1 that the series with $\omega = 0.1$ and 0.9 tend to contain more information than $\omega = 0.5$.

4.2. Method

4.2.1. Factor loadings at each region b and quantile level ω

For each ω and a single region b, we consider two approaches to construct the variable Y_{ij}:

- **(C1)** Whole time series: set $Y_{ij}^{b,\omega} = v_i(j, b, \omega)$, where $i = 1, \ldots, n$ with $n = N = 1400$, $j = 1, \ldots, 19$ (subject). Thus, we have $m = 19$ curves in each region b and at each quantile level ω.
- **(C2)** Analyzing each task separately (task-wise): we divide the whole time series in each region b and at each quantile level ω into subseries based on the beginning and the end of each task. Let $\mathcal{I}_q \subset \{1, \ldots, N\}$ be the set containing the indices of the images taken during the qth task. In our data, each $|\mathcal{I}_q| = 3$ or 4. We linearly interpolate the points $\{v_i(s, b, \omega)\}_{i \in \mathcal{I}_q}$ for each fixed s, b, and ω. Denote $\widetilde{v}_i(s, b, q, \omega)$ by the value on the interpolated curve at the ith point in n equally distant grid on the interval $(\min(\mathcal{I}_q), \max(\mathcal{I}_q))$, where $i = 1, \ldots, n = 50$. Let $Y_{ij}^{b,\omega} = \widetilde{v}_i(s, b, q, \omega)$ with $j = 256(s-1) + q$, where $s = 1, \ldots, 19$ (index for subject) and $q = 1, \ldots, 256$ (index for task) for each ω, b. Thus, there are $m = 19 \times 256 = 4864$ curves in each b and ω.

The variable $\boldsymbol{X}_i$ is a vector of basis functions that need to be flexible enough to capture the various shapes of the fMRI BOLD sequences. For this purpose, we use the cubic B-spline basis $\{B_k\}_{k=1}^{p}$ with equally spaced knots on [0, 1], and set $\boldsymbol{X}_i = \big(B_1(i/n), B_2(i/n), \ldots, B_p(i/n)\big)^\top$, where $i = 1, \ldots, n$. Note that $n = 1400$ in (**C1**) and $n = 50$ in (**C2**). B-splines are

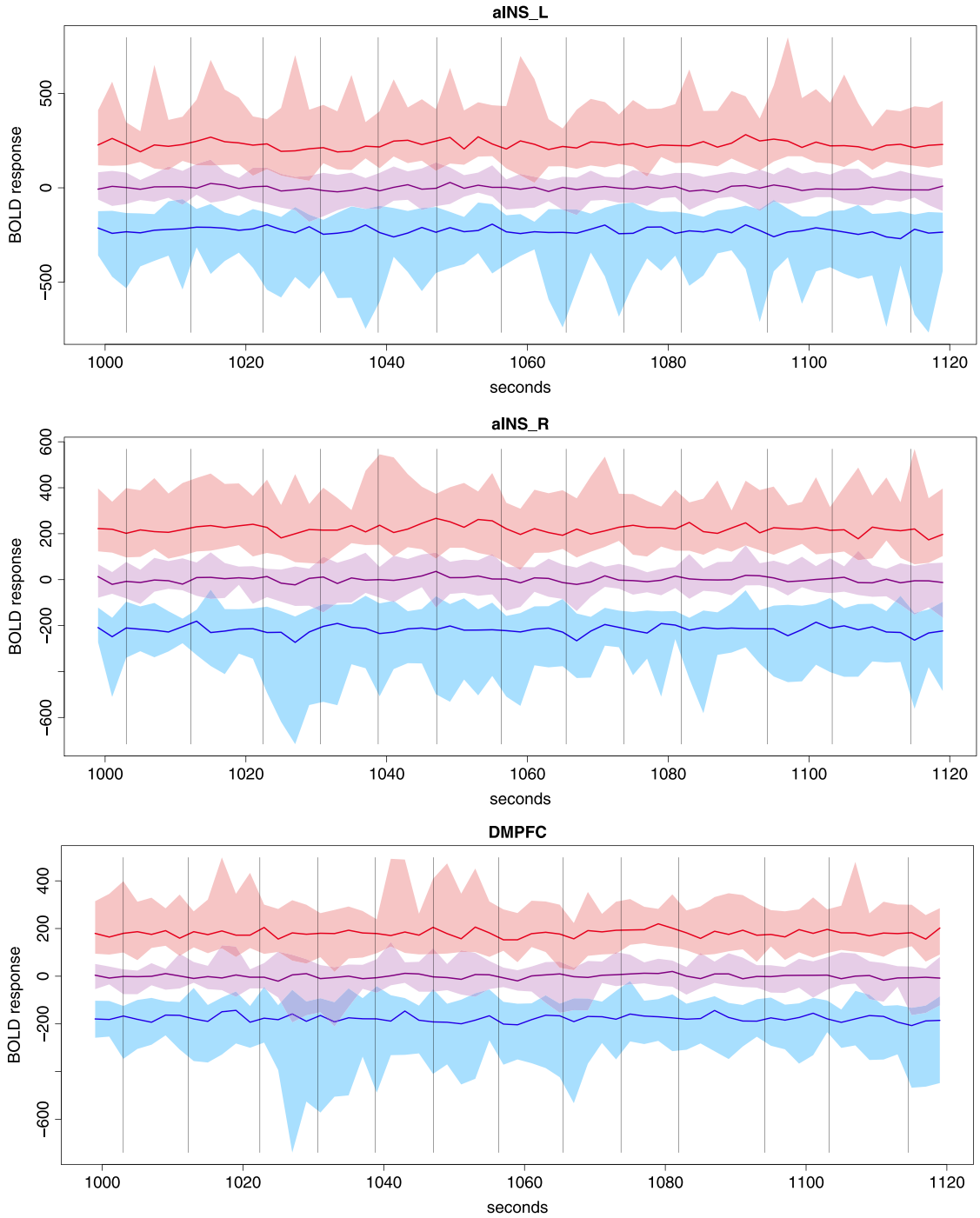


Fig. 4.1. In each region, the ω quantiles of the BOLD responses over all the voxels between 1000 and 1120 s of the experiment are shown. In each subfigure (region), lowest (resp., middle, highest) solid lines represent the median of $\omega = 0.1$ (resp., $\omega = 0.5, 0.9$) quantiles of all 19 subjects, and the upper and lower boundaries of the bands present the maximum and the minimum of the ω quantiles of the 19 subjects. Vertical lines indicate the occurrences of the stimuli (the beginning of each task). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

suitable for estimating the hemodynamic response function, see Degras and Lindquist (2014) for more details. We select $p = \lceil n^{0.8} \rceil$ of basis functions in each approach above, where $\lceil \cdot \rceil$ takes the smallest integer that is greater than the argument. The power 0.8 is greater than the (asymptotic) optimal rate 0.4, because the nuclear norm penalty alleviates the issue of overfitting. As the result, there are 329 basis functions in the approach (**C1**) and 23 in (**C2**).

We compute the matrix $\widehat{\mathbf{\Gamma}}^{b,\omega}$ with expectile level $\tau = 0.1, 0.5, 0.9$ using Y_{ij} and $\boldsymbol{X}_i$ by Algorithm 1 , where Y_{ij} is chosen under either (**C1**) or (**C2**) with $\lambda^{b,\omega}$ selected by the standard 5-fold cross-validation for each region b and each quantile level ω. Please see Appendix D.1 for the exact value of λ for each pair (b, ω). Using SVD $\widehat{\mathbf{\Gamma}}^{b,\omega} = \widehat{\mathbf{U}}_\tau^{b,\omega}\widehat{\mathbf{D}}_\tau^{b,\omega}(\widehat{\mathbf{V}}_\tau^{b,\omega})^\top$, where $(\widehat{\mathbf{V}}_\tau^{b,\omega})^\top$ is regarded as the factor loadings. We note that the size of the matrix $\widehat{\mathbf{V}}_\tau^{b,\omega}$ is 19×19 if we define $Y_{ij}^{b,\omega}$ by following (**C1**), and 4864×4864 by following (**C2**). Note that the sign of the factor loadings cannot be determined exactly (see Remark 2.1).

Remark 4.1 (*On the Computation of* λ)**.** The model error of the BOLD signals typically demonstrates autocorrelation following AR(k) or ARMA(1,1) (Lindquist, 2008 page 446) under the temporal resolution 2 s. A major consequence of the presence of temporal correlation is that the usual cross-validation could potentially underestimate λ, which leads to undersmoothing and overfitting (Opsomer et al., 2001 Section 2). This problem is especially important for the setting (**C2**), where the dimensionality is high because we separate each task. However, we observe that the estimated number of factors for the setting (**C2**) is typically very sparse (less than five factors). Overall, the overfitting does not cause a big issue and the usual cross-validation works well in our model.

4.2.2. Predicting risk attitude

To evaluate the prediction performance, we need to obtain the subjects' risk attitude β_s, where $s = 1, \ldots, 19$ denotes the subject. We follow the approach of Majer et al. (2016) and estimate β_s using the investment decisions made by the subjects to each task with logistic regression; see Appendix D.2 for more details. In essence, higher β_s means the subject s is *less* risk-averse.

In order to use the loadings $\widehat{\mathbf{V}}_\tau^{b,\omega}$ to predict β_s, we apply the standard linear regression models. In particular, in the case (**C1**), we construct a model for β_s using the first two factor loadings

$$\begin{aligned}\beta_s = \alpha_0^{\omega,\tau} &+ \alpha_{11}^{\omega,\tau}\left|(\widehat{\mathbf{V}}_\tau^{1,\omega})_{s1}\right| + \alpha_{12}^{\omega,\tau}\left|(\widehat{\mathbf{V}}_\tau^{2,\omega})_{s1}\right| + \alpha_{13}^{\omega,\tau}\left|(\widehat{\mathbf{V}}_\tau^{3,\omega})_{s1}\right| \\ &+ \alpha_{21}^{\omega,\tau}\left|(\widehat{\mathbf{V}}_\tau^{1,\omega})_{s2}\right| + \alpha_{22}^{\omega,\tau}\left|(\widehat{\mathbf{V}}_\tau^{2,\omega})_{s2}\right| + \alpha_{23}^{\omega,\tau}\left|(\widehat{\mathbf{V}}_\tau^{3,\omega})_{s2}\right| + \varepsilon_s, \quad s = 1, \ldots, 19, \end{aligned} \tag{M1}$$

where $\{\alpha_0^{\omega,\tau}, \alpha_{11}^{\omega,\tau}, \alpha_{12}^{\omega,\tau}, \alpha_{13}^{\omega,\tau}, \alpha_{21}^{\omega,\tau}, \alpha_{22}^{\omega,\tau}, \alpha_{23}^{\omega,\tau}\} \in \mathbb{R}^7$ are the intercept and the coefficients associated with the regions left and right Anterior insula, and dorsomedial prefrontal cortex.

In the case (**C2**), define the averaged loadings of all tasks for each s

$$\mu_{s,k}^{b,\omega,\tau} \overset{\text{def}}{=} \frac{1}{256}\sum_{q=1}^{256}\left|(\widehat{\mathbf{V}}_\tau^{b,\omega})_{256(s-1)+q,k}\right|.$$

We construct another model for β_s using $\mu_{s,k}^{b,\omega,\tau}$:

$$\begin{aligned}\beta_s = \bar{\alpha}_0^{\omega,\tau} &+ \bar{\alpha}_{11}^{\omega,\tau}\mu_{s,1}^{1,\omega,\tau} + \bar{\alpha}_{12}^{\omega,\tau}\mu_{s,1}^{2,\omega,\tau} + \bar{\alpha}_{13}^{\omega,\tau}\mu_{s,1}^{3,\omega,\tau} \\ &+ \bar{\alpha}_{21}^{\omega,\tau}\mu_{s,2}^{1,\omega,\tau} + \bar{\alpha}_{22}^{\omega,\tau}\mu_{s,2}^{2,\omega,\tau} + \bar{\alpha}_{23}^{\omega,\tau}\mu_{s,2}^{3,\omega,\tau} + \varepsilon_s, \quad s = 1, \ldots, 19, \end{aligned} \tag{M2}$$

where $\{\bar{\alpha}_0^{\omega,\tau}, \bar{\alpha}_{11}^{\omega,\tau}, \bar{\alpha}_{12}^{\omega,\tau}, \bar{\alpha}_{13}^{\omega,\tau}, \bar{\alpha}_{21}^{\omega,\tau}, \bar{\alpha}_{22}^{\omega,\tau}, \bar{\alpha}_{23}^{\omega,\tau}\} \in \mathbb{R}^7$. We take the absolute value of the loadings $\widehat{\mathbf{V}}_\tau^{b,\omega}$ because we are only interested in the magnitude of the loadings, which describes the importance of the factors.

Remark 4.2. If sufficiently many subjects are available, then ideally we could use all the estimated factors as suggested by one of our referees. However, because we have only 19 subjects, the number of factor loadings that can be included is very limited. For example, according to the results of an extensive simulation study shown in Table 1 on page 438 in Knofczynski and Mundfrom (2008), the maximum number of predictors that guarantees the best prediction performance is perhaps only around 9 to 12, given the sample size 19. In an unreported analysis, we checked the out-of-sample performance of the models that include up to 3 and 4 factors loadings. We are not able to find strong evidences that more factor loadings improve the prediction performance.

4.2.3. In-sample and out-of-sample performance

We compare the in-sample and out-of-sample performance of the models (M1) and (M2). For the in-sample performance, R^2 of both regressions (M1) and (M2) are computed. In addition, in order to determine whether (M1) and (M2) correctly predict the *order* of risk-aversion of the subjects (rather than the exact value of β_s), we calculate the Spearman's and Kendall's rank correlations between the fitted $\widehat{\beta}_s$ (in-sample) and β_s.

To measure the out-of-sample performance, we calculate $\{\widetilde{\beta}_s\}_{s=1}^{19}$ by a leave-one-out procedure. The steps are as below:

(1) Fix s, where $s = 1, \ldots, 19$. Use the values of the remaining 18 subjects to compute the coefficients $\{\alpha_0^{\omega,\tau}, \alpha_{11}^{\omega,\tau}, \alpha_{12}^{\omega,\tau}, \alpha_{13}^{\omega,\tau}, \alpha_{21}^{\omega,\tau}, \alpha_{22}^{\omega,\tau}, \alpha_{23}^{\omega,\tau}\}$ in model (M1) or $\{\bar{\alpha}_0^{\omega,\tau}, \bar{\alpha}_{11}^{\omega,\tau}, \bar{\alpha}_{12}^{\omega,\tau}, \bar{\alpha}_{13}^{\omega,\tau}, \bar{\alpha}_{21}^{\omega,\tau}, \bar{\alpha}_{22}^{\omega,\tau}, \bar{\alpha}_{23}^{\omega,\tau}\}$ in model (M2) by the standard linear regression.
(2) Compute $\widetilde{\beta}_s$ by plugging in the coefficients computed in the last step in models (M1) and (M2), and input the loadings of the sth subject.
(3) Repeat steps (1) and (2) for each $s = 1, \ldots, 19$.
(4) Calculate the Spearman's and Kendall's rank correlations between $\{\widetilde{\beta}_s\}_{s=1}^{19}$ and $\{\beta_s\}_{s=1}^{19}$.

4.3. Empirical results

In Table 4.1, we present the in-sample fitting and out-of-sample performance for models (M1) and (M2) with the constrained model that uses only the 1st factor ($\alpha_{21}^{\omega,\tau} = \alpha_{22}^{\omega,\tau} = \alpha_{23}^{\omega,\tau} = 0$ in (M1) and $\bar{\alpha}_{21}^{\omega,\tau} = \bar{\alpha}_{22}^{\omega,\tau} = \bar{\alpha}_{23}^{\omega,\tau} = 0$ in (M2)) and the unconstrained model, under various (τ, ω) pairs.

For the in-sample fitting, cases $\omega = 0.1$ and $\omega = 0.9$ outperform the case $\omega = 0.5$. This shows that both extreme negative or positive BOLD can lead to good fitting for models (M1) and (M2). In particular, the fitting performance is the best when $\tau = 0.9$ for $\omega = 0.9$ and $\tau = 0.1$ for $\omega = 0.1$, which correspond to the upper boundary of the red area and the lower boundary of the blue area in each of the three panels in Fig. 4.1.

For the out-of-sample performance, the constrained (M2) using only the first factor with the negative BOLD ($\omega = 0.1$, $\tau = 0.1$) nearly always outperforms all the other cases. In contrast, positive BOLD ($\omega = 0.9$) under the same model performs poorly. Moreover, the unconstrained model improves the prediction performance in most cases, particularly for (M2) under $\omega = 0.9$ and $\tau = 0.9$.

Majer et al. (2016) estimate a dynamic semiparametric factor model and extract the resulting factor loadings to predict the subjects' risk attitude. They evaluate the in-sample fitting (with all 19 subjects) by $R^2 = 0.47$ for a special case of our (M1) ($\tau = 0.5$ and $\alpha_{21}^{\omega,\tau} = \alpha_{22}^{\omega,\tau} = \alpha_{23}^{\omega,\tau} = 0$). Their fitting performance beats all the R^2 in our results, but we are able to describe the *predictive* abilities at several levels of τ, instead of only looking at $\tau = 0.5$. Our findings successfully confirm that the tails of the BOLD signals are more informative than their means in predicting the risk attitude.

5. Conclusions

In this paper, we propose a factorizable multivariate expectile regression method for the high-dimensional cross-sectional or spatial data with sparse latent factors. Fast iterative shrinkage-thresholding algorithm is applied to estimate the model. The convergence of the algorithm and the non-asymptotic theoretical guarantee of the estimator are established. We apply our method on the fMRI data obtained from an investment decisions making experiment, and study the ranking accuracy of the subjects' risk preference using the factor loadings of the extreme BOLD responses. The results show that the negative BOLD signals could provide comparable prediction performance as the positive BOLD signals. This provides insights into the on-going debate on the meaning of the negative BOLD responses.

There are several possibilities for the future research. As many data in practice are time series, there is a need to relax the i.i.d. assumption and make our method compatible with richer temporal structure. Statistical inference is also an important issue for many applications.

Acknowledgments

Financial support from the Deutsche Forschungsgemeinschaft via CRC 649 "Economic Risk" and IRTG 1792 "High Dimensional Non Stationary Time Series", Humboldt-Universität zu Berlin, is gratefully acknowledged. Shih-Kang Chao is partially supported by the Office of Naval Research of the U.S.A (ONR N00014-15-1-2331).

Appendix A. Proofs for Section 2.2

A.1. Proof for Theorem 2.1

Theorem 4.4 in Beck and Teboulle (2009) gives the upper bound of the loss difference at iteration t by

$$|F(\mathbf{\Gamma}_t) - F(\widehat{\mathbf{\Gamma}})| \leq \frac{2L_{\nabla g}\|\mathbf{\Gamma}_0 - \widehat{\mathbf{\Gamma}}\|_{\mathrm{F}}^2}{(t+1)^2}, \tag{A.1}$$

where $L_{\nabla g}$ is the Lipschitz constant of $\nabla g(\mathbf{\Gamma})$ defined in (2.9).

We note that

$$\rho_\tau'(u) = \begin{cases} 2\tau u & \text{for } u \geq 0; \\ 2(1-\tau)u & \text{for } u < 0. \end{cases} \tag{A.2}$$

Hence, the gradient is

$$\nabla g(\mathbf{\Gamma}) = -(mn)^{-1}\mathbf{X}^\top\left\{\mathbf{W} \circ (\mathbf{Y} - \mathbf{X}\mathbf{\Gamma})\right\}, \tag{A.3}$$

where $\mathbf{W}(\mathbf{\Gamma}) = (w_{ij}) \in \mathbb{R}^{n\times m}$, $w_{ij} \stackrel{\text{def}}{=} 2\left\{\tau + \mathbf{1}(Y_{ij} \leq \boldsymbol{X}_i^\top \mathbf{\Gamma}_{\cdot j})(1-2\tau)\right\}$, "$\circ$" represents the Hadamard product.

To simplify the notations, define $\mathbf{U}(\mathbf{\Gamma}) = (Y_{ij} - \boldsymbol{X}_i^\top \mathbf{\Gamma}_{\cdot j}) \in \mathbb{R}^{n\times m}$. For all $\mathbf{\Gamma}_1, \mathbf{\Gamma}_2 \in \mathbb{R}^{p\times m}$, let $\mathbf{U}_1 = \mathbf{U}(\mathbf{\Gamma}_1)$, $\mathbf{U}_2 = \mathbf{U}(\mathbf{\Gamma}_2)$, $\mathbf{W}_1 = \mathbf{W}(\mathbf{\Gamma}_1)$ and $\mathbf{W}_2 = \mathbf{W}(\mathbf{\Gamma}_2)$.

$$\begin{aligned} \|\nabla g(\mathbf{\Gamma}_1) - \nabla g(\mathbf{\Gamma}_2)\|_{\mathrm{F}} &= (mn)^{-1}\|\mathbf{X}^\top(\mathbf{W}_1 \circ \mathbf{U}_1) - \mathbf{X}^\top(\mathbf{W}_2 \circ \mathbf{U}_2)\|_{\mathrm{F}} \\ &\leq (mn)^{-1}\|\mathbf{X}\|_{\mathrm{F}}\|\mathbf{W}_1 \circ \mathbf{U}_1 - \mathbf{W}_2 \circ \mathbf{U}_2\|_{\mathrm{F}} \quad \text{(by submultiplicity)} \end{aligned}$$

$$
\begin{aligned}
&= (mn)^{-1}\|\mathbf{X}\|_{\mathrm{F}}\Big[\sum_{i=1}^{n}\sum_{j=1}^{m}\big\{\rho'_\tau(u_{1,ij}) - \rho'_\tau(u_{2,ij})\big\}^2\Big]^{1/2} \\
&\le (mn)^{-1}\|\mathbf{X}\|_{\mathrm{F}}\Big[\sum_{i=1}^{n}\sum_{j=1}^{m}\big\{2\max(\tau, 1-\tau)\big\}^2(u_{1,ij} - u_{2,ij})^2\Big]^{1/2} \\
&= 2(mn)^{-1}\max(\tau, 1-\tau)\|\mathbf{X}\|_{\mathrm{F}}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\Gamma}_1 - (\mathbf{Y} - \mathbf{X}\boldsymbol{\Gamma}_2)\|_{\mathrm{F}} \\
&\le 2(mn)^{-1}\max(\tau, 1-\tau)\|\mathbf{X}\|_{\mathrm{F}}^2\|\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2\|_{\mathrm{F}} \quad \text{(by submultiplicity)}, \qquad \text{(A.4)}
\end{aligned}
$$

where the fourth line makes use of the fact that $\rho'_\tau(u)$ is Lipschitz continuous with Lipschitz constant $2\max(\tau, 1-\tau)$, see Chao et al. (2017).

Plug $L_{\nabla g} = 2(mn)^{-1}\max(\tau, 1-\tau)\|\mathbf{X}\|_{\mathrm{F}}^2$ into (A.1) yields

$$
|F(\boldsymbol{\Gamma}_t) - F(\widehat{\boldsymbol{\Gamma}})| \le \frac{4(mn)^{-1}\max(\tau, 1-\tau)\|\mathbf{X}\|_{\mathrm{F}}^2\|\boldsymbol{\Gamma}_0 - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}^2}{(t+1)^2}. \qquad \text{(A.5)}
$$

Moreover, setting the right hand side of (A.5) to be ϵ ($\forall \epsilon > 0$) and solving for t gives

$$
t \ge \frac{2\sqrt{\max(\tau, 1-\tau)}\|\mathbf{X}\|_{\mathrm{F}}\|\boldsymbol{\Gamma}_0 - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}}{\sqrt{mn\epsilon}} - 1. \quad \square \qquad \text{(A.6)}
$$

A.2. Proof for Theorem 2.2

Following the proof of Theorem 1 in Fadili and Peyré (2011), define

$$
I(\boldsymbol{\Gamma}_t) \stackrel{\text{def}}{=} g(\boldsymbol{\Gamma}_t) - g(\widehat{\boldsymbol{\Gamma}}) - \langle\langle \nabla g(\boldsymbol{\Gamma}_t), \boldsymbol{\Gamma}_t - \widehat{\boldsymbol{\Gamma}}\rangle\rangle, \qquad \text{(A.7)}
$$

$$
J(\boldsymbol{\Gamma}_t) \stackrel{\text{def}}{=} h(\boldsymbol{\Gamma}_t) - h(\widehat{\boldsymbol{\Gamma}}) + \langle\langle \nabla g(\boldsymbol{\Gamma}_t), \boldsymbol{\Gamma}_t - \widehat{\boldsymbol{\Gamma}}\rangle\rangle, \qquad \text{(A.8)}
$$

the sum of them gives

$$
I(\boldsymbol{\Gamma}_t) + J(\boldsymbol{\Gamma}_t) = F(\boldsymbol{\Gamma}_t) - F(\widehat{\boldsymbol{\Gamma}}). \qquad \text{(A.9)}
$$

According to Lemma C.2, we have

$$
\begin{aligned}
I(\boldsymbol{\Gamma}_t) &\ge \kappa\|\boldsymbol{\Gamma}_t - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}^2 \\
&= \frac{1}{9}m^{-1}\min(\tau, 1-\tau)\sigma_{\min}(\boldsymbol{\Sigma})\|\boldsymbol{\Gamma}_t - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}^2, \qquad \text{(A.10)}
\end{aligned}
$$

where the second line holds with probability greater than $1 - a_n$ under (A1).

Since $\widehat{\boldsymbol{\Gamma}}$ is the optimizer of (2.6), therefore,

$$
\mathbf{0} \in \nabla g(\widehat{\boldsymbol{\Gamma}}) + \nabla h(\widehat{\boldsymbol{\Gamma}}), \qquad \text{(A.11)}
$$

which implies

$$
-\nabla g(\widehat{\boldsymbol{\Gamma}}) \in \nabla h(\widehat{\boldsymbol{\Gamma}}). \qquad \text{(A.12)}
$$

As a result, we have

$$
h(\boldsymbol{\Gamma}_t) - h(\widehat{\boldsymbol{\Gamma}}) \ge \langle\langle -\nabla g(\boldsymbol{\Gamma}_t), \boldsymbol{\Gamma}_t - \widehat{\boldsymbol{\Gamma}}\rangle\rangle, \qquad \text{(A.13)}
$$

i.e., $J(\boldsymbol{\Gamma}_t) \ge 0$.

Plugging (A.10) and (A.13) into (A.9) yields,

$$
\begin{aligned}
\|\boldsymbol{\Gamma}_t - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}^2 &\le \frac{9\,m}{\min(\tau, 1-\tau)\sigma_{\min}(\boldsymbol{\Sigma})}\big\{F(\boldsymbol{\Gamma}_t) - F(\widehat{\boldsymbol{\Gamma}})\big\} \\
&\le \frac{36}{n(t+1)^2}\frac{\max(\tau, 1-\tau)}{\min(\tau, 1-\tau)}\frac{\|\mathbf{X}\|_{\mathrm{F}}^2}{\sigma_{\min}(\boldsymbol{\Sigma})}\|\boldsymbol{\Gamma}_0 - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}^2, \qquad \text{(A.14)}
\end{aligned}
$$

with probability greater than $1 - a_n$. The second line comes from the result of Theorem 2.1. $\square$

Appendix B. Proof for Theorem 2.3

By triangle inequality, we have

$$
\|\boldsymbol{\Gamma}_t - \boldsymbol{\Gamma}\|_{\mathrm{F}}^2 = \|\boldsymbol{\Gamma}_t - \widehat{\boldsymbol{\Gamma}} + \widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}\|_{\mathrm{F}}^2 \le 2\|\boldsymbol{\Gamma}_t - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}^2 + 2\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}\|_{\mathrm{F}}^2. \qquad \text{(B.1)}
$$

Combining the results of Lemma B.2 and Theorem 2.2, it follows that

$$\begin{aligned}\|\mathbf{\Gamma}_t - \mathbf{\Gamma}\|_{\mathrm{F}}^2 \leq & 18^3 c^2 \frac{p+m}{n} \frac{\max(\tau, 1-\tau)^2}{\min(\tau, 1-\tau)^2} \frac{\|\mathbf{\Sigma}\|}{\sigma_{\min}(\mathbf{\Sigma})^2} K_u^2 \dim(\overline{\mathcal{M}}) \\ & + 144c \sqrt{\frac{p+m}{n}} \frac{\max(\tau, 1-\tau)}{\min(\tau, 1-\tau)} \frac{\sqrt{\|\mathbf{\Sigma}\|}}{\sigma_{\min}(\mathbf{\Sigma})} K_u \|\mathbf{\Gamma}_{\mathcal{M}^\perp}\|_* \\ & + \frac{72}{n(t+1)^2} \frac{\max(\tau, 1-\tau)}{\min(\tau, 1-\tau)} \frac{\|\mathbf{X}\|_{\mathrm{F}}^2}{\sigma_{\min}(\mathbf{\Sigma})} \|\mathbf{\Gamma}_0 - \widehat{\mathbf{\Gamma}}\|_{\mathrm{F}}^2, \end{aligned} \tag{B.2}$$

holds with probability greater than $1 - 3 \times 8^{-(p+m)} - a_n$.
Furthermore, given

$$\|\mathbf{\Gamma}_0 - \widehat{\mathbf{\Gamma}}\|_{\mathrm{F}}^2 = \|\mathbf{\Gamma}_0 - \mathbf{\Gamma} + \mathbf{\Gamma} - \widehat{\mathbf{\Gamma}}\|_{\mathrm{F}}^2 \leq 2\|\mathbf{\Gamma}_0 - \mathbf{\Gamma}\|_{\mathrm{F}}^2 + 2\|\mathbf{\Gamma} - \widehat{\mathbf{\Gamma}}\|_{\mathrm{F}}^2, \tag{B.3}$$

and applying Lemma B.2 again we complete the proof of Theorem 2.3. □

Now we show auxiliary results used in the proof of Theorem 2.3. The next theorem is an application of Theorem 1 of Negahban et al. (2012).

Theorem B.1 (*Error Bounds for the Estimator*). *Under (A1), for any $q \in \{1, \ldots, p \wedge m\}$, any optimal solution $\widehat{\mathbf{\Gamma}}$ in the problem (2.6) with $\lambda \geq 2\|\nabla g(\mathbf{\Gamma})\|$ satisfies the bound*

$$\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_{\mathrm{F}}^2 \leq \frac{9\, m^2 \lambda^2}{\left\{c_1 \min(\tau, 1-\tau)\sigma_{\min}(\mathbf{\Sigma})\right\}^2} q + \frac{36\, m\lambda}{\min(\tau, 1-\tau)\sigma_{\min}(\mathbf{\Sigma})} \sum_{j=q+1}^{p \wedge m} \sigma_j(\mathbf{\Gamma}), \tag{B.4}$$

with probability greater than $1 - a_n$, where $\sigma_j(\mathbf{\Gamma})$ is the jth singular value of $\mathbf{\Gamma}$.

Proof for Theorem B.1. The proof is an application of Theorem 1 of Negahban et al. (2012). First, we observe that the nuclear norm is *decomposable* in the sense that

$$\|\mathbf{\Gamma} + \mathbf{\Delta}\|_* = \|\mathbf{\Gamma}\|_* + \|\mathbf{\Delta}\|_*, \forall \mathbf{\Gamma} \in \mathcal{M}_q, \mathbf{\Delta} \in \overline{\mathcal{M}}_q^\perp, \tag{B.5}$$

where

$$\begin{aligned} \mathcal{M}_q &= \mathcal{M}(U_q, V_q) \stackrel{\text{def}}{=} \{\mathbf{\Theta} \in \mathbb{R}^{p \times m} | \operatorname{col}(\mathbf{\Theta}) \subseteq U_q, \operatorname{row}(\mathbf{\Theta}) \subseteq V_q\}, \\ \overline{\mathcal{M}}_q^\perp &= \overline{\mathcal{M}}^\perp(U_q, V_q) \stackrel{\text{def}}{=} \{\mathbf{\Theta} \in \mathbb{R}^{p \times m} | \operatorname{col}(\mathbf{\Theta}) \subseteq U_q^\perp, \operatorname{row}(\mathbf{\Theta}) \subseteq V_q^\perp\}, \end{aligned} \tag{B.6}$$

where $\operatorname{row}(\mathbf{\Theta})$ and $\operatorname{col}(\mathbf{\Theta})$ denote the row and column spaces of $\mathbf{\Theta}$. It can be seen that $\mathcal{M}_q \subset \overline{\mathcal{M}}_q$ where $\overline{\mathcal{M}}_q \stackrel{\text{def}}{=} \{\mathbf{\Theta} \in \mathbb{R}^{p \times m} | \operatorname{tr}(\mathbf{\Theta}^\top \mathbf{S}) = 0,\ \forall \mathbf{S} \in \overline{\mathcal{M}}_q^\perp\}$. Similarly, $\mathcal{M}_q^\perp \stackrel{\text{def}}{=} \{\mathbf{\Theta} \in \mathbb{R}^{p \times m} | \operatorname{tr}(\mathbf{\Theta}^\top \mathbf{S}) = 0,\ \forall \mathbf{S} \in \mathcal{M}_q\}$.

We will verify its conditions (G1) and (G2). For condition (G1), it is already mentioned above that the nuclear norm $\|\cdot\|_*$ is decomposable with respect to $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$ defined in (B.6). For condition (G2), note that on the event

$$\Omega_1 \stackrel{\text{def}}{=} \left\{\sigma_{\min}\left(\frac{\mathbf{X}^\top \mathbf{X}}{n}\right) \geq c_1 \sigma_{\min}(\mathbf{\Sigma}), \sigma_{\max}\left(\frac{\mathbf{X}^\top \mathbf{X}}{n}\right) \leq c_2 \sigma_{\max}(\mathbf{\Sigma})\right\}, \tag{B.7}$$

the loss function g is restrictive strongly convex with coefficients κ and $\xi = 0$ (we replace $\tau_{\mathcal{L}}$ in Negahban et al. (2012) by ξ) shown in Lemma C.2. Since we measure the error in the Frobenius norm $\|\cdot\|_{\mathrm{F}}$, the subspace compatibility constant (Definition 3 of Negahban et al., 2012) is

$$\Psi(\overline{\mathcal{M}}_q) \stackrel{\text{def}}{=} \sup_{\mathbf{S} \in \overline{\mathcal{M}}_q} \frac{\|\mathbf{S}\|_*}{\|\mathbf{S}\|_{\mathrm{F}}} \leq \sqrt{q}.$$

The conclusion of this theorem follows from Theorem 1 of Negahban et al. (2012). □

Lemma B.1. *Under (A1)–(A3),*

$$\mathrm{P}\left(\|\nabla g(\mathbf{\Gamma})\| \leq cm^{-1} \max(\tau, 1-\tau) K_u \sqrt{\|\mathbf{\Sigma}\|} \sqrt{\frac{p+m}{n}}\right) \geq 1 - 3 \times 8^{-(p+m)} - a_n, \tag{B.8}$$

where $c > 0$ is an absolute constant.

Proof for Lemma B.1. Throughout the proof, we restrict on the event Ω_1 in (B.7). Recall the expression from (A.3) that

$$\nabla g(\mathbf{\Gamma}) = -(mn)^{-1} \mathbf{X}^\top \left\{\mathbf{W} \circ (\mathbf{Y} - \mathbf{X}\mathbf{\Gamma})\right\}$$

and the matrix $\mathbf{U}(\mathbf{\Gamma}) = (u_{ij}) = (Y_{ij} - \boldsymbol{X}_i^\top \mathbf{\Gamma}_{\cdot j}) \in \mathbb{R}^{n\times m}$. Following the proof of Lemma 3 in Negahban and Wainwright (2011), we have

$$\begin{aligned}
\mathrm{P}\left(n^{-1}\|\mathbf{X}^\top(\mathbf{W}\circ\mathbf{U})\| \geq 4s\right) &= \mathrm{P}\left(\sup_{\substack{\boldsymbol{\beta}\in\mathcal{S}^{p-1},\\ \boldsymbol{\alpha}\in\mathcal{S}^{m-1}}} n^{-1}|\boldsymbol{\beta}^\top\mathbf{X}^\top(\mathbf{W}\circ\mathbf{U})\boldsymbol{\alpha}| \geq 4s\right)\\
&\leq 8^{p+m}\sup_{\substack{\boldsymbol{\beta}\in\mathcal{S}^{p-1},\\ \boldsymbol{\alpha}\in\mathcal{S}^{m-1}}} \mathrm{P}\left(n^{-1}|\langle\mathbf{X}\boldsymbol{\beta}, (\mathbf{W}\circ\mathbf{U})\boldsymbol{\alpha}\rangle| \geq s\right)\\
&\leq 8^{p+m}\sup_{\substack{\boldsymbol{\beta}\in\mathcal{S}^{p-1},\\ \boldsymbol{\alpha}\in\mathcal{S}^{m-1}}} \mathrm{P}\left(n^{-1}\sum_{i=1}^{n}\langle\boldsymbol{\beta}, \boldsymbol{X}_i\rangle\langle\boldsymbol{\alpha}, (\mathbf{W}\circ\mathbf{U})_i\rangle \geq s\right),
\end{aligned} \tag{B.9}$$

where $\mathcal{S}^{m-1} \overset{\text{def}}{=} \{\boldsymbol{\alpha} \in \mathbb{R}^m : \|\boldsymbol{\alpha}\|_2 = 1\}$ is the Euclidean sphere in m-dimensions. $\forall s \geq 0$, there exists $C > 0$ such that $\mathrm{P}(|u_{ij}| > s) \leq \exp(1 - s^2/C^2)$. Since $|w_{ij}| \leq \max(\tau, 1-\tau)$, we have

$$\begin{aligned}
\mathrm{P}\left(|w_{ij}u_{ij}| > s\right) &\leq \mathrm{P}\left(\max(\tau, 1-\tau)|u_{ij}| > s\right)\\
&= \mathrm{P}\left(|u_{ij}| > \frac{s}{\max(\tau, 1-\tau)}\right)\\
&\leq \exp\left(1 - \frac{s^2}{\max(\tau, 1-\tau)^2 C^2}\right).
\end{aligned} \tag{B.10}$$

It means for each $j \in \{1, \ldots, m\}$, $w_{ij}u_{ij}$ are sub-gaussian. Moreover, the maximal sub-gaussian norm is bounded by

$$\begin{aligned}
\max_{1\leq j\leq m}\|w_{ij}u_{ij}\|_{\psi_2} &= \max_{1\leq j\leq m}\sup_{p\geq 1} p^{-1/2}\left(\mathsf{E}|w_{ij}u_{ij}|^p\right)^{1/p}\\
&\leq \max(\tau, 1-\tau)\max_{1\leq j\leq m}\sup_{p\geq 1} p^{-1/2}\left(\mathsf{E}|u_{ij}|^p\right)^{1/p}\\
&= \max(\tau, 1-\tau)K_u.
\end{aligned} \tag{B.11}$$

Then by Hoeffding's inequality (Proposition 5.10 of Vershynin, 2012b), we can conclude that $\langle\boldsymbol{\alpha}, (\mathbf{W}\circ\mathbf{U})_i\rangle$ is also sub-guassian,

$$\begin{aligned}
\mathrm{P}\left(\langle\boldsymbol{\alpha}, (\mathbf{W}\circ\mathbf{U})_i\rangle \geq s\right) &= \mathrm{P}\left(\left|\sum_{j=1}^{m}\alpha_j w_{ij}u_{ij}\right| \geq s\right)\\
&\leq \exp\left(1 - \frac{C's^2}{\max(\tau, 1-\tau)^2 K_u^2\|\boldsymbol{\alpha}\|_2^2}\right)\\
&= \exp\left(1 - \frac{C's^2}{\max(\tau, 1-\tau)^2 K_u^2}\right),
\end{aligned} \tag{B.12}$$

where $C' > 0$ is an absolute constant. Furthermore, its sub-gaussian norm is bounded by

$$\begin{aligned}
\left\|\langle\boldsymbol{\alpha}, (\mathbf{W}\circ\mathbf{U})_i\rangle\right\|_{\psi_2} &= \sup_{p\geq 1} p^{-1/2}\left\{\mathsf{E}\left|\langle\boldsymbol{\alpha}, (\mathbf{W}\circ\mathbf{U})_i\rangle\right|^p\right\}^{1/p}\\
&= \sup_{p\geq 1} p^{-1/2}\left(\mathsf{E}\left|\sum_{j=1}^{m}\alpha_j w_{ij}u_{ij}\right|^p\right)^{1/p}\\
&\leq \max(\tau, 1-\tau)\sup_{p\geq 1} p^{-1/2}\left(\mathsf{E}\left|\sum_{j=1}^{m}\alpha_j u_{ij}\right|^p\right)^{1/p}\\
&\leq \max(\tau, 1-\tau)MK_u,
\end{aligned} \tag{B.13}$$

where $M > 0$ is an absolute constant. The last line comes from Khintchine inequality (Corollary 5.12 of Vershynin, 2012b) and recall that $\|\boldsymbol{\alpha}\|_2 = 1$. Applying Hoeffding's inequality again we can obtain

$$\begin{aligned}
\mathrm{P}\left(n^{-1}\sum_{i=1}^{n}\langle\boldsymbol{\beta}, \boldsymbol{X}_i\rangle\langle\boldsymbol{\alpha}, (\mathbf{W}\circ\mathbf{U})_i\rangle \geq s\right) &\leq \exp\left(1 - \frac{C''s^2 n}{\max(\tau, 1-\tau)^2 M^2 K_u^2 n^{-1}\sum_{i=1}^{n}\langle\boldsymbol{\beta}, \boldsymbol{X}_i\rangle^2}\right)\\
&\leq \exp\left(1 - \frac{C''s^2 n}{\max(\tau, 1-\tau)^2 M^2 K_u^2 n^{-1}\|\mathbf{X}\boldsymbol{\beta}\|_2^2}\right),
\end{aligned}$$

$$\leq \exp\left(1 - \frac{C''s^2n}{c_2 \max(\tau, 1-\tau)^2 M^2 K_u^2 \|\mathbf{\Sigma}\|}\right), \tag{B.14}$$

where C'' is an absolute constant. Combining (B.9) and (B.14) gives

$$\mathrm{P}\left(n^{-1}\left\|\mathbf{X}^\top(\mathbf{W}\circ\mathbf{U})\right\| \geq 4s\right) \leq \exp\left(1 - \frac{C''s^2n}{9\max(\tau, 1-\tau)^2 M^2 K_u^2 \|\mathbf{\Sigma}\|} + (p+m)\log 8\right). \tag{B.15}$$

Set $s = \frac{1}{4}c\max(\tau, 1-\tau)K_u\sqrt{\|\mathbf{\Sigma}\|}\sqrt{\frac{p+m}{n}}$, where $c \overset{\text{def}}{=} 4\cdot\sqrt{2\log 8\frac{9M^2}{C''}}$, then we can conclude from the fact $P(\Omega_1) \geq 1 - a_n$,

$$\begin{aligned}
&\mathrm{P}\left(n^{-1}\left\|\mathbf{X}^\top(\mathbf{W}\circ\mathbf{U})\right\| \leq c\max(\tau, 1-\tau)K_u\sqrt{\|\mathbf{\Sigma}\|}\sqrt{\frac{p+m}{n}}\right)\\
&\quad\geq \left[1 - \exp\big(1 - (p+m)\log 8\big)\right] \times (1 - a_n)\\
&\quad\geq \left[1 - 3\times 8^{-(p+m)}\right] \times (1 - a_n)\\
&\quad\geq 1 - 3\times 8^{-(p+m)} - a_n \quad (\text{as } p+m > 1).
\end{aligned} \tag{B.16}$$

This finishes the proof. □

Lemma B.2. *Under (A1)–(A3), selecting* $\lambda = 2\,\mathrm{cm}^{-1}\max(\tau, 1-\tau)K_u\sqrt{\|\mathbf{\Sigma}\|}\sqrt{\frac{p+m}{n}}$, *for* $n \geq 2\min(m, p)$, *any optimal solution* $\widehat{\mathbf{\Gamma}}$ *in the problem (2.6) satisfies the bound*

$$\begin{aligned}
\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_{\mathrm{F}}^2 \leq& c'\frac{p+m}{n}\frac{\max(\tau, 1-\tau)^2}{\min(\tau, 1-\tau)^2}\frac{\|\mathbf{\Sigma}\|}{\sigma_{\min}(\mathbf{\Sigma})^2}K_u^2\dim(\overline{\mathcal{M}})\\
&+ c'\sqrt{\frac{p+m}{n}}\frac{\max(\tau, 1-\tau)}{\min(\tau, 1-\tau)}\frac{\sqrt{\|\mathbf{\Sigma}\|}}{\sigma_{\min}(\mathbf{\Sigma})}K_u\|\mathbf{\Gamma}_{\mathcal{M}^\perp}\|_*,
\end{aligned} \tag{B.17}$$

with probability greater than $1 - 3\times 8^{-(p+m)} - a_n$, *where* $c, c' > 0$ *are absolute constants.*

Proof of Lemma B.2. Recall that Ω_1 is defined as (B.7), and let the event that (B.8) holds as Ω_2. On event $\Omega_1 \cap \Omega_2$, (B.17) can be achieved by simply plugging $\lambda = 2\,\mathrm{cm}^{-1}\max(\tau, 1-\tau)K_u\sqrt{\|\mathbf{\Sigma}\|}\sqrt{\frac{p+m}{n}}$ into (B.4). We note that

$$\begin{aligned}
\mathrm{P}(\Omega_2\cap\Omega_1) = \mathrm{P}(\Omega_2|\Omega_1)\,\mathrm{P}(\Omega_1) &\geq \left[1 - 3\times 8^{-(p+m)}\right] \times (1 - a_n)\\
&\geq 1 - 3\times 8^{-(p+m)} - a_n \quad (\text{as } p+m > 1). \quad \square
\end{aligned} \tag{B.18}$$

Appendix C. Auxiliary results

Lemma C.1. *For any* $u, \delta \in \mathbb{R}$ *and* $\tau \in (0, 1)$,

$$\rho_\tau(u+\delta) - \rho_\tau(u) - \rho_\tau'(u)\delta \geq \min(\tau, 1-\tau)\delta^2. \tag{C.1}$$

Proof of Lemma C.1. When $u = 0$, we have $\rho_\tau(u) = \rho_\tau'(u) = 0$, therefore

$$\rho_\tau(\delta) = |\tau - \mathbf{1}\{\delta < 0\}|\delta^2 \geq \min(\tau, 1-\tau)\delta^2.$$

If $u > 0, u + \delta < 0$ $(\delta < 0)$, we have

$$\rho_\tau(u+\delta) - \rho_\tau(u) - \rho_\tau'(u)\delta - \min(\tau, 1-\tau)\delta^2 = \begin{cases}(1-2\tau)(\delta+u)^2 \geq 0 & \text{for } \tau \leq 1-\tau;\\ (1-2\tau)(u+2\delta)u > 0 & \text{for } \tau > 1-\tau.\end{cases}$$

If $u > 0, u + \delta > 0$ $(\delta > 0)$, we have

$$\rho_\tau(u+\delta) - \rho_\tau(u) - \rho_\tau'(u)\delta - \min(\tau, 1-\tau)\delta^2 = \begin{cases}(2\tau-1)(u+2\delta)u \geq 0 & \text{for } \tau \leq 1-\tau;\\ (2\tau-1)(u+\delta)^2 u > 0 & \text{for } \tau > 1-\tau.\end{cases}$$

In the other two cases,

$$\rho_\tau(u+\delta) - \rho_\tau(u) - \rho_\tau'(u)\delta = \begin{cases}\tau\delta^2 \geq \min(\tau, 1-\tau)\delta^2 & \text{for } u > 0, u+\delta \geq 0;\\ (1-\tau)\delta^2 \geq \min(\tau, 1-\tau)\delta^2 & \text{for } u < 0, u+\delta \leq 0.\end{cases}$$

Therefore, we can conclude that

$$\rho_\tau(u+\delta) - \rho_\tau(u) - \rho_\tau'(u)\delta \geq \min(\tau, 1-\tau)\delta^2.$$

Table D.1
Tuning parameters by 5-fold cross validation.

		Whole series (C1)			Task-wise (C2)		
τ		0.1	0.5	0.9	0.1	0.5	0.9
$\omega = 0.1$	$aINS_L$	0.0442	0.0552	0.0383	0.0008	0.0006	0.0008
	$aINS_R$	0.0303	0.0421	0.0293	0.0004	0.0008	0.0004
	DMPFC	0.0348	0.0504	0.0198	0.0004	0.0007	0.0006
$\omega = 0.5$	$aINS_L$	0.0181	0.0403	0.0153	0.0004	0.0006	0.0003
	$aINS_R$	0.0137	0.0393	0.0157	0.0006	0.0004	0.0005
	DMPFC	0.0195	0.0391	0.0143	0.0006	0.0002	0.0007
$\omega = 0.9$	$aINS_L$	0.0253	0.0408	0.0275	0.0006	0.0004	0.0004
	$aINS_R$	0.0243	0.0442	0.0200	0.0008	0.0002	0.0006
	DMPFC	0.0193	0.0474	0.0206	0.0005	0.0008	0.0008

Lemma C.2. *$g(\mathbf{\Gamma})$ defined in (2.10) is κ-strongly convex and differentiable with $\kappa = m^{-1}\min(\tau, 1-\tau)\sigma_{\min}(\frac{\mathbf{X}^\top\mathbf{X}}{n})$.* □

Proof of Lemma C.2. Denote $\widetilde{u}_{ij} \overset{\text{def}}{=} Y_{ij} - \boldsymbol{X}_i^\top(\mathbf{\Gamma}_{\cdot j} + \mathbf{\Delta}_{\cdot j})$ and $u_{ij} \overset{\text{def}}{=} Y_{ij} - \boldsymbol{X}_i^\top\mathbf{\Gamma}_{\cdot j}$, for $i = 1, \ldots, n, j = 1, \ldots, m$, we have

$$\begin{aligned}
\langle\langle \nabla g(\mathbf{\Gamma}), \mathbf{\Delta}\rangle\rangle &= \operatorname{tr}\big(\nabla g(\mathbf{\Gamma})^\top \mathbf{\Delta}\big) \\
&= -(mn)^{-1}\sum_{j=1}^{m}\sum_{l=1}^{p}\mathbf{\Delta}_{lj}\sum_{i=1}^{n}\rho'(u_{ij})X_{il} \\
&= -(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\Big\{\sum_{l=1}^{p}\mathbf{\Delta}_{lj}\rho'(u_{ij})X_{il}\Big\} \\
&= -(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\big\{\rho'(u_{ij})\boldsymbol{X}_i^\top\mathbf{\Delta}_{\cdot j}\big\}. \qquad \text{(C.2)}
\end{aligned}$$

Therefore,

$$\begin{aligned}
g(\mathbf{\Gamma}+\mathbf{\Delta}) - g(\mathbf{\Gamma}) - \langle\langle \nabla g(\mathbf{\Gamma}), \mathbf{\Delta}\rangle\rangle &= (mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\big\{\rho(\widetilde{u}_{ij}) - \rho(u_{ij}) + \rho'(u_{ij})\boldsymbol{X}_i^\top\mathbf{\Delta}_{\cdot j}\big\} \\
&\geq (mn)^{-1}\min(\tau, 1-\tau)\sum_{i=1}^{n}\sum_{j=1}^{m}(\boldsymbol{X}_i^\top\mathbf{\Delta}_{\cdot j})^2 \text{ (by Lemma C.1)} \\
&= (mn)^{-1}\min(\tau, 1-\tau)\|\mathbf{X}\mathbf{\Delta}\|_{\mathrm{F}}^2 \\
&= (mn)^{-1}\min(\tau, 1-\tau)\operatorname{tr}(\mathbf{\Delta}^\top\mathbf{X}^\top\mathbf{X}\mathbf{\Delta}) \\
&\geq m^{-1}\min(\tau, 1-\tau)\sigma_{\min}\Big(\frac{\mathbf{X}^\top\mathbf{X}}{n}\Big)\|\mathbf{\Delta}\|_{\mathrm{F}}^2. \quad \square \qquad \text{(C.3)}
\end{aligned}$$

Appendix D. Additional details for Section 4

D.1. Tuning parameters by cross-validation

Choosing $\omega = 0.1, b = 1$ (aINS_L cluster) in (C1) case as an example, Fig. D.1 illustrates the cross-validation error function in terms of λ under different τ levels. The optimal tuning parameters determined by 5-fold cross-validation under all cases are reported in Table D.1.

D.2. Risk attitude parameter

The risk attitude parameter β is estimated by logistic model via maximum likelihood estimation (MLE)

$$\begin{aligned}
\mathrm{P}\{\text{risky choice}|x\} &= \frac{1}{1+\exp[-\sigma\{\bar{x} - \beta S(x) - 5\}]}, \\
\mathrm{P}\{\text{sure choice}|x\} &= 1 - \frac{1}{1+\exp[-\sigma\{\bar{x} - \beta S(x) - 5\}]}, \qquad \text{(D.1)}
\end{aligned}$$

where x is the return stream displayed to the individual, its mean and standard deviation are $\bar{x}$ and $S(x)$.

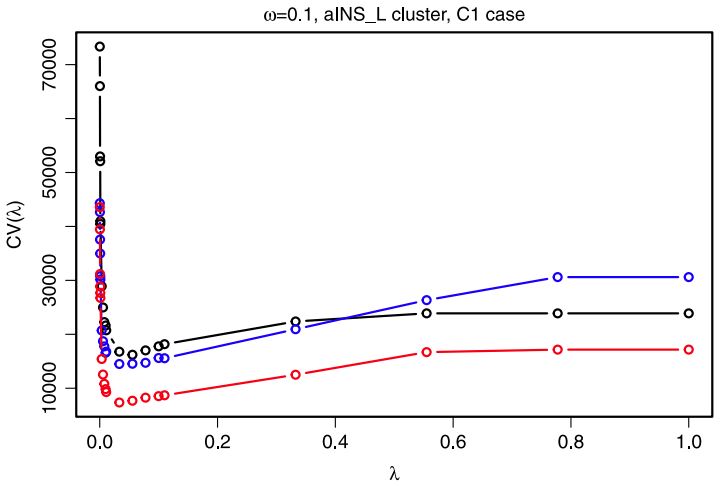


Fig. D.1. The cross-validation error function in terms of tuning parameter λ, with $\tau = 0.1$, 0.5, and 0.9, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

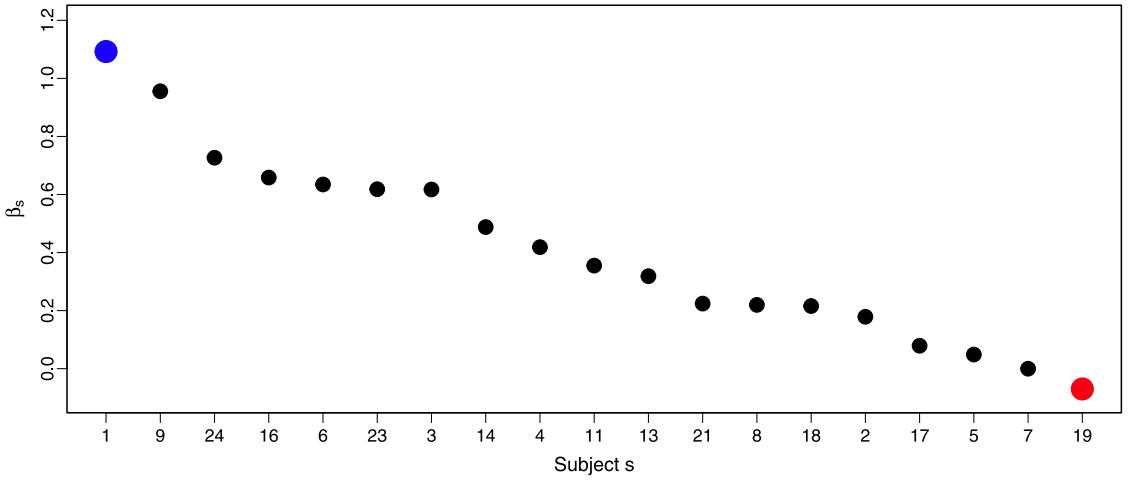


Fig. D.2. Estimated risk attitude for 19 subjects.

The estimated risk attitude parameters for 19 subjects in order are plotted in Fig. D.2, also see Majer et al. (2016). Negative parameters imply risk-seeking behaviors; while positive parameters indicate averse risk patterns. We can see most of the individuals are risk-averse and the two extremes #1 and #19 are the most risk-averse and most risk-seeking persons respectively.

References

Beck, A., Teboulle, M., 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imag. Sci. 2 (1), 183–202.
Camerer, C.F., 2007. Neuroeconomics: Using neuroscience to make economic predictions. Econom. J. 117 (519), C26–C42.
Chao, S.-K., Härdle, W.K., Yuan, M., 2016. Factorisable multi-task quantile regression, SFB 649 Discussion Paper 2016-057, Sonderforschungsbereich 649, Humboldt Universität zu Berlin, Germany. Available at http://sfb649.wiwi.hu-berlin.de/papers/pdf/SFB649DP2016-057.pdf.
Chao, S.-K., Proksch, K., Dette, H., Härdle, W.K., 2017. Confidence corridors for multivariate generalized quantile regression. J. Bus. Econom. Statist. 35 (1), 70–85.
Degras, D., Lindquist, M.A., 2014. A hierarchical model for simultaneous detection and estimation in multi-subject fMRI studies. NeuroImage 98, 61–72.
Fadili, J.M., Peyré, G., 2011. Total variation projection with first order schemes. IEEE Trans. Image Process. 20 (3), 657–669.
Heekeren, H.R., Marrett, S., Ungerleider, L.G., 2008. The neural systems that mediate human perceptual decision making. Nat. Rev. Neurosci. 9 (6), 467–479.
Helfinstein, S.M., Schonberg, T., Congdon, E., Karlsgodt, K.H., Mumford, J.A., Sabb, F.W., Cannon, T.D., London, E.D., Bilder, R.M., Poldrack, R.A., 2014. Predicting risky choices from brain activity patterns. Proc. Natl. Acad. Sci. 111 (7), 2470–2475.
Izenman, A.J., 1975. Reduced-rank regression for the multivariate linear model. J. Multivariate Anal. 5 (2), 248–264.
Ji, S., Ye, J., 2009. An accelerated gradient method for trace norm minimization. In: Proceedings of the 26th International Conference on Machine Learning.

Knofczynski, G.T., Mundfrom, D., 2008. Sample sizes when using multiple linear regression for prediction. Educ. Psychol. Meas. 68 (3), 431–442.
Koenker, R., Bassett, G.W., 1978. Regression quantiles. Econometrica 46 (1), 33–50.
Koenker, R., Portnoy, S., 1990. *M* Estimation of multivariate regressions. J. Amer. Statist. Assoc. 85 (412), 1060–1068.
Lindquist, M.A., 2008. The statistical analysis of fMRI data. Statist. Sci. 23 (4), 439–464.
Majer, P., Mohr, P.N.C., Heekeren, H., Härdle, W.K., 2016. Portfolio decisions and brain reactions via the CEAD method. Psychometrika 81 (3), 881–903.
Mullinger, K., Mayhew, S., Bagshaw, A., Bowtell, R., Francis, S., 2014. Evidence that the negative BOLD response is neuronal in origin: A simultaneous EEG–BOLD–CBF study in humans. NeuroImage 94, 263–274.
Negahban, S.N., Ravikumar, P., Wainwright, M.J., Yu, B., 2012. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. Statist. Sci. 27 (4), 538–557.
Negahban, S.N., Wainwright, M.J., 2011. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. Ann. Statist. 39 (2), 1069–1097.
Newey, W.K., Powell, J.L., 1987. Asymmetric least squares estimation and testing. Econometrica 55 (4), 819–847.
Opsomer, J., Wang, Y., Yang, Y., 2001. Nonparametric regression with correlated errors. Statist. Sci. 16 (2), 134–153.
Reinsel, G.C., Velu, R.P., 1998. Multivariate Reduced-Rank Regression. Springer, New York.
Rossi, G.D., Harvey, A., 2009. Quantiles, expectiles and splines. J. Econometrics 152 (2), 179–185.
Schultz, W., 2015. Neuronal reward and decision signals: From theories to data. Physiol. Rev. 95 (3), 853–951.
van Bömmel, A., Song, S., Majer, P., Mohr, P.N.C., Heekeren, H.R., Härdle, W.K., 2014. Risk patterns and correlated brain activities. multidimensional statistical analysis of fMRI data in economic decision making study. Psychometrika 79 (3), 489–514.
Vershynin, R., 2012a. How close is the sample covariance matrix to the actual covariance matrix? J. Theoret. Probab. 25 (3), 655–686.
Vershynin, R., 2012b. Introduction to the non-asymptotic analysis of random matrices. In: Eldar, Y., Kutyniok, G. (Eds.), Compressed Sensing, Theory and Applications. Cambridge University Press, pp. 210–268 (Chapter 5).
Wainwright, M.J., 2009. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). IEEE Trans. Inform. Theory 55, 2183–2202.
Yuan, M., Ekici, A., Lu, Z., Monteiro, R., 2007. Dimension reduction and coefficient estimation in multivariate linear regression. J. R. Stat. Soc. Ser. B Stat. Methodol. 69 (3), 329–346.

Computational Statistics and Data Analysis 121 (2018) 1–19

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Multivariate factorizable expectile regression with application to fMRI data☆

Shih-Kang Chao [a], Wolfgang K. Härdle [b,c], Chen Huang [d,*]

[a] Department of Statistics, Purdue University, 250 N University St., West Lafayette, IN 47907-2066, USA
[b] Ladislaus von Bortkiewicz Chair of Statistics, C.A.S.E. - Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany
[c] Sim Kee Boon Institute for Financial Economics, Singapore Management University, 50 Stamford Road, Singapore 178899, Singapore
[d] Faculty of Mathematics and Statistics, University of St. Gallen, Bodanstrasse 6, 9000 St. Gallen, Switzerland

ARTICLE INFO

Article history:
Received 10 March 2017
Received in revised form 28 October 2017
Accepted 3 December 2017
Available online 12 December 2017

Keywords:
Multivariate regression
Factor analysis
Expectile regression
Functional magnetic resonance imaging
Risk preference

ABSTRACT

A multivariate expectile regression model is proposed to analyze the tail events of large cross-sectional and spatial data, where the tail events are linked by a latent factor structure. The computational advantage of the method is demonstrated, and the estimation risk is analyzed for every *fixed number of iteration* and *fixed sample size*, when the latent factors are either exactly or approximately sparse. The proposed method is applied on the functional magnetic resonance imaging (fMRI) data taken during an experiment of investment decisions making. It is shown that the negative extreme blood oxygenation level dependent (BOLD) responses may be relevant to the risk preferences.

1. Introduction

Analyzing cross-sectional or spatial data is of critical interest in many scientific fields. Particularly, the interests in these fields are mostly in the *tail events*, which are the extreme events that occur with very small (or very large) probability. For example, in finance, Value-at-Risk (VaR) defined by the 1% quantile of the distribution of investment portfolio is widely used for measuring the market risk. In climatology, one of the major interests is the prediction of extreme precipitation defined by the tail quantile with level very close to 1. The estimation or prediction of tail events is often complicated by high dimensionality, which is common in many modern applications. However, the latent factors that influence all the cross-sections or spatial points may be sparse.

Multivariate regression (Izenman, 1975; Reinsel and Velu, 1998) is a classical tool for analyzing the cross-sectional or spatial data, and the penalization methods with matrix nuclear norm (Yuan et al., 2007; Negahban and Wainwright, 2011; Negahban et al., 2012) is applied to handle high dimensionality. However, the literature in multivariate regression is mostly silent about the estimation and prediction of tail events. On the other hand, quantile regression proposed by Koenker and Bassett (1978) is a well-known method for estimating the conditional quantiles, which is done through optimizing a non-differentiable loss function. Koenker and Portnoy (1990) generalize the quantile regression to a multivariate regression framework, but it cannot be applied to modern high dimensional data. To deal with high dimensional data with certain sparsity structure, it seems necessary to use some penalization methods, but the non-differentiable loss function of quantile regression is less convenient when being optimized with a penalty that is also non-differentiable, such as the nuclear norm.

☆ The codes to implement the algorithms are publicly accessible via the website Qwww.quantlet.de.

* Corresponding author.
E-mail address: chen.huang@unisg.ch (C. Huang).

https://doi.org/10.1016/j.csda.2017.12.001
0167-9473/

In this paper, we propose to estimate the tail events of a factorizable multivariate model using expectile regression (see (2.3) in Section 2 for the specific form of the model). Expectiles illustrate the tail events, and are closely related to quantiles (see, e.g. Section 2 of Rossi and Harvey, 2009). The expectile regression is proposed by Newey and Powell (1987) and is done through optimizing a smooth loss function. The smooth loss function of expectile regression yields computational advantages when being combined with a non-differentiable penalty, which will be shown in the algorithmic convergence analysis in Section 2.2. Furthermore, our method can be easily and efficiently implemented with the fast iterative shrinkage-thresholding algorithm of Beck and Teboulle (2009).

In addition to the convergence analysis, we *jointly* analyze the algorithmic and stochastic risk of our iterative estimator in Theorem 2.3, which characterizes the estimation error for each *fixed sample size* and *fixed number of iteration*. In particular, the theorem shows that our estimator is consistent as long as $\max\{p,\ m\} \ll n$ while $p,\ m \to \infty$, where p is the dimension of the covariates, m is the number of cross-sections or spatial points, and n is the sample size obtained in each cross-section or spatial point. The theorem is established under the weak assumption that the number of latent factors jointly influencing all the cross-sections or spatial points are approximately sparse.

Much interest has concentrated on using the functional magnetic resonance imaging (fMRI) data to understand the risk perception of humans (Heekeren et al., 2008). While the positive blood oxygenation level dependent (BOLD) signals are the focus in most studies, an increasing number of researchers are intrigued by the observed negative BOLD signals and their implications. Many hypotheses on the causes and implications of the negative BOLD are proposed, but they are still highly debatable (Mullinger et al., 2014).

We apply our method on the BOLD signals measured on the human subjects during an experiment on investment decisions making, and shed light on how the negative BOLD responses may be relevant in the decision making process. Using the same data, Majer et al. (2016) retrieve factor loadings from a dynamic factor model of the BOLD signals, and apply these loadings on explaining the subjects' risk attitude. However, their analysis only focus on the mean, and neglect the tail information of the BOLD signals. We apply our method on the BOLD responses obtained from 19 subjects, and estimate the factors and loadings at both high and low extreme expectile levels. We find that the factor loadings from the negative tail of the BOLD signals could not only well explain the revealed risk preference of the subjects in terms of R^2, but also *predict* the revealed risk preference. The prediction performance of the negative extreme BOLD is generally similar to that of the positive extreme BOLD, but sometimes they can be more accurate. Nonetheless, we note that our results do not yield any conclusions on the source of the negative BOLD responses.

The rest of the paper is arranged as follows. Section 2 introduces the model setting, estimation method and theoretical properties of the estimator. Simulation studies of our method are shown in Section 3. Section 4 illustrates the empirical application with the fMRI data. Section 5 concludes this paper. Proofs and auxiliary results are provided in Appendices.

2. Method

2.1. Model

We start with defining some notations. Denote a matrix $\mathbf{S} = (s_{lj}) = [\mathbf{S}_{\cdot 1} \ldots \mathbf{S}_{\cdot m}] \in \mathbb{R}^{p\times m}$, where $\mathbf{S}_{\cdot j} \in \mathbb{R}^p$ are the column vectors. Let $\|\mathbf{S}\|_{\mathrm{F}}$, $\|\mathbf{S}\|_*$ and $\|\mathbf{S}\|$ be the matrix Frobenius, nuclear and spectral norm. Denote $\sigma_{\min}(\mathbf{S})$ and $\sigma_{\max}(\mathbf{S})$ the smallest and largest singular values. For a vector $\mathbf{v} \in \mathbb{R}^p$, $\|\mathbf{v}\|_2$ is the Euclidean norm. Define $\langle\langle \mathbf{A}, \mathbf{B}\rangle\rangle \stackrel{\text{def}}{=} \mathrm{tr}(\mathbf{A}^\top \mathbf{B})$. $\mathbf{I}_m$ is the identity matrix with dimension m.

Let $\{(\boldsymbol{X}_i, Y_{i1}, \ldots, Y_{im})\}_{1\le i\le n}$ be the samples with $Y_{ij} \in \mathbb{R}$ and $\boldsymbol{X}_i \in \mathbb{R}^p$. Specifically, Y_{ij} represents the value observed from the response j at the time point i, and $\{\boldsymbol{X}_i\}_{i=1}^n$ are the covariates. For simplicity, we assume that the samples are i.i.d. over i. For $\tau \in (0, 1)$, the conditional expectile $e_j(\tau|\boldsymbol{X}_i)$ of Y_{ij} given $\boldsymbol{X}_i$ is defined by

$$e_j(\tau|\boldsymbol{X}_i) = \boldsymbol{X}_i^\top \boldsymbol{\gamma}_j(\tau), \tag{2.1}$$

where

$$\boldsymbol{\gamma}_j(\tau) \stackrel{\text{def}}{=} \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^p} \mathsf{E}\big[\rho_\tau(Y_{ij} - \boldsymbol{X}^\top \boldsymbol{\gamma})\big], \tag{2.2}$$

and $\rho_\tau(u) \stackrel{\text{def}}{=} |\tau - \mathbf{1}\{u < 0\}|\, |u|^2$. Define the coefficient matrix

$$\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_\tau \stackrel{\text{def}}{=} [\boldsymbol{\gamma}_1(\tau) \ldots \boldsymbol{\gamma}_m(\tau)].$$

We assume that the expectiles $e_1(\tau|\boldsymbol{X}_i), \ldots, e_m(\tau|\boldsymbol{X}_i)$ are related through a factor model:

$$e_j(\tau|\boldsymbol{X}_i) = \sum_{k=1}^{r} \psi_{j,k}(\tau) f_k^\tau(\boldsymbol{X}_i), \tag{2.3}$$

where $f_k^\tau(\boldsymbol{X}_i)$ is the kth factor, r is the number of factors, and $\psi_{j,k}(\tau)$ are the factor loadings. Furthermore, factors are constructed by linear combinations of covariates $\boldsymbol{X}_i$:

$$f_k^\tau(\boldsymbol{X}_i) = \boldsymbol{\varphi}_k(\tau)^\top \boldsymbol{X}_i \tag{2.4}$$

where $\boldsymbol{\varphi}_k(\tau) = (\varphi_{k,1}(\tau), \ldots, \varphi_{k,p}(\tau))^\top$. By substituting (2.4) into (2.3), it can be seen that the factor structure yields the reparametrization $\boldsymbol{\Gamma}^\top = \boldsymbol{\Psi}_\tau \boldsymbol{\Phi}_\tau$, where the matrix $\boldsymbol{\Psi}_\tau = (\psi_{j,k}(\tau))_{j\le m, k\le r}$ and $\boldsymbol{\Phi}_\tau = (\varphi_{k,l}(\tau))_{k\le r, l\le p}$. Unfortunately, the matrix factorization is in general not unique, so the factors and loadings may not be identifiable from $\boldsymbol{\Gamma}$. We alleviate the identifiability issue by imposing the normalization restrictions as Eq. (2.14) on page 28 of Reinsel and Velu (1998):

$$\boldsymbol{\Psi}_\tau^\top \boldsymbol{\Psi}_\tau = \mathbf{I}_m, \quad \boldsymbol{\Phi}_\tau \boldsymbol{\Phi}_\tau^\top = \operatorname{diag}(\sigma_1(\boldsymbol{\Gamma}), \ldots, \sigma_{p\wedge m}(\boldsymbol{\Gamma})). \tag{2.5}$$

The restrictions (2.5) make the factors and loadings associated with the nonzero singular values of $\boldsymbol{\Gamma}$ identifiable up to sign, if the nonzero singular values are distinct. When there exist repeated singular values, $\boldsymbol{\Psi}_\tau$ and $\boldsymbol{\Phi}_\tau$ cannot be uniquely identified; see Remark 2.1. Given the singular value decomposition $\boldsymbol{\Gamma} = \mathbf{UDV}^\top$, we have $\boldsymbol{\Psi}_\tau = \mathbf{V}$ and $\boldsymbol{\Phi}_\tau = \mathbf{D}^\top \mathbf{U}^\top$. Suppose an estimator $\widehat{\boldsymbol{\Gamma}}$ is available, we can estimate the kth factor by $\widehat{f}_k^\tau(\boldsymbol{X}_i) = \boldsymbol{X}_i^\top \widehat{\boldsymbol{\varphi}}_k(\tau) = \widehat{\sigma}_k \boldsymbol{X}_i^\top \widehat{\mathbf{U}}_{\cdot k}$ and the factor loadings for the jth response by $\widehat{\boldsymbol{\psi}}_j(\tau) = \widehat{\mathbf{V}}_{j\cdot}$, where $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ are unitary matrices obtained from the singular value decomposition $\widehat{\boldsymbol{\Gamma}} = \widehat{\mathbf{U}}\widehat{\mathbf{D}}\widehat{\mathbf{V}}^\top$.

Remark 2.1 (*Identifiability and Free Parameters*)**.** If there exist repeated singular values, then the singular vectors associated with these repeated singular values are not unique, and the factors and loadings are not uniquely identifiable. In particular, suppose the multiplicity of lth singular value $\mu_l > 1$, the number of free parameters for factor loadings (eigenvectors of the right singular spaces) is $\mu_l^2 - \binom{\mu_l}{2} - \mu_l = \mu_l(\mu_l - 1)/2$, where "$\mu_l^2$" is the total number of coefficients that determine the factor loadings associated with the lth singular value, "$-\binom{\mu_l}{2}$" is from the orthogonality constraints and "$-\mu_l$" is from the normalization constraints. Since the sign in the matrix factorization cannot be determined, the sign of the loadings and factors are not identifiable. In our empirical analysis in Section 4, we only use the absolute value of the loadings.

The factor model (2.3) implies that $\boldsymbol{\Gamma}$ is of rank r, and the model (2.1) corresponds to a multivariate linear regression model. For the standard regression with square loss, Reinsel and Velu (1998) propose to estimate $\boldsymbol{\Gamma}$ with the reduced-rank regression under the knowledge of r. However, r is usually unknown in practice. Yuan et al. (2007) propose to perform the multivariate regression with the nuclear norm penalty, which does not require the knowledge of r. The latter inspired the use of the nuclear norm penalty in the next section. However, Yuan et al. (2007) do not provide an algorithm that can scale up to large dimensions.

2.2. Algorithm

To estimate our model under the factor model (2.3), we combine an asymmetric loss with the nuclear norm penalty. To be more specific, we estimate $\boldsymbol{\Gamma}$ (defined in Section 2.1) by solving:

$$\widehat{\boldsymbol{\Gamma}}_{\tau,\lambda} \overset{\text{def}}{=} \arg\min_{\boldsymbol{\Gamma}\in\mathbb{R}^{p\times m}} F(\boldsymbol{\Gamma}), \tag{2.6}$$

$$F(\boldsymbol{\Gamma}) \overset{\text{def}}{=} (mn)^{-1} \sum_{i=1}^{n}\sum_{j=1}^{m} \rho_\tau(Y_{ij} - \boldsymbol{X}_i^\top \boldsymbol{\Gamma}_{\cdot j}) + \lambda \|\boldsymbol{\Gamma}\|_*, \tag{2.7}$$

where λ is a tuning parameter, $\boldsymbol{\Gamma}_{\cdot j}$ is the jth column of $\boldsymbol{\Gamma}$. The second term $\|\boldsymbol{\Gamma}\|_* = \sum_{l=1}^{\min(p,\,m)} \sigma_l(\boldsymbol{\Gamma})$, where the singular values $\sigma_1(\boldsymbol{\Gamma}) \ge \sigma_2(\boldsymbol{\Gamma}) \ge \cdots \ge \sigma_{\min(p,\,m)}(\boldsymbol{\Gamma})$. We note that (2.7) is a convex optimization problem. The number of factors r in (2.3) does not need to be specified. To simplify the notation, we denote $\widehat{\boldsymbol{\Gamma}}$ for $\widehat{\boldsymbol{\Gamma}}_{\tau,\lambda}$ hereinafter.

To solve the optimization problem (2.7), we apply the fast iterative shrinkage-thresholding algorithm (FISTA) of Beck and Teboulle (2009). FISTA solves the optimization problems of the form:

$$\min_{\boldsymbol{\Gamma}} \{g(\boldsymbol{\Gamma}) + h(\boldsymbol{\Gamma})\}, \tag{2.8}$$

where g is a smooth convex function with Lipschitz continuous gradient ∇g,

$$\|\nabla g(\boldsymbol{\Gamma}_1) - \nabla g(\boldsymbol{\Gamma}_2)\|_{\mathrm{F}} \le L_{\nabla g} \|\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2\|_{\mathrm{F}}, \forall \boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2 \in \mathbb{R}^{p\times m}, \tag{2.9}$$

where $L_{\nabla g}$ is the Lipschitz constant of ∇g and h is a continuous convex (possibly non-smooth) function (Ji and Ye, 2009).

In view of (2.7), this corresponds to

$$g(\boldsymbol{\Gamma}) \overset{\text{def}}{=} (mn)^{-1} \sum_{i=1}^{n}\sum_{j=1}^{m} \rho_\tau(Y_{ij} - \boldsymbol{X}_i^\top \boldsymbol{\Gamma}_{\cdot j}), \tag{2.10}$$

$$h(\boldsymbol{\Gamma}) \overset{\text{def}}{=} \lambda \|\boldsymbol{\Gamma}\|_*. \tag{2.11}$$

The Lipschitz constant of ∇g is $L_{\nabla g} = 2(mn)^{-1} \max(\tau, 1-\tau)\|\mathbf{X}\|_{\mathrm{F}}^2$; see Appendix A.1.

Algorithm 1 is an application of FISTA, with g and h chosen as (2.10) and (2.11).

The subroutine $\mathtt{SVT}_{\lambda,g}$ in Algorithm 1 is the singular value thresholding operator given by $\mathtt{SVT}_{\lambda,g}(\mathbf{S}) \overset{\text{def}}{=} \mathbf{U}_\mathbf{S}(\mathbf{D}_\mathbf{S} - (\lambda/L_{\nabla g})\mathbf{I}_{p\times m})_+ \mathbf{V}_\mathbf{S}^\top$, where SVD implies $\mathbf{S} = \mathbf{U}_\mathbf{S}\mathbf{D}_\mathbf{S}\mathbf{V}_\mathbf{S}^\top$, $\mathbf{I}_{p\times m}$ is a rectangular identity matrix with main diagonal elements equal to 1, and $(\mathbf{S})_+ = (\max\{0, s_{ij}\})$.

Algorithm 1: FISTA for expectile regression with nuclear norm penalty.

Input: $\{\boldsymbol{Y}_i\}_{i=1}^n$, $\{\boldsymbol{X}_i\}_{i=1}^n$, λ
Output: $\widehat{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}_T$
Initialization: $\boldsymbol{\Gamma}_0 = 0$, $\boldsymbol{\Omega}_1 = 0$, step size $\delta_1 = 1$;
for $t = 1, 2, \ldots, T$ **do**
  $\boldsymbol{\Gamma}_t = \mathrm{SVT}_{\lambda, g}\big(\boldsymbol{\Omega}_t - L_{\nabla g}^{-1} \nabla g(\boldsymbol{\Omega}_t)\big)$;
  $\delta_{t+1} = \frac{1+\sqrt{1+4\delta_t^2}}{2}$;
  $\boldsymbol{\Omega}_{t+1} = \boldsymbol{\Gamma}_t + \frac{\delta_t - 1}{\delta_{t+1}}(\boldsymbol{\Gamma}_t - \boldsymbol{\Gamma}_{t-1})$;
end

Remark 2.2 (*Initialization and the Stopping Rule*)**.** We suggest to initialize the algorithm with $\boldsymbol{\Gamma}_0 = 0$ in Algorithm 1 , but because the optimization problem is convex, this can be replaced by any matrix. Of course, the algorithm converges faster if we initialize it with a matrix that is close to the minimizer. We suggest to stop the algorithm at iteration T satisfying $|F(\boldsymbol{\Gamma}_{T+1}) - F(\boldsymbol{\Gamma}_T)| \le \epsilon$, for some small $\epsilon > 0$. In the simulation and empirical analysis of this paper, $\epsilon = 10^{-6}$.

The convergence of Algorithm 1 in terms of the loss function is guaranteed by the following theorem.

Theorem 2.1 (*Bounds for the Loss Difference and Convergence Rate in Algorithm 1*)**.** *Let* $\{\boldsymbol{\Gamma}_t\}_{t=0}^T$ *be the sequence obtained by the iteration of Algorithm 1 . Then*

$$|F(\boldsymbol{\Gamma}_t) - F(\widehat{\boldsymbol{\Gamma}})| \le \frac{4(mn)^{-1} \max(\tau, 1-\tau) \|\mathbf{X}\|_{\mathrm{F}}^2 \|\boldsymbol{\Gamma}_0 - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}^2}{(t+1)^2}. \tag{2.12}$$

In particular, if for $\epsilon > 0$,

$$t \ge \frac{2\sqrt{\max(\tau, 1-\tau)} \|\mathbf{X}\|_{\mathrm{F}} \|\boldsymbol{\Gamma}_0 - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}}{\sqrt{mn\epsilon}} - 1, \tag{2.13}$$

then $|F(\boldsymbol{\Gamma}_t) - F(\widehat{\boldsymbol{\Gamma}})| \le \epsilon$.

The bound (2.12) comes from a careful calculation of the Lipschitz constant of the gradient of g. The proof of Theorem 2.1 can be found in Appendix A.1.

Theorem 2.1 shows that to get an ϵ-accurate solution, it requires $1/\sqrt{\epsilon}$ steps when holding other parameters fixed. This is smaller than $1/\epsilon$ steps given by quantile regression and $1/\epsilon^2$ by the general subgradient methods, see Theorem 2.3 and Remark 2.4 in Chao et al. (2016). In view of (2.13), when τ is approaching 0 or 1, the number of iterations that is required to achieve an ϵ-accurate solution would increase.

Furthermore, utilizing the strong convexity of g, we can obtain a bound for $\|\boldsymbol{\Gamma}_t - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}^2$. For this purpose, additional assumptions on the design $\boldsymbol{X}$ are required.

(A1) Suppose $\mathrm{E}\boldsymbol{X}_i = 0$, $\mathrm{E}\boldsymbol{X}_i\boldsymbol{X}_i^\top = \boldsymbol{\Sigma}$ with $\sigma_{\min}(\boldsymbol{\Sigma}) > C_1$ and $\sigma_{\max}(\boldsymbol{\Sigma}) < C_2$ for some constants $C_1, C_2 > 0$ uniformly in p. For some sequence $0 < a_n < 1$, constants $c_1, c_2 > 0$,

$$\mathrm{P}\left[\sigma_{\min}\left(\frac{\mathbf{X}^\top\mathbf{X}}{n}\right) \ge c_1\sigma_{\min}(\boldsymbol{\Sigma}), \sigma_{\max}\left(\frac{\mathbf{X}^\top\mathbf{X}}{n}\right) \le c_2\sigma_{\max}(\boldsymbol{\Sigma})\right] \ge 1 - a_n. \tag{2.14}$$

Assumption (A1) holds for Gaussian design $\boldsymbol{X}$ with $c_1 = 1/9$, $c_2 = 9$ and $a_n = 4\exp(-n/2)$. See Wainwright (2009). It can be shown that (A1) holds for the sub-gaussian designs; see Vershynin (2012a) for details.

The following theorem characterizes the convergence in the Frobenius norm.

Theorem 2.2. *Given (A1), the sequence* $\boldsymbol{\Gamma}_t$ *obtained from Algorithm 1 satisfy*

$$\|\boldsymbol{\Gamma}_t - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}^2 \le \frac{36}{n(t+1)^2} \frac{\max(\tau, 1-\tau)}{\min(\tau, 1-\tau)} \frac{\|\mathbf{X}\|_{\mathrm{F}}^2}{\sigma_{\min}(\boldsymbol{\Sigma})} \|\boldsymbol{\Gamma}_0 - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}^2, \tag{2.15}$$

with probability greater than $1 - a_n$. *In particular, if for* $\epsilon > 0$,

$$t \ge 6\sqrt{\frac{\max(\tau, 1-\tau)}{\min(\tau, 1-\tau)}} \frac{\|\mathbf{X}\|_{\mathrm{F}} \|\boldsymbol{\Gamma}_0 - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}}{\sqrt{n\sigma_{\min}(\boldsymbol{\Sigma})\epsilon}} - 1, \tag{2.16}$$

then $\|\boldsymbol{\Gamma}_t - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}^2 \le \epsilon$ *holds with probability greater than* $1 - a_n$.

The proof of Theorem 2.2 is in Appendix A.2. We discuss the estimation of the number of factors in the following remark.

Remark 2.3 (*Estimation of the Number of Factors*). The number of factors r defined in Section 2.1 can be estimated by $\text{rank}(\mathbf{\Gamma}_T)$, which is the estimator generated by Algorithm 1 . If the number of factors is exactly sparse, $\text{rank}(\mathbf{\Gamma}_T)$ is usually a good estimator; see the simulation study in Section 3.

2.3. Oracle inequalities

In this section, we derive the bounds for the difference between the sequence $\mathbf{\Gamma}_t$ generated by Algorithm 1 and the true matrix $\mathbf{\Gamma}$. These results heavily rely on the strong convexity of ρ_τ.

We make the following assumptions.

(A2) There exists $C > 0$ such that for $u_{ij} \overset{\text{def}}{=} Y_{ij} - \boldsymbol{X}_i^\top \mathbf{\Gamma}_{\cdot j}$, $\mathrm{P}(|u_{ij}| > s) \le \exp(1 - s^2/C^2)$, $\forall s \ge 0$) with sub-gaussian norm $\|u_{ij}\|_{\psi_2} \overset{\text{def}}{=} \sup_{p\ge 1} p^{-1/2}(\mathsf{E}|u_{ij}|^p)^{1/p}$, and let $K_u \overset{\text{def}}{=} \max_{1\le j\le m} \|u_{ij}\|_{\psi_2}$.

(A3) Conditional on $\boldsymbol{X}_i$, Y_{ij} are independent over j.

(A2) regulates the tails of Y_{ij}. (A3) is required for obtaining the bounds on the tail probabilities of the estimation error. In Theorem 2.3, we state a non-asymptotic bound for $\|\mathbf{\Gamma}_t - \mathbf{\Gamma}\|_{\mathrm{F}}$ in the general situation that the number of factors can be increasing with n.

Theorem 2.3 (*Approximately Sparse Factors*). *Under (A1)– (A3), $\lambda = 2\ \mathrm{cm}^{-1} \max(\tau, 1-\tau) K_u \sqrt{\|\mathbf{\Sigma}\|}\sqrt{\frac{p+m}{n}}$ for some absolute constant $c > 0$. Then for any $q \in \{1, \dots, p \wedge m\}$, the sequence $\mathbf{\Gamma}_t$ obtained by Algorithm 1 satisfy*

$$\|\mathbf{\Gamma}_t - \mathbf{\Gamma}\|_{\mathrm{F}}^2 \le c''\Big(\frac{R_t}{n} + 1\Big)\sqrt{\frac{p+m}{n}}\zeta_\tau \left\{\sqrt{\frac{p+m}{n}}\zeta_\tau q + \sum_{j=q+1}^{p\wedge m} \sigma_j(\mathbf{\Gamma})\right\} + \frac{c'' R_t}{n}\|\mathbf{\Gamma}_0 - \mathbf{\Gamma}\|_{\mathrm{F}}^2, \tag{2.17}$$

with probability greater than $1 - 3\cdot 8^{-(p+m)} - a_n$, where $c'' > 0$ is an absolute constant, $R_t \overset{\text{def}}{=} \frac{1}{(t+1)^2}\frac{\max(\tau,1-\tau)}{\min(\tau,1-\tau)}\frac{\|\mathbf{X}\|_{\mathrm{F}}^2}{\sigma_{\min}(\mathbf{\Sigma})}$ and $\zeta_\tau \overset{\text{def}}{=} \frac{\max(\tau,1-\tau)}{\min(\tau,1-\tau)}\frac{\sqrt{\|\mathbf{\Sigma}\|}}{\sigma_{\min}(\mathbf{\Sigma})}K_u$.

Please see Appendix B for a proof of Theorem 2.3. Note that (2.17) holds for any $q \in \{1, \dots, p \wedge m\}$. The optimal bound is obtained by selecting q that balances $\sqrt{\frac{p+m}{n}}\zeta_\tau q$ and $\sum_{j=q+1}^{p\wedge m}\sigma_j(\mathbf{\Gamma})$. For a fixed number of iterations t in Algorithm 1 and τ, a sufficient condition for (2.17) tending to zero is that the number of factors r is *approximately sparse* ($\mathbf{\Gamma}$ is *approximately low rank*): there exists an increasing sequence $q = q_n \in \mathbb{N}$ such that

$$\lim_{n\to\infty}\frac{p+m}{n}\zeta_\tau^2 q = 0 \quad \text{and} \quad \lim_{n\to\infty}\left\{\sum_{j=q+1}^{p\wedge m}\sqrt{\frac{p+m}{n}}\zeta_\tau\sigma_j(\mathbf{\Gamma})\right\} = 0, \tag{2.18}$$

where p and m can be growing sequences in n. The quantity R_t characterizes how the computational cost influences the error bound. We can increase the number of iterations in Algorithm 1 to shrink R_t, but this also increases the computational cost. Similar to Theorems 2.1 and B.1, when τ is approaching to the boundaries of (0, 1), the bound in (2.17) will increase. Furthermore, heavier tails for Y_{ij} make higher K_u, and lead to higher error bounds.

If the number of factors is fixed and is not increasing with n ($\text{rank}(\mathbf{\Gamma})$ is fixed), then (2.17) is minimized by selecting $q = \text{rank}(\mathbf{\Gamma})$ and $\sum_{j=q+1}^{p\wedge m}\sqrt{\frac{p+m}{n}}\zeta_\tau\sigma_j(\mathbf{\Gamma}) = 0$ in (2.17). Hence, we have the following corollary.

Corollary 2.1 (*Exactly Sparse Factors*). *Under the conditions of Theorem 2.3,*

$$\|\mathbf{\Gamma}_t - \mathbf{\Gamma}\|_{\mathrm{F}}^2 \le c''\Big(\frac{R_t}{n} + 1\Big)\frac{p+m}{n}\zeta_\tau^2\,\text{rank}(\mathbf{\Gamma}) + \frac{c'' R_t}{n}\|\mathbf{\Gamma}_0 - \mathbf{\Gamma}\|_{\mathrm{F}}^2, \tag{2.19}$$

with probability greater than $1 - 3\cdot 8^{-(p+m)} - a_n$, where $c'' > 0$ is an absolute constant, $R_t = \frac{1}{(t+1)^2}\frac{\max(\tau,1-\tau)}{\min(\tau,1-\tau)}\frac{\|\mathbf{X}\|_{\mathrm{F}}^2}{\sigma_{\min}(\mathbf{\Sigma})}$ and $\zeta_\tau = \frac{\max(\tau,1-\tau)}{\min(\tau,1-\tau)}\frac{\sqrt{\|\mathbf{\Sigma}\|}}{\sigma_{\min}(\mathbf{\Sigma})}K_u$.

Remark 2.4. As explained in Section 2.1, we estimate $\mathbf{V}_{\cdot k,t}$ (the loadings corresponding to the kth factor for all responses) in the SVD $\mathbf{\Gamma}_t = \mathbf{U}_t\mathbf{D}_t\mathbf{V}_t^\top$. By Theorem 3.10 of Chao et al. (2016), we have:

$$1 - |\mathbf{V}_{\cdot k}^\top \mathbf{V}_{\cdot k,t}| \le \frac{2(2\|\mathbf{\Gamma}\| + \|\mathbf{\Gamma}_t - \mathbf{\Gamma}\|_{\mathrm{F}})\|\mathbf{\Gamma}_t - \mathbf{\Gamma}\|_{\mathrm{F}}}{\min\{\sigma_{k-1}^2(\mathbf{\Gamma}) - \sigma_k^2(\mathbf{\Gamma}), \sigma_k^2(\mathbf{\Gamma}) - \sigma_{k+1}^2(\mathbf{\Gamma})\}}, \tag{2.20}$$

where $\mathbf{V}_{\cdot k}$ are the true loadings. Theorem 2.3 (or Corollary 2.1) can be used with (2.20) to get an explicit bound.

Table 3.1
The averaged estimated number of factors $\widehat{r}$ over simulation repetitions with respect to τ and c. Values in the parentheses are the standard deviations over the simulation repetitions.

τ	0.05	0.3	0.5	0.7	0.95
	$r = 10$				
$c = 1.3$	10.95 (0.22)	11.00 (0.00)	10.00 (0.00)	11.00 (0.00)	10.94 (0.23)
$c = 1.5$	10.70 (0.47)	11.00 (0.00)	10.00 (0.00)	11.00 (0.00)	10.71 (0.46)
$c = 1.7$	10.19 (0.61)	11.00 (0.00)	10.00 (0.00)	11.00 (0.00)	10.20 (0.60)
	$r = 5$				
$c = 1.3$	6.00 (0.00)	6.00 (0.00)	5.00 (0.04)	6.00 (0.00)	6.00 (0.00)
$c = 1.5$	6.00 (0.00)	6.00 (0.00)	5.00 (0.00)	6.00 (0.00)	6.00 (0.00)
$c = 1.7$	6.00 (0.00)	6.00 (0.06)	5.00 (0.00)	6.00 (0.04)	6.00 (0.00)
	$r = 2$				
$c = 1.3$	3.00 (0.00)	3.00 (0.00)	2.03 (0.18)	3.00 (0.00)	3.00 (0.00)
$c = 1.5$	3.00 (0.00)	2.99 (0.12)	2.00 (0.00)	2.99 (0.09)	3.00 (0.00)
$c = 1.7$	3.00 (0.00)	2.72 (0.45)	2.00 (0.00)	2.78 (0.41)	3.00 (0.00)

3. Simulation study

In this section, we apply our method on the simulated data to evaluate the estimation performance on the factors and loadings, as the number of factors varies.

Set $n = m = p = 100$. For $i = 1, \ldots, n$, $j = 1, \ldots, m$, let $\boldsymbol{X}_i \sim \mathcal{N}(0, \Sigma_{p\times p})$ with $\Sigma_{jk} = 0.5^{|j-k|}$ and $\boldsymbol{\varepsilon}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_{m\times m})$, the response variables are generated by

$$Y_{ij} = \boldsymbol{X}_i^\top \boldsymbol{\Gamma}_{\cdot j} + \varepsilon_{ij} = \sum_{k=1}^{r} \psi_{jk} f_k(\boldsymbol{X}_i) + \varepsilon_{ij} = \sum_{k=1}^{r} \mathbf{V}_{jk} \mathbf{D}_{kk} \boldsymbol{X}_i^\top \mathbf{U}_{\cdot k} + \varepsilon_{ij}, \tag{3.1}$$

where $r = \operatorname{rank}(\boldsymbol{\Gamma})$. We will set $r = 2, 5, 10$, and the nonzero diagonal components of $\mathbf{D}$ are (19.01, 18.74, 18.65, 18.22, 17.80, 17.50, 17.21, 17.02, 16.57, 16.49). The columns of $\mathbf{V}$ and $\mathbf{U}$ are the orthonormal singular vectors of a matrix with components chosen from $\mathcal{N}(0, 1)$. We repeat the data generation 500 times.

We apply Algorithm 1 with $\boldsymbol{Y}$ and $\widetilde{\boldsymbol{X}} = (\mathbf{l}_n, \widetilde{\boldsymbol{X}})$, where $\mathbf{l}_n = (1, \ldots, 1)$ is the intercept. The tuning parameter λ is selected according to Lemma B.1, i.e., $\lambda = 2\ \mathrm{cm}^{-1} \max(\tau, 1-\tau) K_u \sqrt{\|\boldsymbol{\Sigma}\|} \sqrt{\frac{p+m}{n}}$. We stop the algorithm as described in Remark 2.2. Denote the resulting estimator $\widetilde{\boldsymbol{\Gamma}}_1$, and obtain $\widetilde{\boldsymbol{\Gamma}}$ by removing the first row (the intercept) of $\widetilde{\boldsymbol{\Gamma}}_1$.

Table 3.1 reports the results for the estimated number of factors $\widehat{r}$, which is the number of nonzero singular values of $\widetilde{\boldsymbol{\Gamma}}$ that are greater than 10^{-10}. That is, the singular values smaller than 10^{-10} are treated as zero. We try several values of c in the formula for λ because we do not know its exact value. The true number of factors are generally well recovered by our algorithm, except for the expectiles that deviate more from $\tau = 0.5$. Furthermore, the estimated number of factors is robust to the model randomness as the standard deviations are very small. The results are similar for different values of c, so we fix $c = 1.3$ for all the later analysis.

The Frobenius error $\|\widetilde{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}\|_{\mathrm{F}}$ is shown in Fig. 3.1. The results are symmetric in τ around $\tau = 0.5$, and the estimation errors tend to be larger for the tail τ. In the models where r is larger, the Frobenius error is also larger. Our findings in the simulation studies are consistent with the roles of τ and r in the error bound in Corollary 2.1.

We measure the estimation performance of the factors and loadings by

$$\begin{aligned} &\|\boldsymbol{\Delta}^{\mathrm{fac}}_{k\cdot}\|_2 / \mathbf{D}_{kk}, \text{ where } \boldsymbol{\Delta}^{\mathrm{fac}} \overset{\text{def}}{=} |\widetilde{\mathbf{D}}\widetilde{\mathbf{U}}^\top| - |\mathbf{D}\mathbf{U}^\top|, \\ &\|\boldsymbol{\Delta}^{\mathrm{load}}_{\cdot k}\|_2, \text{ where } \boldsymbol{\Delta}^{\mathrm{load}} \overset{\text{def}}{=} |\widetilde{\mathbf{V}}| - |\mathbf{V}|, \\ \text{and} \quad &1 - |\mathbf{V}_{\cdot k}^\top \widetilde{\mathbf{V}}_{\cdot k}|, \end{aligned} \tag{3.2}$$

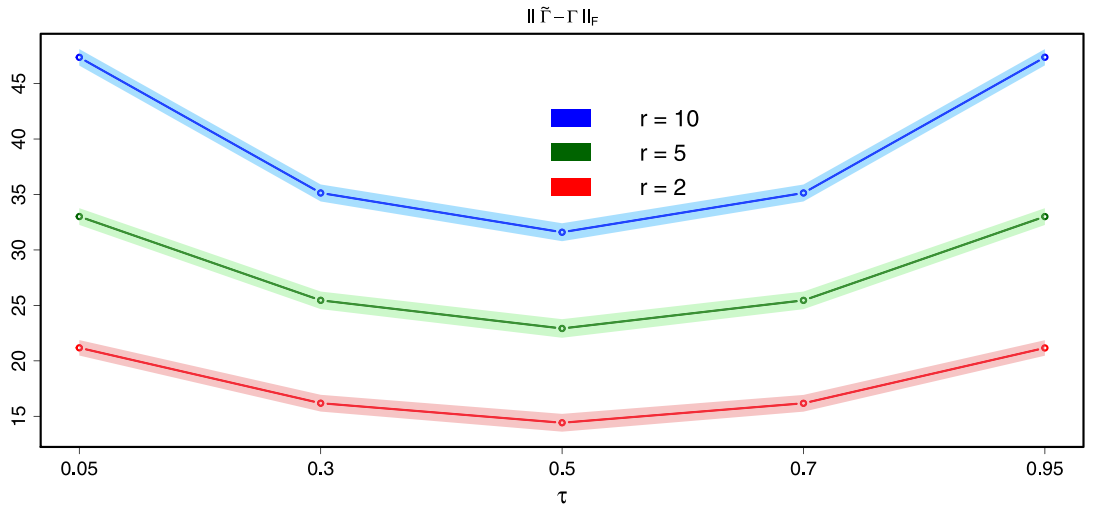


Fig. 3.1. The averaged estimation error $\|\widetilde{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_F$ ($c = 1.3$ in λ). The solid lines represent the averaged Frobenius errors over simulation repetitions, and the bands describe the standard deviations over the simulation repetitions.

for $k = 1, \ldots, r$, where $\widetilde{\mathbf{V}}$, $\widetilde{\mathbf{D}}$ and $\widetilde{\mathbf{U}}$ are based on the SVD $\widetilde{\mathbf{\Gamma}} = \widetilde{\mathbf{U}}\widetilde{\mathbf{D}}\widetilde{\mathbf{V}}^\top$, and the absolute value is taken componentwisely to the matrix. We do not include the covariate $\boldsymbol{X}_i$ in the measure for the estimation error of the factors because all factors share the same $\boldsymbol{X}_i$. We choose two measures for the estimation performance of the loadings. The $\|\mathbf{\Delta}^{\text{load}}_{\cdot k}\|_2$ measures the performance on the recovery of the absolute values of the loadings, which will be relevant in the empirical analysis in Section 4. On the other hand, $1 - |\mathbf{V}_{\cdot k}^\top \widetilde{\mathbf{V}}_{\cdot k}|$ corresponds to the theory that we stated in (2.20), which can be regarded as another measure for the recovery performance. We have also performed the analysis for $\tau = 0.05$ and 0.3, but we do not include their results in the paper because they are similar to $\tau = 0.95$ and 0.7. The results are presented in Fig. 3.2. Some general patterns are observed for the three panels. Smaller r gives smaller estimation error, but the associated standard deviation is larger. When τ deviates from 0.5, the error is larger, and this effect is particularly for $\|\mathbf{\Delta}^{\text{fac}}_{k\cdot}\|_2/\mathbf{D}_{kk}$. $\|\mathbf{\Delta}^{\text{load}}_{\cdot k}\|_2$ shows similar pattern to $1 - |\mathbf{V}_{\cdot k}^\top \widetilde{\mathbf{V}}_{\cdot k}|$, but the variance for $1 - |\mathbf{V}_{\cdot k}^\top \widetilde{\mathbf{V}}_{\cdot k}|$ is overall larger.

4. Empirical analysis: predicting risk attitude with fMRI Data

In this section, we apply our method on the fMRI data to predict the risk attitude on the investment decisions making. To understand how human brain responds to reward and risk is an important research topic in neuropsychology, financial economics and neuroeconomics (Heekeren et al., 2008; Camerer, 2007; Schultz, 2015). Previous research mainly focuses on the identification of the region of interest (ROI) with Blood Oxygenation Level Dependent (BOLD) signals (see Schultz, 2015 and the references therein). However, only a few research uses fMRI on predicting the risk attitude of subjects. Helfinstein et al. (2014) train support vector machines with the BOLD signals recorded in a Ballon Analog Risk Task (BART) on several combinations of brain regions, and this classifier can predict subjects' next choice with over 70% accuracy; van Bömmel et al. (2014) and Majer et al. (2016) retrieve factor loadings from a dynamic factor model on BOLD and apply these loadings on predicting subjects' risk attitude.

We focus on predicting the risk attitude of the subjects using the BOLD signals, but we differ from the previous studies in that we separately analyze the *positive* and *negative* BOLD signals observed in the cortical regions. The positive BOLD signals are known to be closely associated with increased neuronal activities, but the interpretation of large *negative* BOLD responses (NBR) is still controversial. Mullinger et al. (2014) argue that the best explanation for NBR at the cortical layer might be a decrease in cerebral blood flow (CBF) with a lesser reduction in the neuronal activities, which is measured by the cerebral metabolic rate of oxygen consumption ($CMRO_2$). This explanation is proven to be an important complement of the more classical "blood/vascular stealing" hypothesis (see p. 263 of Mullinger et al., 2014). However, Mullinger et al. (2014) also argue that there may exist deeper neuronal reasons for NBR than simply the inversion of the neurovascular coupling mechanism of the positive BOLD responses. Following the interpretation of NBR of Mullinger et al. (2014), we suspect that NBR also contain valuable information for predicting the risk attitude. Using our expectile based approach, we study whether the positive and negative extreme BOLD responses are relevant to the risk attitude.

4.1. Data

Our data come from a rapid event-related design experiment on investment decisions making, and this data set is firstly analyzed in Majer et al. (2016). The experiment was done as follows: 19 subjects were requested to make choices in 256

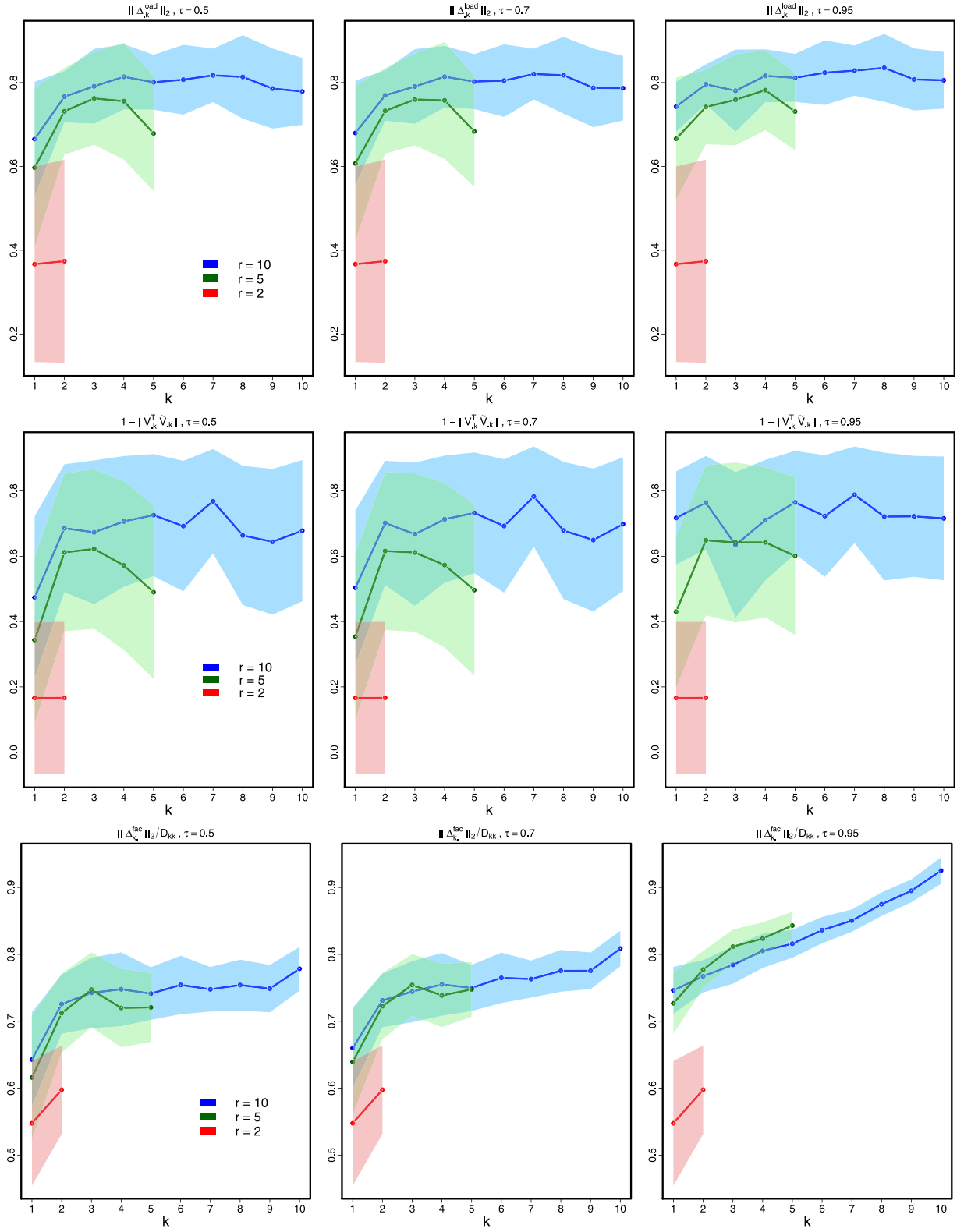


Fig. 3.2. The estimation errors $\|\boldsymbol{\Delta}^{\text{load}}_{.k}\|_2$, $1-|\mathbf{V}^{\top}_{.k}\widetilde{\mathbf{V}}_{.k}|$ for the loadings and $\|\boldsymbol{\Delta}^{\text{fac}}_{k.}\|_2/\mathbf{D}_{kk}$ for the factors, defined in (3.2). The solid lines represent the averaged errors, and the bands describe the standard deviations over simulation repetitions; $c = 1.3$ in λ.

Table 4.1
The goodness of fit R^2, Spearman's and Kendall's rank correlations for the in-sample fitting and out-of-sample prediction by (M1) or (M2) with/without constrains, under different τ, ω levels.

		Constrained model (only 1st factor)						Unconstrained model (2 factors)					
		Whole series (M1)			Task-wise (M2)			Whole series (M1)			Task-wise (M2)		
τ		0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
		In-sample fitting											
$\omega = 0.1$	R^2	0.084	0.158	0.101	0.412	0.412	**0.413**	0.312	0.263	0.226	**0.455**	0.454	0.454
	Spearman's rank corr	0.149	0.377	0.328	0.595	0.595	**0.604**	0.532	0.526	0.396	**0.618**	**0.618**	**0.618**
	Kendall's rank corr	0.076	0.263	0.228	0.462	0.462	**0.474**	0.333	0.357	0.275	**0.474**	**0.474**	**0.474**
$\omega = 0.5$	R^2	0.070	0.043	0.030	0.134	0.136	0.135	0.307	0.260	0.352	**0.445**	0.440	0.441
	Spearman's rank corr	0.177	0.140	0.226	0.335	0.316	0.326	0.547	0.528	0.596	0.533	0.544	0.544
	Kendall's rank corr	0.135	0.088	0.135	0.205	0.193	0.205	0.427	0.333	**0.415**	0.368	0.380	0.380
$\omega = 0.9$	R^2	0.199	0.238	0.148	0.206	0.205	0.205	0.393	0.367	0.229	**0.487**	**0.496**	**0.500**
	Spearman's rank corr	0.435	0.540	0.181	0.412	0.412	0.412	0.588	**0.628**	0.582	0.596	**0.637**	**0.637**
	Kendall's rank corr	0.333	0.391	0.135	0.298	0.298	0.298	**0.439**	**0.439**	**0.439**	0.462	**0.497**	**0.497**
		Out-of-sample predicting											
$\omega = 0.1$	Spearman's rank corr	−0.453	−0.181	−0.321	**0.454**	0.451	0.440	−0.079	−0.133	0.072	0.298	0.298	0.298
	Kendall's rank corr	−0.322	−0.111	−0.240	**0.357**	0.345	0.345	−0.076	−0.088	0.041	0.216	0.216	0.216
$\omega = 0.5$	Spearman's rank corr	−0.444	−0.700	−0.658	−0.119	−0.119	−0.119	−0.035	−0.196	0.247	0.205	0.204	0.212
	Kendall's rank corr	−0.275	−0.509	−0.450	−0.064	−0.064	−0.064	−0.006	−0.146	0.135	0.123	0.111	0.123
$\omega = 0.9$	Spearman's rank corr	−0.207	0.204	−0.493	0.023	0.023	0.023	0.161	0.072	−0.447	0.293	0.307	0.307
	Kendall's rank corr	−0.170	0.135	−0.345	0.006	0.006	0.006	0.076	0.041	−0.298	0.205	0.216	0.216

investment decision tasks and each task lasts 7 s. The fMRI was taken every two seconds (temporal resolution = 2 s), and this resulted in 1400 images for each subject. We have also acquired the answer for each task from each subject. Before applying our method, it is necessary to identify the region of interest (ROI), because the BOLD responses in non-ROIs are generated by noise (under the generalized linear model; see Section 6.2.1 of Lindquist (2008)) and do not have a sparse factor structure. For our data, Majer et al. (2016) identify three brain regions Anterior insula (left and right aINS) and dorsomedial prefrontal cortex (DMPFC) as the active regions related to investment decisions via spectral clustering method. We will only focus on the BOLD responses of the voxels in these three regions.

We integrate the information of each region (left and right aINS and DMPFC) spatially by taking the *quantiles* of the BOLD responses over all voxels in these regions. At each fMRI scan i of the sth subject, we take the quantiles with levels $\omega \in \{0.1, 0.5, 0.9\}$ of the BOLD responses over all voxels in the regions $b = 1$ (aINS_L), $b = 2$ (aINS_R) and $b = 3$ (DMPFC) to construct a single time series $v_i(s, b, \omega)$, where $i = 1, \ldots, N = 1400$. Fig. 4.1 gives an illustration of the BOLD time series of each cluster. For each cluster, the series of 19 subjects at ω are averaged (the solid lines) and the bands show the dispersion of the 19 time series. We observe that the series for $\omega = 0.9$ is positive, which summarizes the information of the positive BOLD responses, while the series for $\omega = 0.1$ is negative, which corresponds to the negative BOLD responses. The series for $\omega = 0.5$ is stationary and varying around the origin. From Fig. 4.1, we observe that the series with each different ω shows different volatility, and this may imply that the series with different ω contains different information. We will show in Table 4.1 that the series with $\omega = 0.1$ and 0.9 tend to contain more information than $\omega = 0.5$.

4.2. Method

4.2.1. Factor loadings at each region b and quantile level ω

For each ω and a single region b, we consider two approaches to construct the variable Y_{ij}:

- **(C1)** Whole time series: set $Y_{ij}^{b,\omega} = v_i(j, b, \omega)$, where $i = 1, \ldots, n$ with $n = N = 1400$, $j = 1, \ldots, 19$ (subject). Thus, we have $m = 19$ curves in each region b and at each quantile level ω.
- **(C2)** Analyzing each task separately (task-wise): we divide the whole time series in each region b and at each quantile level ω into subseries based on the beginning and the end of each task. Let $\mathcal{I}_q \subset \{1, \ldots, N\}$ be the set containing the indices of the images taken during the qth task. In our data, each $|\mathcal{I}_q| = 3$ or 4. We linearly interpolate the points $\{v_i(s, b, \omega)\}_{i \in \mathcal{I}_q}$ for each fixed s, b, and ω. Denote $\widetilde{v}_i(s, b, q, \omega)$ by the value on the interpolated curve at the ith point in n equally distant grid on the interval $(\min(\mathcal{I}_q), \max(\mathcal{I}_q))$, where $i = 1, \ldots, n = 50$. Let $Y_{ij}^{b,\omega} = \widetilde{v}_i(s, b, q, \omega)$ with $j = 256(s-1) + q$, where $s = 1, \ldots, 19$ (index for subject) and $q = 1, \ldots, 256$ (index for task) for each ω, b. Thus, there are $m = 19 \times 256 = 4864$ curves in each b and ω.

The variable $\boldsymbol{X}_i$ is a vector of basis functions that need to be flexible enough to capture the various shapes of the fMRI BOLD sequences. For this purpose, we use the cubic B-spline basis $\{B_k\}_{k=1}^{p}$ with equally spaced knots on [0, 1], and set $\boldsymbol{X}_i = \big(B_1(i/n), B_2(i/n), \ldots, B_p(i/n)\big)^\top$, where $i = 1, \ldots, n$. Note that $n = 1400$ in (**C1**) and $n = 50$ in (**C2**). B-splines are

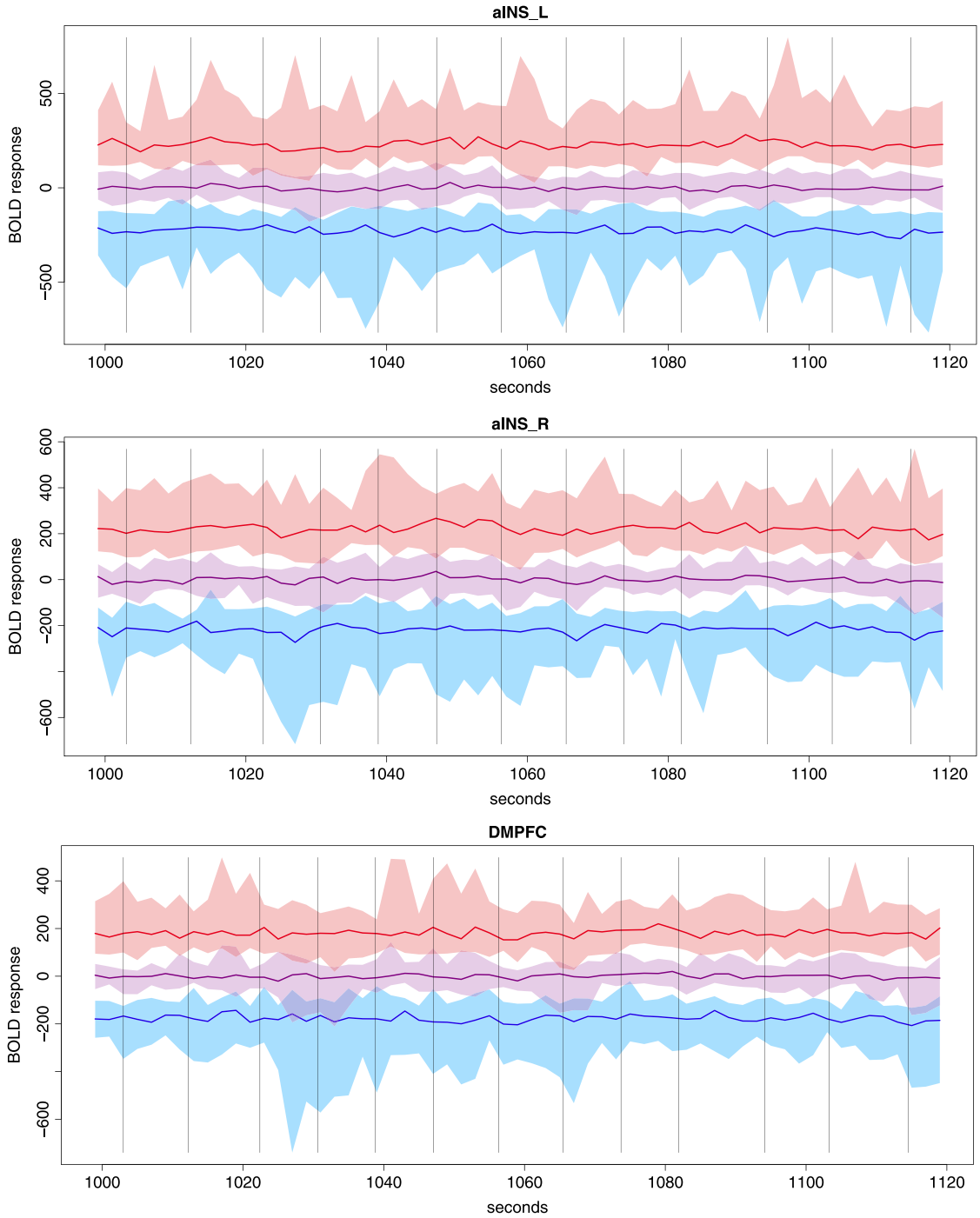


Fig. 4.1. In each region, the ω quantiles of the BOLD responses over all the voxels between 1000 and 1120 s of the experiment are shown. In each subfigure (region), lowest (resp., middle, highest) solid lines represent the median of $\omega = 0.1$ (resp., $\omega = 0.5, 0.9$) quantiles of all 19 subjects, and the upper and lower boundaries of the bands present the maximum and the minimum of the ω quantiles of the 19 subjects. Vertical lines indicate the occurrences of the stimuli (the beginning of each task). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

suitable for estimating the hemodynamic response function, see Degras and Lindquist (2014) for more details. We select $p = \lceil n^{0.8} \rceil$ of basis functions in each approach above, where $\lceil \cdot \rceil$ takes the smallest integer that is greater than the argument. The power 0.8 is greater than the (asymptotic) optimal rate 0.4, because the nuclear norm penalty alleviates the issue of overfitting. As the result, there are 329 basis functions in the approach (**C1**) and 23 in (**C2**).

We compute the matrix $\widehat{\mathbf{\Gamma}}^{b,\omega}$ with expectile level $\tau = 0.1, 0.5, 0.9$ using Y_{ij} and $\boldsymbol{X}_i$ by Algorithm 1 , where Y_{ij} is chosen under either (**C1**) or (**C2**) with $\lambda^{b,\omega}$ selected by the standard 5-fold cross-validation for each region b and each quantile level ω. Please see Appendix D.1 for the exact value of λ for each pair (b, ω). Using SVD $\widehat{\mathbf{\Gamma}}^{b,\omega} = \widehat{\mathbf{U}}_\tau^{b,\omega}\widehat{\mathbf{D}}_\tau^{b,\omega}(\widehat{\mathbf{V}}_\tau^{b,\omega})^\top$, where $(\widehat{\mathbf{V}}_\tau^{b,\omega})^\top$ is regarded as the factor loadings. We note that the size of the matrix $\widehat{\mathbf{V}}_\tau^{b,\omega}$ is 19×19 if we define $Y_{ij}^{b,\omega}$ by following (**C1**), and 4864×4864 by following (**C2**). Note that the sign of the factor loadings cannot be determined exactly (see Remark 2.1).

Remark 4.1 (*On the Computation of* λ)**.** The model error of the BOLD signals typically demonstrates autocorrelation following AR(k) or ARMA(1,1) (Lindquist, 2008 page 446) under the temporal resolution 2 s. A major consequence of the presence of temporal correlation is that the usual cross-validation could potentially underestimate λ, which leads to undersmoothing and overfitting (Opsomer et al., 2001 Section 2). This problem is especially important for the setting (**C2**), where the dimensionality is high because we separate each task. However, we observe that the estimated number of factors for the setting (**C2**) is typically very sparse (less than five factors). Overall, the overfitting does not cause a big issue and the usual cross-validation works well in our model.

4.2.2. Predicting risk attitude

To evaluate the prediction performance, we need to obtain the subjects' risk attitude β_s, where $s = 1, \ldots, 19$ denotes the subject. We follow the approach of Majer et al. (2016) and estimate β_s using the investment decisions made by the subjects to each task with logistic regression; see Appendix D.2 for more details. In essence, higher β_s means the subject s is *less* risk-averse.

In order to use the loadings $\widehat{\mathbf{V}}_\tau^{b,\omega}$ to predict β_s, we apply the standard linear regression models. In particular, in the case (**C1**), we construct a model for β_s using the first two factor loadings

$$\begin{aligned}\beta_s = \alpha_0^{\omega,\tau} &+ \alpha_{11}^{\omega,\tau}\left|(\widehat{\mathbf{V}}_\tau^{1,\omega})_{s1}\right| + \alpha_{12}^{\omega,\tau}\left|(\widehat{\mathbf{V}}_\tau^{2,\omega})_{s1}\right| + \alpha_{13}^{\omega,\tau}\left|(\widehat{\mathbf{V}}_\tau^{3,\omega})_{s1}\right| \\ &+ \alpha_{21}^{\omega,\tau}\left|(\widehat{\mathbf{V}}_\tau^{1,\omega})_{s2}\right| + \alpha_{22}^{\omega,\tau}\left|(\widehat{\mathbf{V}}_\tau^{2,\omega})_{s2}\right| + \alpha_{23}^{\omega,\tau}\left|(\widehat{\mathbf{V}}_\tau^{3,\omega})_{s2}\right| + \varepsilon_s, \quad s = 1, \ldots, 19, \end{aligned} \tag{M1}$$

where $\{\alpha_0^{\omega,\tau}, \alpha_{11}^{\omega,\tau}, \alpha_{12}^{\omega,\tau}, \alpha_{13}^{\omega,\tau}, \alpha_{21}^{\omega,\tau}, \alpha_{22}^{\omega,\tau}, \alpha_{23}^{\omega,\tau}\} \in \mathbb{R}^7$ are the intercept and the coefficients associated with the regions left and right Anterior insula, and dorsomedial prefrontal cortex.

In the case (**C2**), define the averaged loadings of all tasks for each s

$$\mu_{s,k}^{b,\omega,\tau} \overset{\text{def}}{=} \frac{1}{256}\sum_{q=1}^{256}\left|(\widehat{\mathbf{V}}_\tau^{b,\omega})_{256(s-1)+q,k}\right|.$$

We construct another model for β_s using $\mu_{s,k}^{b,\omega,\tau}$:

$$\begin{aligned}\beta_s = \bar{\alpha}_0^{\omega,\tau} &+ \bar{\alpha}_{11}^{\omega,\tau}\mu_{s,1}^{1,\omega,\tau} + \bar{\alpha}_{12}^{\omega,\tau}\mu_{s,1}^{2,\omega,\tau} + \bar{\alpha}_{13}^{\omega,\tau}\mu_{s,1}^{3,\omega,\tau} \\ &+ \bar{\alpha}_{21}^{\omega,\tau}\mu_{s,2}^{1,\omega,\tau} + \bar{\alpha}_{22}^{\omega,\tau}\mu_{s,2}^{2,\omega,\tau} + \bar{\alpha}_{23}^{\omega,\tau}\mu_{s,2}^{3,\omega,\tau} + \varepsilon_s, \quad s = 1, \ldots, 19, \end{aligned} \tag{M2}$$

where $\{\bar{\alpha}_0^{\omega,\tau}, \bar{\alpha}_{11}^{\omega,\tau}, \bar{\alpha}_{12}^{\omega,\tau}, \bar{\alpha}_{13}^{\omega,\tau}, \bar{\alpha}_{21}^{\omega,\tau}, \bar{\alpha}_{22}^{\omega,\tau}, \bar{\alpha}_{23}^{\omega,\tau}\} \in \mathbb{R}^7$. We take the absolute value of the loadings $\widehat{\mathbf{V}}_\tau^{b,\omega}$ because we are only interested in the magnitude of the loadings, which describes the importance of the factors.

Remark 4.2. If sufficiently many subjects are available, then ideally we could use all the estimated factors as suggested by one of our referees. However, because we have only 19 subjects, the number of factor loadings that can be included is very limited. For example, according to the results of an extensive simulation study shown in Table 1 on page 438 in Knofczynski and Mundfrom (2008), the maximum number of predictors that guarantees the best prediction performance is perhaps only around 9 to 12, given the sample size 19. In an unreported analysis, we checked the out-of-sample performance of the models that include up to 3 and 4 factors loadings. We are not able to find strong evidences that more factor loadings improve the prediction performance.

4.2.3. In-sample and out-of-sample performance

We compare the in-sample and out-of-sample performance of the models (M1) and (M2). For the in-sample performance, R^2 of both regressions (M1) and (M2) are computed. In addition, in order to determine whether (M1) and (M2) correctly predict the *order* of risk-aversion of the subjects (rather than the exact value of β_s), we calculate the Spearman's and Kendall's rank correlations between the fitted $\widehat{\beta}_s$ (in-sample) and β_s.

To measure the out-of-sample performance, we calculate $\{\widetilde{\beta}_s\}_{s=1}^{19}$ by a leave-one-out procedure. The steps are as below:

(1) Fix s, where $s = 1, \ldots, 19$. Use the values of the remaining 18 subjects to compute the coefficients $\{\alpha_0^{\omega,\tau}, \alpha_{11}^{\omega,\tau}, \alpha_{12}^{\omega,\tau}, \alpha_{13}^{\omega,\tau}, \alpha_{21}^{\omega,\tau}, \alpha_{22}^{\omega,\tau}, \alpha_{23}^{\omega,\tau}\}$ in model (M1) or $\{\bar{\alpha}_0^{\omega,\tau}, \bar{\alpha}_{11}^{\omega,\tau}, \bar{\alpha}_{12}^{\omega,\tau}, \bar{\alpha}_{13}^{\omega,\tau}, \bar{\alpha}_{21}^{\omega,\tau}, \bar{\alpha}_{22}^{\omega,\tau}, \bar{\alpha}_{23}^{\omega,\tau}\}$ in model (M2) by the standard linear regression.
(2) Compute $\widetilde{\beta}_s$ by plugging in the coefficients computed in the last step in models (M1) and (M2), and input the loadings of the sth subject.
(3) Repeat steps (1) and (2) for each $s = 1, \ldots, 19$.
(4) Calculate the Spearman's and Kendall's rank correlations between $\{\widetilde{\beta}_s\}_{s=1}^{19}$ and $\{\beta_s\}_{s=1}^{19}$.

4.3. Empirical results

In Table 4.1, we present the in-sample fitting and out-of-sample performance for models (M1) and (M2) with the constrained model that uses only the 1st factor ($\alpha_{21}^{\omega,\tau} = \alpha_{22}^{\omega,\tau} = \alpha_{23}^{\omega,\tau} = 0$ in (M1) and $\bar{\alpha}_{21}^{\omega,\tau} = \bar{\alpha}_{22}^{\omega,\tau} = \bar{\alpha}_{23}^{\omega,\tau} = 0$ in (M2)) and the unconstrained model, under various (τ, ω) pairs.

For the in-sample fitting, cases $\omega = 0.1$ and $\omega = 0.9$ outperform the case $\omega = 0.5$. This shows that both extreme negative or positive BOLD can lead to good fitting for models (M1) and (M2). In particular, the fitting performance is the best when $\tau = 0.9$ for $\omega = 0.9$ and $\tau = 0.1$ for $\omega = 0.1$, which correspond to the upper boundary of the red area and the lower boundary of the blue area in each of the three panels in Fig. 4.1.

For the out-of-sample performance, the constrained (M2) using only the first factor with the negative BOLD ($\omega = 0.1$, $\tau = 0.1$) nearly always outperforms all the other cases. In contrast, positive BOLD ($\omega = 0.9$) under the same model performs poorly. Moreover, the unconstrained model improves the prediction performance in most cases, particularly for (M2) under $\omega = 0.9$ and $\tau = 0.9$.

Majer et al. (2016) estimate a dynamic semiparametric factor model and extract the resulting factor loadings to predict the subjects' risk attitude. They evaluate the in-sample fitting (with all 19 subjects) by $R^2 = 0.47$ for a special case of our (M1) ($\tau = 0.5$ and $\alpha_{21}^{\omega,\tau} = \alpha_{22}^{\omega,\tau} = \alpha_{23}^{\omega,\tau} = 0$). Their fitting performance beats all the R^2 in our results, but we are able to describe the *predictive* abilities at several levels of τ, instead of only looking at $\tau = 0.5$. Our findings successfully confirm that the tails of the BOLD signals are more informative than their means in predicting the risk attitude.

5. Conclusions

In this paper, we propose a factorizable multivariate expectile regression method for the high-dimensional cross-sectional or spatial data with sparse latent factors. Fast iterative shrinkage-thresholding algorithm is applied to estimate the model. The convergence of the algorithm and the non-asymptotic theoretical guarantee of the estimator are established. We apply our method on the fMRI data obtained from an investment decisions making experiment, and study the ranking accuracy of the subjects' risk preference using the factor loadings of the extreme BOLD responses. The results show that the negative BOLD signals could provide comparable prediction performance as the positive BOLD signals. This provides insights into the on-going debate on the meaning of the negative BOLD responses.

There are several possibilities for the future research. As many data in practice are time series, there is a need to relax the i.i.d. assumption and make our method compatible with richer temporal structure. Statistical inference is also an important issue for many applications.

Acknowledgments

Financial support from the Deutsche Forschungsgemeinschaft via CRC 649 "Economic Risk" and IRTG 1792 "High Dimensional Non Stationary Time Series", Humboldt-Universität zu Berlin, is gratefully acknowledged. Shih-Kang Chao is partially supported by the Office of Naval Research of the U.S.A (ONR N00014-15-1-2331).

Appendix A. Proofs for Section 2.2

A.1. Proof for Theorem 2.1

Theorem 4.4 in Beck and Teboulle (2009) gives the upper bound of the loss difference at iteration t by

$$|F(\mathbf{\Gamma}_t) - F(\widehat{\mathbf{\Gamma}})| \leq \frac{2L_{\nabla g}\|\mathbf{\Gamma}_0 - \widehat{\mathbf{\Gamma}}\|_{\mathrm{F}}^2}{(t+1)^2}, \tag{A.1}$$

where $L_{\nabla g}$ is the Lipschitz constant of $\nabla g(\mathbf{\Gamma})$ defined in (2.9).

We note that

$$\rho_\tau'(u) = \begin{cases} 2\tau u & \text{for } u \geq 0; \\ 2(1-\tau)u & \text{for } u < 0. \end{cases} \tag{A.2}$$

Hence, the gradient is

$$\nabla g(\mathbf{\Gamma}) = -(mn)^{-1}\mathbf{X}^\top\left\{\mathbf{W} \circ (\mathbf{Y} - \mathbf{X}\mathbf{\Gamma})\right\}, \tag{A.3}$$

where $\mathbf{W}(\mathbf{\Gamma}) = (w_{ij}) \in \mathbb{R}^{n\times m}$, $w_{ij} \stackrel{\text{def}}{=} 2\left\{\tau + \mathbf{1}(Y_{ij} \leq \boldsymbol{X}_i^\top \mathbf{\Gamma}_{\cdot j})(1-2\tau)\right\}$, "$\circ$" represents the Hadamard product.

To simplify the notations, define $\mathbf{U}(\mathbf{\Gamma}) = (Y_{ij} - \boldsymbol{X}_i^\top \mathbf{\Gamma}_{\cdot j}) \in \mathbb{R}^{n\times m}$. For all $\mathbf{\Gamma}_1, \mathbf{\Gamma}_2 \in \mathbb{R}^{p\times m}$, let $\mathbf{U}_1 = \mathbf{U}(\mathbf{\Gamma}_1)$, $\mathbf{U}_2 = \mathbf{U}(\mathbf{\Gamma}_2)$, $\mathbf{W}_1 = \mathbf{W}(\mathbf{\Gamma}_1)$ and $\mathbf{W}_2 = \mathbf{W}(\mathbf{\Gamma}_2)$.

$$\begin{aligned} \|\nabla g(\mathbf{\Gamma}_1) - \nabla g(\mathbf{\Gamma}_2)\|_{\mathrm{F}} &= (mn)^{-1}\|\mathbf{X}^\top(\mathbf{W}_1 \circ \mathbf{U}_1) - \mathbf{X}^\top(\mathbf{W}_2 \circ \mathbf{U}_2)\|_{\mathrm{F}} \\ &\leq (mn)^{-1}\|\mathbf{X}\|_{\mathrm{F}}\|\mathbf{W}_1 \circ \mathbf{U}_1 - \mathbf{W}_2 \circ \mathbf{U}_2\|_{\mathrm{F}} \quad \text{(by submultiplicity)} \end{aligned}$$

$$
\begin{aligned}
&= (mn)^{-1}\|\mathbf{X}\|_{\mathrm{F}}\Big[\sum_{i=1}^{n}\sum_{j=1}^{m}\big\{\rho'_\tau(u_{1,ij}) - \rho'_\tau(u_{2,ij})\big\}^2\Big]^{1/2} \\
&\le (mn)^{-1}\|\mathbf{X}\|_{\mathrm{F}}\Big[\sum_{i=1}^{n}\sum_{j=1}^{m}\big\{2\max(\tau, 1-\tau)\big\}^2(u_{1,ij} - u_{2,ij})^2\Big]^{1/2} \\
&= 2(mn)^{-1}\max(\tau, 1-\tau)\|\mathbf{X}\|_{\mathrm{F}}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\Gamma}_1 - (\mathbf{Y} - \mathbf{X}\boldsymbol{\Gamma}_2)\|_{\mathrm{F}} \\
&\le 2(mn)^{-1}\max(\tau, 1-\tau)\|\mathbf{X}\|_{\mathrm{F}}^2\|\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2\|_{\mathrm{F}} \quad \text{(by submultiplicity)}, \qquad \text{(A.4)}
\end{aligned}
$$

where the fourth line makes use of the fact that $\rho'_\tau(u)$ is Lipschitz continuous with Lipschitz constant $2\max(\tau, 1-\tau)$, see Chao et al. (2017).

Plug $L_{\nabla g} = 2(mn)^{-1}\max(\tau, 1-\tau)\|\mathbf{X}\|_{\mathrm{F}}^2$ into (A.1) yields

$$
|F(\boldsymbol{\Gamma}_t) - F(\widehat{\boldsymbol{\Gamma}})| \le \frac{4(mn)^{-1}\max(\tau, 1-\tau)\|\mathbf{X}\|_{\mathrm{F}}^2\|\boldsymbol{\Gamma}_0 - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}^2}{(t+1)^2}. \qquad \text{(A.5)}
$$

Moreover, setting the right hand side of (A.5) to be ϵ ($\forall \epsilon > 0$) and solving for t gives

$$
t \ge \frac{2\sqrt{\max(\tau, 1-\tau)}\|\mathbf{X}\|_{\mathrm{F}}\|\boldsymbol{\Gamma}_0 - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}}{\sqrt{mn\epsilon}} - 1. \quad \square \qquad \text{(A.6)}
$$

A.2. Proof for Theorem 2.2

Following the proof of Theorem 1 in Fadili and Peyré (2011), define

$$
I(\boldsymbol{\Gamma}_t) \stackrel{\text{def}}{=} g(\boldsymbol{\Gamma}_t) - g(\widehat{\boldsymbol{\Gamma}}) - \langle\langle \nabla g(\boldsymbol{\Gamma}_t), \boldsymbol{\Gamma}_t - \widehat{\boldsymbol{\Gamma}}\rangle\rangle, \qquad \text{(A.7)}
$$

$$
J(\boldsymbol{\Gamma}_t) \stackrel{\text{def}}{=} h(\boldsymbol{\Gamma}_t) - h(\widehat{\boldsymbol{\Gamma}}) + \langle\langle \nabla g(\boldsymbol{\Gamma}_t), \boldsymbol{\Gamma}_t - \widehat{\boldsymbol{\Gamma}}\rangle\rangle, \qquad \text{(A.8)}
$$

the sum of them gives

$$
I(\boldsymbol{\Gamma}_t) + J(\boldsymbol{\Gamma}_t) = F(\boldsymbol{\Gamma}_t) - F(\widehat{\boldsymbol{\Gamma}}). \qquad \text{(A.9)}
$$

According to Lemma C.2, we have

$$
\begin{aligned}
I(\boldsymbol{\Gamma}_t) &\ge \kappa\|\boldsymbol{\Gamma}_t - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}^2 \\
&= \frac{1}{9}m^{-1}\min(\tau, 1-\tau)\sigma_{\min}(\boldsymbol{\Sigma})\|\boldsymbol{\Gamma}_t - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}^2, \qquad \text{(A.10)}
\end{aligned}
$$

where the second line holds with probability greater than $1 - a_n$ under (A1).

Since $\widehat{\boldsymbol{\Gamma}}$ is the optimizer of (2.6), therefore,

$$
\mathbf{0} \in \nabla g(\widehat{\boldsymbol{\Gamma}}) + \nabla h(\widehat{\boldsymbol{\Gamma}}), \qquad \text{(A.11)}
$$

which implies

$$
-\nabla g(\widehat{\boldsymbol{\Gamma}}) \in \nabla h(\widehat{\boldsymbol{\Gamma}}). \qquad \text{(A.12)}
$$

As a result, we have

$$
h(\boldsymbol{\Gamma}_t) - h(\widehat{\boldsymbol{\Gamma}}) \ge \langle\langle -\nabla g(\boldsymbol{\Gamma}_t), \boldsymbol{\Gamma}_t - \widehat{\boldsymbol{\Gamma}}\rangle\rangle, \qquad \text{(A.13)}
$$

i.e., $J(\boldsymbol{\Gamma}_t) \ge 0$.

Plugging (A.10) and (A.13) into (A.9) yields,

$$
\begin{aligned}
\|\boldsymbol{\Gamma}_t - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}^2 &\le \frac{9\,m}{\min(\tau, 1-\tau)\sigma_{\min}(\boldsymbol{\Sigma})}\big\{F(\boldsymbol{\Gamma}_t) - F(\widehat{\boldsymbol{\Gamma}})\big\} \\
&\le \frac{36}{n(t+1)^2}\frac{\max(\tau, 1-\tau)}{\min(\tau, 1-\tau)}\frac{\|\mathbf{X}\|_{\mathrm{F}}^2}{\sigma_{\min}(\boldsymbol{\Sigma})}\|\boldsymbol{\Gamma}_0 - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}^2, \qquad \text{(A.14)}
\end{aligned}
$$

with probability greater than $1 - a_n$. The second line comes from the result of Theorem 2.1. $\square$

Appendix B. Proof for Theorem 2.3

By triangle inequality, we have

$$
\|\boldsymbol{\Gamma}_t - \boldsymbol{\Gamma}\|_{\mathrm{F}}^2 = \|\boldsymbol{\Gamma}_t - \widehat{\boldsymbol{\Gamma}} + \widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}\|_{\mathrm{F}}^2 \le 2\|\boldsymbol{\Gamma}_t - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}^2 + 2\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}\|_{\mathrm{F}}^2. \qquad \text{(B.1)}
$$

Combining the results of Lemma B.2 and Theorem 2.2, it follows that

$$\begin{aligned}\|\mathbf{\Gamma}_t - \mathbf{\Gamma}\|_{\mathrm{F}}^2 \leq & 18^3 c^2 \frac{p+m}{n} \frac{\max(\tau, 1-\tau)^2}{\min(\tau, 1-\tau)^2} \frac{\|\mathbf{\Sigma}\|}{\sigma_{\min}(\mathbf{\Sigma})^2} K_u^2 \dim(\overline{\mathcal{M}}) \\ & + 144c \sqrt{\frac{p+m}{n}} \frac{\max(\tau, 1-\tau)}{\min(\tau, 1-\tau)} \frac{\sqrt{\|\mathbf{\Sigma}\|}}{\sigma_{\min}(\mathbf{\Sigma})} K_u \|\mathbf{\Gamma}_{\mathcal{M}^\perp}\|_* \\ & + \frac{72}{n(t+1)^2} \frac{\max(\tau, 1-\tau)}{\min(\tau, 1-\tau)} \frac{\|\mathbf{X}\|_{\mathrm{F}}^2}{\sigma_{\min}(\mathbf{\Sigma})} \|\mathbf{\Gamma}_0 - \widehat{\mathbf{\Gamma}}\|_{\mathrm{F}}^2, \end{aligned} \tag{B.2}$$

holds with probability greater than $1 - 3 \times 8^{-(p+m)} - a_n$.
Furthermore, given

$$\|\mathbf{\Gamma}_0 - \widehat{\mathbf{\Gamma}}\|_{\mathrm{F}}^2 = \|\mathbf{\Gamma}_0 - \mathbf{\Gamma} + \mathbf{\Gamma} - \widehat{\mathbf{\Gamma}}\|_{\mathrm{F}}^2 \leq 2\|\mathbf{\Gamma}_0 - \mathbf{\Gamma}\|_{\mathrm{F}}^2 + 2\|\mathbf{\Gamma} - \widehat{\mathbf{\Gamma}}\|_{\mathrm{F}}^2, \tag{B.3}$$

and applying Lemma B.2 again we complete the proof of Theorem 2.3. □

Now we show auxiliary results used in the proof of Theorem 2.3. The next theorem is an application of Theorem 1 of Negahban et al. (2012).

Theorem B.1 (*Error Bounds for the Estimator*). *Under (A1), for any $q \in \{1, \ldots, p \wedge m\}$, any optimal solution $\widehat{\mathbf{\Gamma}}$ in the problem (2.6) with $\lambda \geq 2\|\nabla g(\mathbf{\Gamma})\|$ satisfies the bound*

$$\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_{\mathrm{F}}^2 \leq \frac{9\, m^2 \lambda^2}{\left\{c_1 \min(\tau, 1-\tau)\sigma_{\min}(\mathbf{\Sigma})\right\}^2} q + \frac{36\, m\lambda}{\min(\tau, 1-\tau)\sigma_{\min}(\mathbf{\Sigma})} \sum_{j=q+1}^{p \wedge m} \sigma_j(\mathbf{\Gamma}), \tag{B.4}$$

with probability greater than $1 - a_n$, where $\sigma_j(\mathbf{\Gamma})$ is the jth singular value of $\mathbf{\Gamma}$.

Proof for Theorem B.1. The proof is an application of Theorem 1 of Negahban et al. (2012). First, we observe that the nuclear norm is *decomposable* in the sense that

$$\|\mathbf{\Gamma} + \mathbf{\Delta}\|_* = \|\mathbf{\Gamma}\|_* + \|\mathbf{\Delta}\|_*, \forall \mathbf{\Gamma} \in \mathcal{M}_q, \mathbf{\Delta} \in \overline{\mathcal{M}}_q^\perp, \tag{B.5}$$

where

$$\begin{aligned} \mathcal{M}_q &= \mathcal{M}(U_q, V_q) \stackrel{\text{def}}{=} \{\mathbf{\Theta} \in \mathbb{R}^{p \times m} | \operatorname{col}(\mathbf{\Theta}) \subseteq U_q, \operatorname{row}(\mathbf{\Theta}) \subseteq V_q\}, \\ \overline{\mathcal{M}}_q^\perp &= \overline{\mathcal{M}}^\perp(U_q, V_q) \stackrel{\text{def}}{=} \{\mathbf{\Theta} \in \mathbb{R}^{p \times m} | \operatorname{col}(\mathbf{\Theta}) \subseteq U_q^\perp, \operatorname{row}(\mathbf{\Theta}) \subseteq V_q^\perp\}, \end{aligned} \tag{B.6}$$

where $\operatorname{row}(\mathbf{\Theta})$ and $\operatorname{col}(\mathbf{\Theta})$ denote the row and column spaces of $\mathbf{\Theta}$. It can be seen that $\mathcal{M}_q \subset \overline{\mathcal{M}}_q$ where $\overline{\mathcal{M}}_q \stackrel{\text{def}}{=} \{\mathbf{\Theta} \in \mathbb{R}^{p \times m} | \operatorname{tr}(\mathbf{\Theta}^\top \mathbf{S}) = 0,\ \forall \mathbf{S} \in \overline{\mathcal{M}}_q^\perp\}$. Similarly, $\mathcal{M}_q^\perp \stackrel{\text{def}}{=} \{\mathbf{\Theta} \in \mathbb{R}^{p \times m} | \operatorname{tr}(\mathbf{\Theta}^\top \mathbf{S}) = 0,\ \forall \mathbf{S} \in \mathcal{M}_q\}$.

We will verify its conditions (G1) and (G2). For condition (G1), it is already mentioned above that the nuclear norm $\|\cdot\|_*$ is decomposable with respect to $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$ defined in (B.6). For condition (G2), note that on the event

$$\Omega_1 \stackrel{\text{def}}{=} \left\{\sigma_{\min}\left(\frac{\mathbf{X}^\top \mathbf{X}}{n}\right) \geq c_1 \sigma_{\min}(\mathbf{\Sigma}), \sigma_{\max}\left(\frac{\mathbf{X}^\top \mathbf{X}}{n}\right) \leq c_2 \sigma_{\max}(\mathbf{\Sigma})\right\}, \tag{B.7}$$

the loss function g is restrictive strongly convex with coefficients κ and $\xi = 0$ (we replace $\tau_{\mathcal{L}}$ in Negahban et al. (2012) by ξ) shown in Lemma C.2. Since we measure the error in the Frobenius norm $\|\cdot\|_{\mathrm{F}}$, the subspace compatibility constant (Definition 3 of Negahban et al., 2012) is

$$\Psi(\overline{\mathcal{M}}_q) \stackrel{\text{def}}{=} \sup_{\mathbf{S} \in \overline{\mathcal{M}}_q} \frac{\|\mathbf{S}\|_*}{\|\mathbf{S}\|_{\mathrm{F}}} \leq \sqrt{q}.$$

The conclusion of this theorem follows from Theorem 1 of Negahban et al. (2012). □

Lemma B.1. *Under (A1)–(A3),*

$$\mathrm{P}\left(\|\nabla g(\mathbf{\Gamma})\| \leq cm^{-1} \max(\tau, 1-\tau) K_u \sqrt{\|\mathbf{\Sigma}\|} \sqrt{\frac{p+m}{n}}\right) \geq 1 - 3 \times 8^{-(p+m)} - a_n, \tag{B.8}$$

where $c > 0$ is an absolute constant.

Proof for Lemma B.1. Throughout the proof, we restrict on the event Ω_1 in (B.7). Recall the expression from (A.3) that

$$\nabla g(\mathbf{\Gamma}) = -(mn)^{-1} \mathbf{X}^\top \left\{\mathbf{W} \circ (\mathbf{Y} - \mathbf{X}\mathbf{\Gamma})\right\}$$

and the matrix $\mathbf{U}(\mathbf{\Gamma}) = (u_{ij}) = (Y_{ij} - \boldsymbol{X}_i^\top \mathbf{\Gamma}_{\cdot j}) \in \mathbb{R}^{n\times m}$. Following the proof of Lemma 3 in Negahban and Wainwright (2011), we have

$$\begin{aligned}
\mathrm{P}\left(n^{-1}\|\mathbf{X}^\top(\mathbf{W}\circ\mathbf{U})\| \geq 4s\right) &= \mathrm{P}\left(\sup_{\substack{\boldsymbol{\beta}\in\mathcal{S}^{p-1},\\ \boldsymbol{\alpha}\in\mathcal{S}^{m-1}}} n^{-1}|\boldsymbol{\beta}^\top\mathbf{X}^\top(\mathbf{W}\circ\mathbf{U})\boldsymbol{\alpha}| \geq 4s\right)\\
&\leq 8^{p+m}\sup_{\substack{\boldsymbol{\beta}\in\mathcal{S}^{p-1},\\ \boldsymbol{\alpha}\in\mathcal{S}^{m-1}}} \mathrm{P}\left(n^{-1}|\langle\mathbf{X}\boldsymbol{\beta}, (\mathbf{W}\circ\mathbf{U})\boldsymbol{\alpha}\rangle| \geq s\right)\\
&\leq 8^{p+m}\sup_{\substack{\boldsymbol{\beta}\in\mathcal{S}^{p-1},\\ \boldsymbol{\alpha}\in\mathcal{S}^{m-1}}} \mathrm{P}\left(n^{-1}\sum_{i=1}^{n}\langle\boldsymbol{\beta}, \boldsymbol{X}_i\rangle\langle\boldsymbol{\alpha}, (\mathbf{W}\circ\mathbf{U})_i\rangle \geq s\right),
\end{aligned} \tag{B.9}$$

where $\mathcal{S}^{m-1} \overset{\text{def}}{=} \{\boldsymbol{\alpha} \in \mathbb{R}^m : \|\boldsymbol{\alpha}\|_2 = 1\}$ is the Euclidean sphere in m-dimensions. $\forall s \geq 0$, there exists $C > 0$ such that $\mathrm{P}(|u_{ij}| > s) \leq \exp(1 - s^2/C^2)$. Since $|w_{ij}| \leq \max(\tau, 1-\tau)$, we have

$$\begin{aligned}
\mathrm{P}\left(|w_{ij}u_{ij}| > s\right) &\leq \mathrm{P}\left(\max(\tau, 1-\tau)|u_{ij}| > s\right)\\
&= \mathrm{P}\left(|u_{ij}| > \frac{s}{\max(\tau, 1-\tau)}\right)\\
&\leq \exp\left(1 - \frac{s^2}{\max(\tau, 1-\tau)^2 C^2}\right).
\end{aligned} \tag{B.10}$$

It means for each $j \in \{1, \ldots, m\}$, $w_{ij}u_{ij}$ are sub-gaussian. Moreover, the maximal sub-gaussian norm is bounded by

$$\begin{aligned}
\max_{1\leq j\leq m}\|w_{ij}u_{ij}\|_{\psi_2} &= \max_{1\leq j\leq m}\sup_{p\geq 1} p^{-1/2}\left(\mathsf{E}|w_{ij}u_{ij}|^p\right)^{1/p}\\
&\leq \max(\tau, 1-\tau)\max_{1\leq j\leq m}\sup_{p\geq 1} p^{-1/2}\left(\mathsf{E}|u_{ij}|^p\right)^{1/p}\\
&= \max(\tau, 1-\tau)K_u.
\end{aligned} \tag{B.11}$$

Then by Hoeffding's inequality (Proposition 5.10 of Vershynin, 2012b), we can conclude that $\langle\boldsymbol{\alpha}, (\mathbf{W}\circ\mathbf{U})_i\rangle$ is also sub-guassian,

$$\begin{aligned}
\mathrm{P}\left(\langle\boldsymbol{\alpha}, (\mathbf{W}\circ\mathbf{U})_i\rangle \geq s\right) &= \mathrm{P}\left(\left|\sum_{j=1}^{m}\alpha_j w_{ij}u_{ij}\right| \geq s\right)\\
&\leq \exp\left(1 - \frac{C's^2}{\max(\tau, 1-\tau)^2 K_u^2\|\boldsymbol{\alpha}\|_2^2}\right)\\
&= \exp\left(1 - \frac{C's^2}{\max(\tau, 1-\tau)^2 K_u^2}\right),
\end{aligned} \tag{B.12}$$

where $C' > 0$ is an absolute constant. Furthermore, its sub-gaussian norm is bounded by

$$\begin{aligned}
\left\|\langle\boldsymbol{\alpha}, (\mathbf{W}\circ\mathbf{U})_i\rangle\right\|_{\psi_2} &= \sup_{p\geq 1} p^{-1/2}\left\{\mathsf{E}\left|\langle\boldsymbol{\alpha}, (\mathbf{W}\circ\mathbf{U})_i\rangle\right|^p\right\}^{1/p}\\
&= \sup_{p\geq 1} p^{-1/2}\left(\mathsf{E}\left|\sum_{j=1}^{m}\alpha_j w_{ij}u_{ij}\right|^p\right)^{1/p}\\
&\leq \max(\tau, 1-\tau)\sup_{p\geq 1} p^{-1/2}\left(\mathsf{E}\left|\sum_{j=1}^{m}\alpha_j u_{ij}\right|^p\right)^{1/p}\\
&\leq \max(\tau, 1-\tau)MK_u,
\end{aligned} \tag{B.13}$$

where $M > 0$ is an absolute constant. The last line comes from Khintchine inequality (Corollary 5.12 of Vershynin, 2012b) and recall that $\|\boldsymbol{\alpha}\|_2 = 1$. Applying Hoeffding's inequality again we can obtain

$$\begin{aligned}
\mathrm{P}\left(n^{-1}\sum_{i=1}^{n}\langle\boldsymbol{\beta}, \boldsymbol{X}_i\rangle\langle\boldsymbol{\alpha}, (\mathbf{W}\circ\mathbf{U})_i\rangle \geq s\right) &\leq \exp\left(1 - \frac{C''s^2 n}{\max(\tau, 1-\tau)^2 M^2 K_u^2 n^{-1}\sum_{i=1}^{n}\langle\boldsymbol{\beta}, \boldsymbol{X}_i\rangle^2}\right)\\
&\leq \exp\left(1 - \frac{C''s^2 n}{\max(\tau, 1-\tau)^2 M^2 K_u^2 n^{-1}\|\mathbf{X}\boldsymbol{\beta}\|_2^2}\right),
\end{aligned}$$

$$\leq \exp\left(1 - \frac{C''s^2n}{c_2 \max(\tau, 1-\tau)^2 M^2 K_u^2 \|\mathbf{\Sigma}\|}\right), \tag{B.14}$$

where C'' is an absolute constant. Combining (B.9) and (B.14) gives

$$\mathrm{P}\left(n^{-1}\left\|\mathbf{X}^\top(\mathbf{W}\circ\mathbf{U})\right\| \geq 4s\right) \leq \exp\left(1 - \frac{C''s^2n}{9\max(\tau, 1-\tau)^2 M^2 K_u^2 \|\mathbf{\Sigma}\|} + (p+m)\log 8\right). \tag{B.15}$$

Set $s = \frac{1}{4}c\max(\tau, 1-\tau)K_u\sqrt{\|\mathbf{\Sigma}\|}\sqrt{\frac{p+m}{n}}$, where $c \overset{\text{def}}{=} 4\cdot\sqrt{2\log 8\frac{9M^2}{C''}}$, then we can conclude from the fact $P(\Omega_1) \geq 1 - a_n$,

$$\begin{aligned}
&\mathrm{P}\left(n^{-1}\left\|\mathbf{X}^\top(\mathbf{W}\circ\mathbf{U})\right\| \leq c\max(\tau, 1-\tau)K_u\sqrt{\|\mathbf{\Sigma}\|}\sqrt{\frac{p+m}{n}}\right)\\
&\quad\geq \left[1 - \exp\big(1 - (p+m)\log 8\big)\right] \times (1 - a_n)\\
&\quad\geq \left[1 - 3\times 8^{-(p+m)}\right] \times (1 - a_n)\\
&\quad\geq 1 - 3\times 8^{-(p+m)} - a_n \quad (\text{as } p+m > 1).
\end{aligned} \tag{B.16}$$

This finishes the proof. □

Lemma B.2. *Under (A1)–(A3), selecting* $\lambda = 2\,\mathrm{cm}^{-1}\max(\tau, 1-\tau)K_u\sqrt{\|\mathbf{\Sigma}\|}\sqrt{\frac{p+m}{n}}$, *for* $n \geq 2\min(m, p)$, *any optimal solution* $\widehat{\mathbf{\Gamma}}$ *in the problem (2.6) satisfies the bound*

$$\begin{aligned}
\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_{\mathrm{F}}^2 \leq& c'\frac{p+m}{n}\frac{\max(\tau, 1-\tau)^2}{\min(\tau, 1-\tau)^2}\frac{\|\mathbf{\Sigma}\|}{\sigma_{\min}(\mathbf{\Sigma})^2}K_u^2\dim(\overline{\mathcal{M}})\\
&+ c'\sqrt{\frac{p+m}{n}}\frac{\max(\tau, 1-\tau)}{\min(\tau, 1-\tau)}\frac{\sqrt{\|\mathbf{\Sigma}\|}}{\sigma_{\min}(\mathbf{\Sigma})}K_u\|\mathbf{\Gamma}_{\mathcal{M}^\perp}\|_*,
\end{aligned} \tag{B.17}$$

with probability greater than $1 - 3\times 8^{-(p+m)} - a_n$, *where* $c, c' > 0$ *are absolute constants.*

Proof of Lemma B.2. Recall that Ω_1 is defined as (B.7), and let the event that (B.8) holds as Ω_2. On event $\Omega_1 \cap \Omega_2$, (B.17) can be achieved by simply plugging $\lambda = 2\,\mathrm{cm}^{-1}\max(\tau, 1-\tau)K_u\sqrt{\|\mathbf{\Sigma}\|}\sqrt{\frac{p+m}{n}}$ into (B.4). We note that

$$\begin{aligned}
\mathrm{P}(\Omega_2\cap\Omega_1) = \mathrm{P}(\Omega_2|\Omega_1)\,\mathrm{P}(\Omega_1) &\geq \left[1 - 3\times 8^{-(p+m)}\right] \times (1 - a_n)\\
&\geq 1 - 3\times 8^{-(p+m)} - a_n \quad (\text{as } p+m > 1). \quad \square
\end{aligned} \tag{B.18}$$

Appendix C. Auxiliary results

Lemma C.1. *For any* $u, \delta \in \mathbb{R}$ *and* $\tau \in (0, 1)$,

$$\rho_\tau(u+\delta) - \rho_\tau(u) - \rho_\tau'(u)\delta \geq \min(\tau, 1-\tau)\delta^2. \tag{C.1}$$

Proof of Lemma C.1. When $u = 0$, we have $\rho_\tau(u) = \rho_\tau'(u) = 0$, therefore

$$\rho_\tau(\delta) = |\tau - \mathbf{1}\{\delta < 0\}|\delta^2 \geq \min(\tau, 1-\tau)\delta^2.$$

If $u > 0, u + \delta < 0$ $(\delta < 0)$, we have

$$\rho_\tau(u+\delta) - \rho_\tau(u) - \rho_\tau'(u)\delta - \min(\tau, 1-\tau)\delta^2 = \begin{cases}(1-2\tau)(\delta+u)^2 \geq 0 & \text{for } \tau \leq 1-\tau;\\ (1-2\tau)(u+2\delta)u > 0 & \text{for } \tau > 1-\tau.\end{cases}$$

If $u > 0, u + \delta > 0$ $(\delta > 0)$, we have

$$\rho_\tau(u+\delta) - \rho_\tau(u) - \rho_\tau'(u)\delta - \min(\tau, 1-\tau)\delta^2 = \begin{cases}(2\tau-1)(u+2\delta)u \geq 0 & \text{for } \tau \leq 1-\tau;\\ (2\tau-1)(u+\delta)^2 u > 0 & \text{for } \tau > 1-\tau.\end{cases}$$

In the other two cases,

$$\rho_\tau(u+\delta) - \rho_\tau(u) - \rho_\tau'(u)\delta = \begin{cases}\tau\delta^2 \geq \min(\tau, 1-\tau)\delta^2 & \text{for } u > 0, u+\delta \geq 0;\\ (1-\tau)\delta^2 \geq \min(\tau, 1-\tau)\delta^2 & \text{for } u < 0, u+\delta \leq 0.\end{cases}$$

Therefore, we can conclude that

$$\rho_\tau(u+\delta) - \rho_\tau(u) - \rho_\tau'(u)\delta \geq \min(\tau, 1-\tau)\delta^2.$$

Table D.1
Tuning parameters by 5-fold cross validation.

		Whole series (C1)			Task-wise (C2)		
τ		0.1	0.5	0.9	0.1	0.5	0.9
$\omega = 0.1$	$aINS_L$	0.0442	0.0552	0.0383	0.0008	0.0006	0.0008
	$aINS_R$	0.0303	0.0421	0.0293	0.0004	0.0008	0.0004
	DMPFC	0.0348	0.0504	0.0198	0.0004	0.0007	0.0006
$\omega = 0.5$	$aINS_L$	0.0181	0.0403	0.0153	0.0004	0.0006	0.0003
	$aINS_R$	0.0137	0.0393	0.0157	0.0006	0.0004	0.0005
	DMPFC	0.0195	0.0391	0.0143	0.0006	0.0002	0.0007
$\omega = 0.9$	$aINS_L$	0.0253	0.0408	0.0275	0.0006	0.0004	0.0004
	$aINS_R$	0.0243	0.0442	0.0200	0.0008	0.0002	0.0006
	DMPFC	0.0193	0.0474	0.0206	0.0005	0.0008	0.0008

Lemma C.2. *$g(\mathbf{\Gamma})$ defined in (2.10) is κ-strongly convex and differentiable with $\kappa = m^{-1}\min(\tau, 1-\tau)\sigma_{\min}(\frac{\mathbf{X}^\top\mathbf{X}}{n})$.* □

Proof of Lemma C.2. Denote $\widetilde{u}_{ij} \overset{\text{def}}{=} Y_{ij} - \boldsymbol{X}_i^\top(\mathbf{\Gamma}_{\cdot j} + \mathbf{\Delta}_{\cdot j})$ and $u_{ij} \overset{\text{def}}{=} Y_{ij} - \boldsymbol{X}_i^\top\mathbf{\Gamma}_{\cdot j}$, for $i = 1, \ldots, n, j = 1, \ldots, m$, we have

$$\begin{aligned}
\langle\langle \nabla g(\mathbf{\Gamma}), \mathbf{\Delta}\rangle\rangle &= \operatorname{tr}\big(\nabla g(\mathbf{\Gamma})^\top \mathbf{\Delta}\big) \\
&= -(mn)^{-1}\sum_{j=1}^{m}\sum_{l=1}^{p}\mathbf{\Delta}_{lj}\sum_{i=1}^{n}\rho'(u_{ij})X_{il} \\
&= -(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\Big\{\sum_{l=1}^{p}\mathbf{\Delta}_{lj}\rho'(u_{ij})X_{il}\Big\} \\
&= -(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\big\{\rho'(u_{ij})\boldsymbol{X}_i^\top\mathbf{\Delta}_{\cdot j}\big\}. \qquad \text{(C.2)}
\end{aligned}$$

Therefore,

$$\begin{aligned}
g(\mathbf{\Gamma}+\mathbf{\Delta}) - g(\mathbf{\Gamma}) - \langle\langle \nabla g(\mathbf{\Gamma}), \mathbf{\Delta}\rangle\rangle &= (mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\big\{\rho(\widetilde{u}_{ij}) - \rho(u_{ij}) + \rho'(u_{ij})\boldsymbol{X}_i^\top\mathbf{\Delta}_{\cdot j}\big\} \\
&\geq (mn)^{-1}\min(\tau, 1-\tau)\sum_{i=1}^{n}\sum_{j=1}^{m}(\boldsymbol{X}_i^\top\mathbf{\Delta}_{\cdot j})^2 \text{ (by Lemma C.1)} \\
&= (mn)^{-1}\min(\tau, 1-\tau)\|\mathbf{X}\mathbf{\Delta}\|_{\mathrm{F}}^2 \\
&= (mn)^{-1}\min(\tau, 1-\tau)\operatorname{tr}(\mathbf{\Delta}^\top\mathbf{X}^\top\mathbf{X}\mathbf{\Delta}) \\
&\geq m^{-1}\min(\tau, 1-\tau)\sigma_{\min}\Big(\frac{\mathbf{X}^\top\mathbf{X}}{n}\Big)\|\mathbf{\Delta}\|_{\mathrm{F}}^2. \quad \square \qquad \text{(C.3)}
\end{aligned}$$

Appendix D. Additional details for Section 4

D.1. Tuning parameters by cross-validation

Choosing $\omega = 0.1, b = 1$ (aINS_L cluster) in (C1) case as an example, Fig. D.1 illustrates the cross-validation error function in terms of λ under different τ levels. The optimal tuning parameters determined by 5-fold cross-validation under all cases are reported in Table D.1.

D.2. Risk attitude parameter

The risk attitude parameter β is estimated by logistic model via maximum likelihood estimation (MLE)

$$\begin{aligned}
\mathrm{P}\{\text{risky choice}|x\} &= \frac{1}{1+\exp[-\sigma\{\bar{x} - \beta S(x) - 5\}]}, \\
\mathrm{P}\{\text{sure choice}|x\} &= 1 - \frac{1}{1+\exp[-\sigma\{\bar{x} - \beta S(x) - 5\}]}, \qquad \text{(D.1)}
\end{aligned}$$

where x is the return stream displayed to the individual, its mean and standard deviation are $\bar{x}$ and $S(x)$.

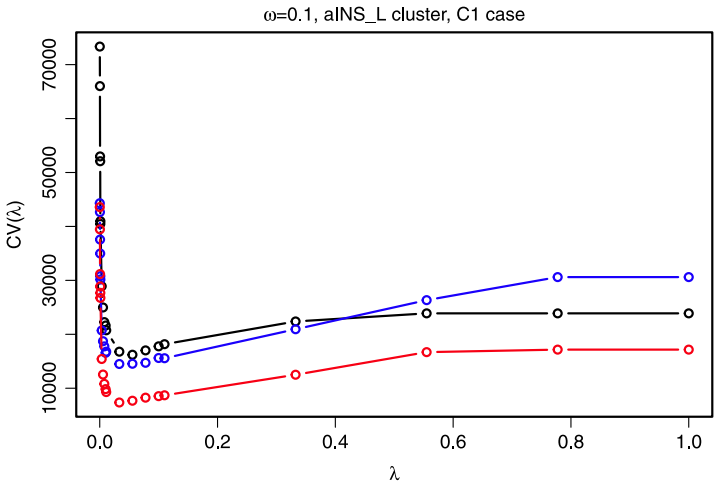


Fig. D.1. The cross-validation error function in terms of tuning parameter λ, with $\tau = 0.1$, 0.5, and 0.9, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

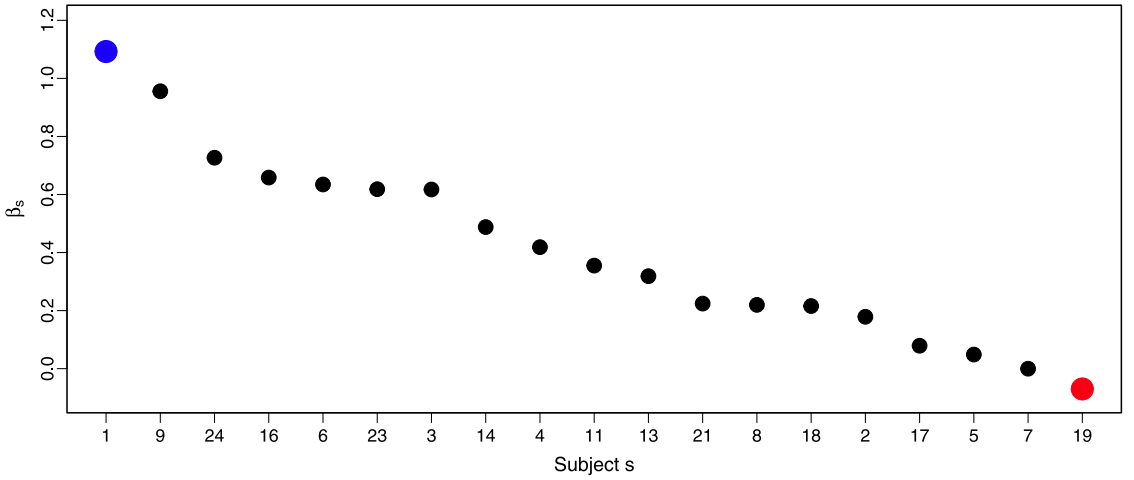


Fig. D.2. Estimated risk attitude for 19 subjects.

The estimated risk attitude parameters for 19 subjects in order are plotted in Fig. D.2, also see Majer et al. (2016). Negative parameters imply risk-seeking behaviors; while positive parameters indicate averse risk patterns. We can see most of the individuals are risk-averse and the two extremes #1 and #19 are the most risk-averse and most risk-seeking persons respectively.

References

Beck, A., Teboulle, M., 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imag. Sci. 2 (1), 183–202.
Camerer, C.F., 2007. Neuroeconomics: Using neuroscience to make economic predictions. Econom. J. 117 (519), C26–C42.
Chao, S.-K., Härdle, W.K., Yuan, M., 2016. Factorisable multi-task quantile regression, SFB 649 Discussion Paper 2016-057, Sonderforschungsbereich 649, Humboldt Universität zu Berlin, Germany. Available at http://sfb649.wiwi.hu-berlin.de/papers/pdf/SFB649DP2016-057.pdf.
Chao, S.-K., Proksch, K., Dette, H., Härdle, W.K., 2017. Confidence corridors for multivariate generalized quantile regression. J. Bus. Econom. Statist. 35 (1), 70–85.
Degras, D., Lindquist, M.A., 2014. A hierarchical model for simultaneous detection and estimation in multi-subject fMRI studies. NeuroImage 98, 61–72.
Fadili, J.M., Peyré, G., 2011. Total variation projection with first order schemes. IEEE Trans. Image Process. 20 (3), 657–669.
Heekeren, H.R., Marrett, S., Ungerleider, L.G., 2008. The neural systems that mediate human perceptual decision making. Nat. Rev. Neurosci. 9 (6), 467–479.
Helfinstein, S.M., Schonberg, T., Congdon, E., Karlsgodt, K.H., Mumford, J.A., Sabb, F.W., Cannon, T.D., London, E.D., Bilder, R.M., Poldrack, R.A., 2014. Predicting risky choices from brain activity patterns. Proc. Natl. Acad. Sci. 111 (7), 2470–2475.
Izenman, A.J., 1975. Reduced-rank regression for the multivariate linear model. J. Multivariate Anal. 5 (2), 248–264.
Ji, S., Ye, J., 2009. An accelerated gradient method for trace norm minimization. In: Proceedings of the 26th International Conference on Machine Learning.

Knofczynski, G.T., Mundfrom, D., 2008. Sample sizes when using multiple linear regression for prediction. Educ. Psychol. Meas. 68 (3), 431–442.
Koenker, R., Bassett, G.W., 1978. Regression quantiles. Econometrica 46 (1), 33–50.
Koenker, R., Portnoy, S., 1990. *M* Estimation of multivariate regressions. J. Amer. Statist. Assoc. 85 (412), 1060–1068.
Lindquist, M.A., 2008. The statistical analysis of fMRI data. Statist. Sci. 23 (4), 439–464.
Majer, P., Mohr, P.N.C., Heekeren, H., Härdle, W.K., 2016. Portfolio decisions and brain reactions via the CEAD method. Psychometrika 81 (3), 881–903.
Mullinger, K., Mayhew, S., Bagshaw, A., Bowtell, R., Francis, S., 2014. Evidence that the negative BOLD response is neuronal in origin: A simultaneous EEG–BOLD–CBF study in humans. NeuroImage 94, 263–274.
Negahban, S.N., Ravikumar, P., Wainwright, M.J., Yu, B., 2012. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. Statist. Sci. 27 (4), 538–557.
Negahban, S.N., Wainwright, M.J., 2011. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. Ann. Statist. 39 (2), 1069–1097.
Newey, W.K., Powell, J.L., 1987. Asymmetric least squares estimation and testing. Econometrica 55 (4), 819–847.
Opsomer, J., Wang, Y., Yang, Y., 2001. Nonparametric regression with correlated errors. Statist. Sci. 16 (2), 134–153.
Reinsel, G.C., Velu, R.P., 1998. Multivariate Reduced-Rank Regression. Springer, New York.
Rossi, G.D., Harvey, A., 2009. Quantiles, expectiles and splines. J. Econometrics 152 (2), 179–185.
Schultz, W., 2015. Neuronal reward and decision signals: From theories to data. Physiol. Rev. 95 (3), 853–951.
van Bömmel, A., Song, S., Majer, P., Mohr, P.N.C., Heekeren, H.R., Härdle, W.K., 2014. Risk patterns and correlated brain activities. multidimensional statistical analysis of fMRI data in economic decision making study. Psychometrika 79 (3), 489–514.
Vershynin, R., 2012a. How close is the sample covariance matrix to the actual covariance matrix? J. Theoret. Probab. 25 (3), 655–686.
Vershynin, R., 2012b. Introduction to the non-asymptotic analysis of random matrices. In: Eldar, Y., Kutyniok, G. (Eds.), Compressed Sensing, Theory and Applications. Cambridge University Press, pp. 210–268 (Chapter 5).
Wainwright, M.J., 2009. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). IEEE Trans. Inform. Theory 55, 2183–2202.
Yuan, M., Ekici, A., Lu, Z., Monteiro, R., 2007. Dimension reduction and coefficient estimation in multivariate linear regression. J. R. Stat. Soc. Ser. B Stat. Methodol. 69 (3), 329–346.

Research Article

Ying Chen*, Wolfgang K. Härdle, Qiang He and Piotr Majer

Risk related brain regions detection and individual risk classification with 3D image FPCA

https://doi.org/10.1515/strm-2017-0011
Received April 14, 2017; revised August 3, 2018; accepted October 13, 2018

Abstract: Understanding how people make decisions from risky choices has attracted increasing attention of researchers in economics, psychology and neuroscience. While economists try to evaluate individual's risk preference through mathematical modeling, neuroscientists answer the question by exploring the neural activities of the brain. We propose a model-free method, 3-dimensional image functional principal component analysis (3DIF), to provide a connection between active risk related brain region detection and individual's risk preference. The 3DIF methodology is directly applicable to 3-dimensional image data without artificial vectorization or mapping and simultaneously guarantees the contiguity of risk related brain regions rather than discrete voxels. Simulation study evidences an accurate and reasonable region detection using the 3DIF method. In real data analysis, five important risk related brain regions are detected, including parietal cortex (PC), ventrolateral prefrontal cortex (VLPFC), lateral orbifrontal cortex (lOFC), anterior insula (aINS) and dorsolateral prefrontal cortex (DLPFC), while the alternative methods only identify limited risk related regions. Moreover, the 3DIF method is useful for extraction of subjective specific signature scores that carry explanatory power for individual's risk attitude. In particular, the 3DIF method perfectly classifies both strongly and weakly risk averse subjects for in-sample analysis. In out-of-sample experiment, it achieves 73 %–88 % overall accuracy, among which 90 %–100 % strongly risk averse subjects and 49 %–71 % weakly risk averse subjects are correctly classified with leave-k-out cross validations.

Keywords: fMRI, FPCA, GLM, risk attitude, SVD

MSC 2010: 62H12, 62P10

1 Introduction

Understanding people's risk preferences and how people make decisions under risk have both attracted much attention in industry and academia alike. Accurate risk classification is of benefit both to creditors including banks, retailers, mail order companies, utilities and various other organizations, and to the applicants avoiding over commitment, see [16]. While the traditional classification approaches rely on expert knowledge,

***Corresponding author: Ying Chen,** Department of Mathematics, National University of Singapore, Singapore; and Department of Statistics and Applied Probability, National University of Singapore, Singapore; and Risk Management Institute, National University of Singapore, Singapore, e-mail: stacheny@nus.edu.sg. http://orcid.org/0000-0002-2577-7348
Wolfgang K. Härdle, Ladislaus von Bortkiewicz Chair of Statistics, C.A.S.E. Center for Applied Statistics & Economics, Humboldt-Universität zu Berlin, Berlin, Germany; and Sim Kee Boon Institute (SKBI) for Financial Economics at Singapore Management University, Singapore, e-mail: haerdle@wiwi.hu-berlin.de
Qiang He, Department of Statistics and Applied Probability, National University of Singapore, Singapore, e-mail: hq19861027@gmail.com
Piotr Majer, Ladislaus von Bortkiewicz Chair of Statistics, C.A.S.E. Center for Applied Statistics & Economics, Humboldt-Universität zu Berlin, Berlin, Germany, e-mail: piotr.majer71@gmail.com

experience and even a subjective feeling to categorize an individual to be risk averse or risk seeking, there has been an increasing demand in statistical methods for quantitative complements to the formal art alike analysis systems. Discriminant analysis, linear regression, logistic regression and decision trees have been developed and implemented in literature.

To explain the decision making behaviors, classical expected utility theory has been widely adopted in economics, see [23, 28, 31, 50]. The utility theory assumes that a rational decision maker chooses a strategy that maximizes the average or expected value of a concave utility function among possible outcomes, see e.g. [36] for the properties of utility functions. The utility functions depend on parameters that represent individual's risk preferences and are estimated based on the individual's characteristics. Alternatively, risk-return models [30] determine the average or expected returns and the associated risks of different choices, and compute a risk-compensated value in the capital asset pricing models, see [43, 51, 52]. The traditional models, though demonstrating some decision making philosophy in a common sense, are unable to explain the heterogeneity in decision-making under similar risk attitudes from person to person in the experiments of behavioral economics and neuroscience, see [3, 5, 10, 21, 44].

Decision-making is indeed a complex neural process involving both cognitive and emotional factors. According to [23] and [44], individuals not only estimate the expected value of utility or return, but more importantly, they seem to adapt these estimates by subjective factors, such as risk preference. It thus becomes scientifically necessary and important to answer which parts of the human brain regulate specific decision-making tasks and which neural processes drive investment decisions, see [25, 33, 37, 41]. It is also interesting to ask whether the identification of the risk related brain regions helps to explain the heterogeneity of individual risk preference and its impact on making decision from the neural aspect.

The recent development on neural image data collection allows quantitative analysis to be possible. In modern risk perception and investment decision (RPID) experiments, subjects are requested to make decisions with uncertain outcomes and simultaneously their brain reactions are recorded as neural images by the functional magnetic resonance imaging (fMRI) scanner. The neural images or fMRI data reflects the changes in the brain's blood flow at volume and oxygen level during neural activities. The blood-oxygen-level-dependent (BOLD) signals are captured on 3-dimensional (3D) spatial maps of brain voxels during the experiments.

Given the fMRI data collected in the risk related experiments, specific brain regions have been found to be associated with risk related decision making. Tobler, O'Doherty, Dolan and Schultz [45] demonstrated that lateral orbifrontal cortex (lOFC) and medial orbifrontal cortex (mOFC) are related to the evaluation and the contrast of risky or sure choices. Mohr, Biele, Krugel, Li and Heekeren [33] discovered that risk averse individuals have greater brain activities in lateral orbifrontal cortex (lOFC) and posterior cingulate cortex (PCC). Mohr, Biele and Heekeren [32] evidenced the importance of anterior insula (aINS) and ventrolateral prefrontal cortex (VLPFC) to value processing, risk and uncertainty. Van Bömmel, Song, Majer, Mohr, Heekeren and Härdle [47] found parietal cortex (PC) is associated with value processing and selective attention. The risk related regions are quantified as the voxels significantly activated by the stimulus, which turn out to be contiguous in modest size relative to the visual or audial cortex. Two techniques – general linear model (GLM) method and principal component analysis (PCA) method – are by far the most popular to identify the risk related regions.

The model-based GLM technique depends on a parametric structure, see e.g. [9, 11, 48]. It only focuses on the neural information with a pre-defined design matrix and ignores any neural activity other than the priori specified modeling. The PCA technique is model free and has potential to detect risk related regions without making any constraint or subjective assumptions, see [2, 4, 27]. Without losing much variability, it extracts spatial factors to represent the risk related brain regions, while the individual risk attitude of the subject is explained by the factor loadings named signature scores via an orthogonal decomposition.

The PCA method however needs a conversion of the fMRI data to a vector of discrete signals, leading to extremely high dimensionality when applied to the high resolution image data. To solve the estimation challenge, singular value decomposition (SVD) has been proposed with a reduced dimension of covariance matrix, see [13]. Nevertheless, the PCA and SVD methods conducted in a discrete framework cannot guarantee the contiguity of risk related regions rather discrete voxels, see [19].

This motivates the adoption of functional principal component analysis (FPCA), see [39, 40]. In FPCA, the vectorized fMRI data is smoothed as a continuous curve, for which eigen-decomposition is performed, see [29, 47, 49]. Zipunnikov, Caffo, Yousem, Davatzikos, Schwartz and Crainiceanu [54] further proposed the functional SVD (FSVD) approach that improved computational efficiency with the utilization of the SVD technique. It is worth noting that the FPCA and FSVD methods both request vectorizing the BOLD signals that are naturally defined on 3D location coordinates to 1D domain. Given the high resolution of fMRI data, without sufficient knowledge of spatial interdependence of the brain, the pre-processing vectorization potentially impairs accuracy and efficiency for the risk related region detection and further for the risk classification.

It is necessary to ask why not directly analyze the fMRI signals in the 3D domain and how much accuracy can be improved by employing such a new technique. In our study, we propose a model-free 3-dimensional image functional principal component analysis (3DIF) method to identify risk related regions and extract subject signature scores. Simulation study and real data analysis demonstrate good quality of the detected risk related regions with stable accuracy and contiguity property. The 3DIF regions are further found to carry explanatory power for subjects' risk attitudes. In the application of risk classification, the 3DIF method reaches 100 % accuracy for in-sample analysis and 73 %–88 % overall accuracy for out-of-sample analysis. In particular, it correctly classifies 90 %–100 % strongly risk averse subjects and 49 %–71 % weakly risk averse subjects by using leave-k-out cross validations.

The remainder of the paper is structured as follows. Section 2 presents the RPID experiment and data. Section 3 details the 3DIF methodology and briefly reviews the alternative methods in literature. Section 4 reports the performance of the proposed 3DIF method under different scenarios. In Section 5, we implement the 3DIF to real data. Section 6 concludes.

2 RPID experiment and data

To investigate the mechanism of brain processes during the process of making decisions under risk, we analyze functional magnetic resonance imaging (fMRI) data on seventeen subjects who were exposed to an RPID experiment designed in [33]. The experiment uses streams of investment returns as stimuli and hypothesizes how individual risk attitude affects decisions in risky choices against sure choices. Figure 1 displays a graphic illustration of the experimental setup. Each experiment trial composes of two phases. The presentation phase displays a random Gaussian distributed return stream with ten observations that are sequentially displayed over 2×10 seconds. After a 2.5 seconds break, subjects are exposed in the decision phase to one of three types of tasks and have to give an answer within the next 7 seconds. The three types of tasks included the *decision* task, where subjects choose either a 5 % fixed return (sure choice) or the investment of the random return stream just shown (risky choice). In the other two tasks subjects report their *subjective expected return* (scaling from 5 % to 15 %) and *perceived risk* (from 0 = no risk to 100 = maximum risk) of the just displayed investment. Each trial is repeated 27 times, with the types of tasks randomly selected. In total, there are 3×27 trails for each subject. During the experiment, subjects were placed in the fMRI scanner and high resolution ($91 \times 109 \times 91$) images were acquired every 2.5 seconds.

The seventeen subjects were native German speakers, healthy and right-handed. All participants had no history of neurological or psychiatric diseases. They were paid for their participation and gave written informed consent. The return streams were independent from trial to trial, randomly drawn from a Gaussian distribution. The expected value of the return streams varied from 6 %, 9 %, to 12 % and standard deviations from 1 %, 5 % to 9 %. The combinations generated in total nine different Gaussian distributions associated with various risk-return relationships, e.g. low return (6 %) and low risk (1 %) as well as high return (12 %) and high risk (9 %).

The same data had been studied by two works in the existing literature. Mohr, Biele, Krugel, Li and Heekeren [33] conducted the general linear model (GLM) with six design factors. The factors are either subject specific values including e.g. return stream, perceived risk, expected value of the return stream, or dummy variables. The study detected value-reward related brain activity in bilateral dorsolateral prefrontal cortex

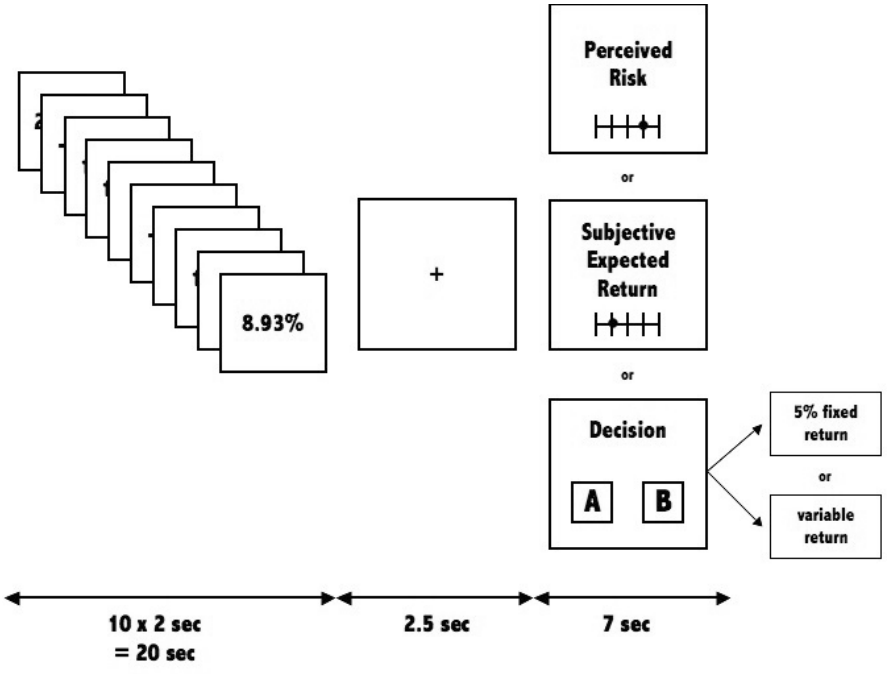


Figure 1: Graphic illustration of one trail of the RPID experiment, see [33].

(DLPFC), posterior cingulate cortex (PCC), ventrolateral prefrontal cortex (VLPFC), and medial prefrontal cortex (MPFC), which is consistent to [1, 22, 24–26, 35, 46]. It also found that perceived risk correlated significantly with the BOLD signal in the anterior insula (aINS), as documented in a variety of studies by [8, 14, 20, 34, 37, 38, 42]. However, GLM detection depends on the significance of statistical tests, which are hard to extract subject specific signals for further analysis.

Van Bömmel, Song, Majer, Mohr, Heekeren and Härdle [47] proposed a panel version of the dynamic semiparametric factor model (PDSFM) to reanalyze the data. The approach however only detected two important risk-related regions and did not contain any activation regions previously reported in [33] except Parietal Cortex (PC). Subject signature scores were extracted and used in risk classification. Using the variance of these stimuli responses as input for the classification algorithm, it obtained very high classification rates at 97 % for strongly risk averse individuals and 75 % for weakly risk averse with the SVM classifier by applying the double leave-one-out cross-validation algorithm.

3 Method

Our interest is to propose a dimension reduction technique on 3D space to improve prediction in the fMRI study of association between risk preferences and brain activity. In this section, we detail the 3D image functional principal component analysis (3DIF) method that is directly applicable to high-dimensional functional data and guarantees the contiguity of detected risk related brain regions. We show how to identify common spatial factors and extract subjective specific scores. The spatial factors are used to construct common risk activation regions that do not dependent on subjects, while the heterogeneity of individual risk attitude is explained by the subjective specific scores.

Let $Y_t^{(j)}(x_1, x_2, x_3)$ denote the observed fMRI signal at time $t = 1, \ldots, N$ for subject $j = 1, \ldots, J$ at 3D spatial location (x_1, x_2, x_3), where $x_1 \in \mathcal{P}_1, x_2 \in \mathcal{P}_2, x_3 \in \mathcal{P}_3$ are defined in a bounded cube $\mathcal{P}_1 \times \mathcal{P}_2 \times \mathcal{P}_3 \subset \mathbb{R}^3$. In our study, $J = 17$ subjects and $N = 1360$ scanned images. The brain is measured in a cube of size $[1, 91] \times [1, 109] \times [1, 91]$, i.e. around 10^6 voxels per scan. A tensor B-spline smoother is used to smooth each time-specific brain image and it leads to continuous 3D functional data, denoted as $f_t^{(j)}(x_1, x_2, x_3)$.

3.1 3D image functional principal component analysis (3DIF)

For any continuous functional data $f_t(\boldsymbol{x})$ with $\boldsymbol{x} = (x_1, x_2, x_3)$, one can represent it in a vector format

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{C}\boldsymbol{\phi}(\boldsymbol{x}), \tag{3.1}$$

where $\boldsymbol{C}$ is an $(N \times K^3)$-dimensional matrix of B-spline coefficients, N is the number of time points in the fMRI data and K refers to the number of knots in each spatial direction, and

$$\boldsymbol{\phi}(\boldsymbol{x}) = [\phi_1(x_1, x_2, x_3), \phi_2(x_1, x_2, x_3), \ldots, \phi_{K^3}(x_1, x_2, x_3)]^\top$$

are the continuous basis functions generated by tensor products of univariate splines. Thus K^3 is the total number of the basis functions.

In the factor extraction experiment, we are able to assume the fMRI images to be independent and identically distributed. Denote the covariance function of the functional data

$$G(\boldsymbol{x}, \boldsymbol{s}) = \mathrm{Cov}\{f(\boldsymbol{x}), f(\boldsymbol{s})\}$$

and its sample estimator

$$\widehat{G}(\boldsymbol{x}, \boldsymbol{s}) = N^{-1} \sum_{t=1}^{N} f_t(\boldsymbol{x}) f_t(\boldsymbol{s}). \tag{3.2}$$

The covariance operator V is defined as

$$Vf = \int_{\mathcal{P}_1} \int_{\mathcal{P}_2} \int_{\mathcal{P}_3} G(\cdot, \boldsymbol{x}) f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}.$$

Similarly to the orthogonal decomposition in the multivariate PCA, we have for the 3D image functional data

$$V\xi = \int_{\mathcal{P}_1} \int_{\mathcal{P}_2} \int_{\mathcal{P}_3} G(\cdot, \boldsymbol{x}) \xi(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \lambda \xi(\boldsymbol{x}),$$

where $\xi(\boldsymbol{x})$ and λ denote the eigenfunction on the 3D domain and the eigenvalue respectively. The eigenvalues are real and non-negative $\lambda_1 > \lambda_2 > \cdots \geq 0$. Without spatial information loss or distortion due to vectorization in e.g. FPCA, the first functional factor $\xi_1(x_1, x_2, x_3)$ corresponding to the largest eigenvalue λ_1 accounts for as much of the variability in the data as possible, and each succeeding functional factor $\xi_\ell(x_1, x_2, x_3)$ in turn has the highest variance possible under the constraint that it is uncorrelated with the preceding ones.

Plugging (3.1) into (3.2), we obtain

$$\widehat{G}(\boldsymbol{s}, \boldsymbol{x}) = N^{-1} \boldsymbol{\phi}^\top(\boldsymbol{s}) \boldsymbol{C}^\top \boldsymbol{C} \boldsymbol{\phi}(\boldsymbol{x}),$$

and the orthogonal decomposition equation as

$$\iiint N^{-1} \boldsymbol{\phi}^\top(\boldsymbol{s}) \boldsymbol{C}^\top \boldsymbol{C} \boldsymbol{\phi}(\boldsymbol{x}) \boldsymbol{\phi}^\top(\boldsymbol{x}) \boldsymbol{b} \, \mathrm{d}(\boldsymbol{x}) = \lambda \boldsymbol{\phi}^\top(\boldsymbol{s}) \boldsymbol{b},$$

where the eigenfunction $\xi = \boldsymbol{\phi}^\top \boldsymbol{b}$ with $\boldsymbol{b}$ being a coefficient vector. Define

$$\boldsymbol{W} = \iiint \boldsymbol{\phi}(\boldsymbol{x}) \boldsymbol{\phi}^\top(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}.$$

By solving

$$N^{-1} \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{C}^\top \boldsymbol{C} \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{u} = \lambda \boldsymbol{u}, \tag{3.3}$$

where $\boldsymbol{u} = \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{b}$ and the coefficient vector $\boldsymbol{b}$ satisfies $\boldsymbol{b}_i^\top \boldsymbol{W} \boldsymbol{b}_i = 1$ and $\boldsymbol{b}_i^\top \boldsymbol{W} \boldsymbol{b}_j = 0$, we obtain the eigenfunction that contains spatial information and hence will be used to construct the common spatial factors of the fMRI data.

3.2 Multilinear model

To obtain common spatial factors across subjects, we borrow the idea of panel data analysis by averaging fMRI signals over subjects at each time t:

$$\bar{Y}_t(x_1, x_2, x_3) = J^{-1} \sum_{j=1}^{J} Y_t^{(j)}(x_1, x_2, x_3), \quad t = 1, \ldots, N.$$

The averaged signals are smoothed over a tensor B-spline regression with $K = 16$ knots in every spatial direction. The eigenfunctions are obtained by the 3DIF orthogonal decomposition in Section 3.1.

The eigenfunctions consist of not only important regions attributed to risk perception and investment decisions but also other neural activities unrelated to the investigated stimuli and possible magnetic noises. To remove the impact of noises, the spatial factors are constructed by trimming the eigenfunctions at extreme quantiles such as [0.05 %, 99.95 %] levels and replacing the "non-active" voxels with zeros. Moreover, we only consider the first L eigenfunctions and denote the trimmed factors as common risk related regions, denoted as $\widehat{\xi}_\ell(x_1, x_2, x_3)$ with $\ell = 1, \ldots, L$, since only the first spatial factors are fundamental and necessary. By doing this, the original high dimensionality is reduced to a small number of common spatial factors.

Heterogeneity of individual risk attitude are extracted in the multilinear regression that projects the raw fMRI signals on the common spatial regions:

$$Y_t^{(j)}(x_1, x_2, x_3) = \sum_{\ell=1}^{L} Z_{\ell,t}^{(j)} \widehat{\xi}_\ell(x_1, x_2, x_3) + \varepsilon_t^{(j)}(x_1, x_2, x_3), \tag{3.4}$$

where $\varepsilon_t^{(j)}(x_1, x_2, x_3)$ denotes the idiosyncratic noise of the j-th subject, which is independently and identically distributed with zero mean and constant variance. The subject-specific factor loadings $Z_{\ell,t}^{(j)}$ are calculated by ordinary least squares regression at time t for subject j:

$$\min_{Z_{\ell,t}^{(j)}} \sum_{x_1, x_2, x_3} \left\{ Y_t^{(j)}(x_1, x_2, x_3) - \sum_{\ell=1}^{L} Z_{\ell,t}^{(j)} \widehat{\xi}_l(x_1, x_2, x_3) \right\}^2.$$

The multi-subject 3DIF estimation procedure can now be summarized as follows:

(1) Take the average $\bar{Y}_t(x_1, x_2, x_3)$ of the raw 3D fMRI data across all subjects and obtain the smoothed 3D image functional data $f_t(x_1, x_2, x_3)$.
(2) Perform 3DIF to construct common spatial functional factors $\widehat{\xi}_\ell(x_1, x_2, x_3)$ via (3.3) and trim out insignificant active regions at e.g. 0.05 %– and 99.95 %+ quantiles.
(3) For every subject, estimate the subject-specific factor loadings $Z_{\ell,t}^{(j)}$ with the multilinear regression (3.4) that will be further used to classify risk attitude of the subject.

4 Simulation

Before implementing the proposed 3DIF method to real data, we perform a simulation study to investigate its performance under known data generating processes. Our primary interest is to see how much the 3DIF method will improve the detection accuracy of the risk related brain regions compared to the alternative 1-dimensional functional approach. Moreover, we study how robust is the region detection with respect to the size of the risk activation brain regions.

Our simulation studies are designed to properly reflect real data at hand. The fMRI signals are generated for a "brain" defined in the dimensions of $[1, 91] \times [9, 100] \times [11, 81]$. In previous literature five regions including PC, VLPFC, lOFC, aINS and DLPFC have been identified to be active under risk related tasks. In the first simulation study, we consider five regions that are contained in the literature documented places and specify each of them to a $3 \times 3 \times 3$ cube for a simple demonstration. In particular, PC is defined at location $[51, 53] \times [25, 27] \times [60, 62]$, VLPFC at $[27, 29] \times [89, 91] \times [38, 40]$, lOFC at $[54, 56] \times [97, 99] \times [30, 32]$,

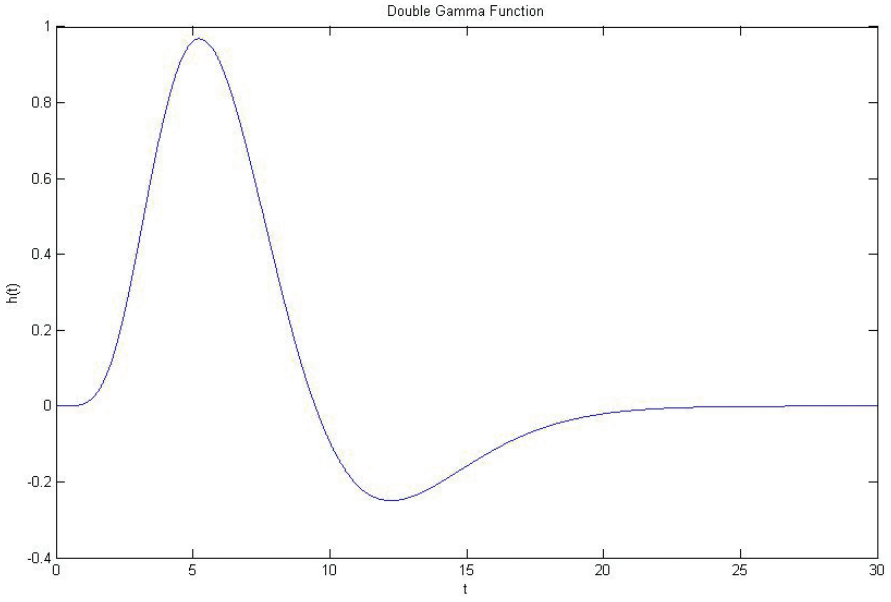


Figure 2: Visualization of the double gamma function.

aINS at $[63, 65] \times [75, 77] \times [37, 39]$, and DLPFC at $[66, 68] \times [77, 79] \times [53, 55]$. The regions are constant in the data generation.

Two kinds of factor loadings are considered: Gaussian distributed random loadings, and a more realistic situation by incorporating the haemodynamic response function (HRF) in the random loadings. The HRF is generated by a double gamma function (see [12, 15, 19, 53]):

$$h(t) = \left(\frac{t}{a_1 b_1}\right)^{a_1} e^{-\frac{t-a_1 b_1}{b_1}} - c\left(\frac{t}{a_2 b_2}\right)^{a_2} e^{-\frac{t-a_2 b_2}{b_2}},$$

where $a_1 = 6$, $a_2 = 12$, $b_1 = b_2 = 0.9$ and $c = 0.35$. Compared to the pure random factor loadings, the HRF scenario mimics the working process of the fMRI scanners, where HRF triggers brain activities. Figure 2 illustrates how the double gamma function reflects the haemodynamic response function (HRF).

Figure 3 gives an illustration of one simulated convolution of double gamma function and the generated factor loadings with HRF.

The 3D image signals are generated to represent brain signals recorded by the fMRI scanner during an RPID experiment:

$$f_t^{(\mathrm{NFL})}(x_1, x_2, x_3) = \sum_{\ell=1}^{5} Z_{\ell t}\xi_\ell(x_1, x_2, x_3) + \varepsilon_t(x_1, x_2, x_3),$$

$$f_t^{(\mathrm{HRF})}(x_1, x_2, x_3) = \sum_{\ell=1}^{5} \{Z_{\ell t} + h(t)\}\xi_\ell(x_1, x_2, x_3) + \varepsilon_t(x_1, x_2, x_3),$$

where NFL refers to the scenario with only normal random factor loadings, while HRF incorporates the impact of HRF in the fMRI signals. The five functional factors $\xi_\ell(x_1, x_2, x_3)$ have been defined in the locations (x_1, x_2, x_3) as mentioned before and are constant over time. The factor loading $Z_{\ell t}$ corresponds to the ℓ-th spatial factor at time point $t = 1, \ldots, 1\,000$. In both the NFL and HRF scenario, the factor loadings are Gaussian distributed with mean zero and standard deviations of 7.6, 5.8, 5.2, 1.8, and 1.7 respectively learned from the real data. The random noise $\varepsilon_t(x_1, x_2, x_3)$ is standard normal distributed and independent from each other. Each generation is repeated 100 times.

We implement two methods to identify the common spatial factors: 3DIF and FSVD proposed by [54]. Both methods handle continuous functional data, however 3DIF directly analyze the fMRI signals in 3D space while FSVD is only applicable for 1D functional data though the latter employs the singular value decomposition (SVD) approach to achieve better estimation feasibility and accuracy. In the simulation study, we chose

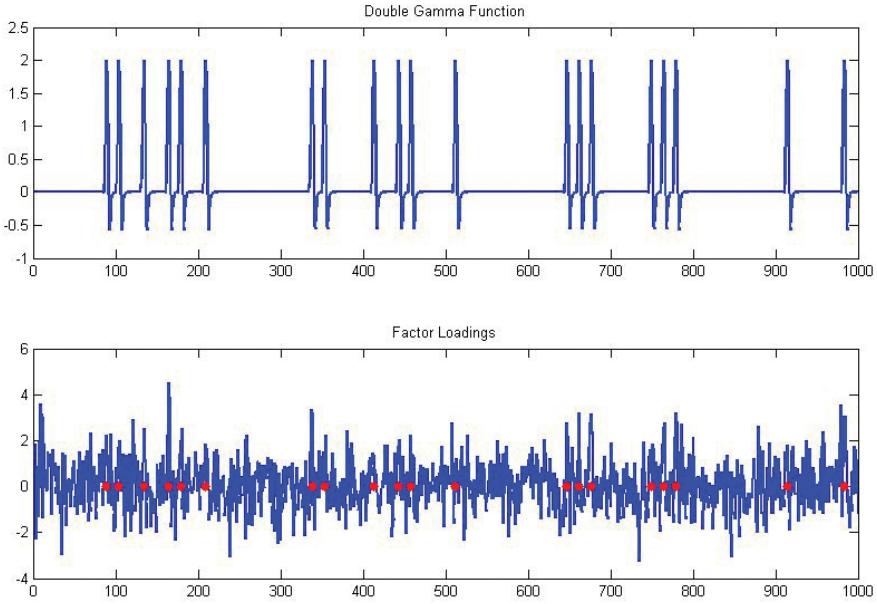


Figure 3: Simulated factor loadings. On top is the double gamma function. The bottom is the simulated factor loadings, which are the sum of the double gamma function and the normal random loadings. The red dots highlight time points when the stimulus are triggered.

$K = 16$ in each direction leading to $K^3 = 4096$ basis functions to utilize the largest computational power for each direction. It is worth noting that the designed risk related regions are only used in the fMRI data generation and will not be utilized in the following decomposition and factor computation. Instead, they are retained to evaluate the detection accuracy. In both methods, the active regions are defined as the trimmed spatial functional factors over the 99.999 % quantile and below the 0.001 % quantile.

As an illustration, Figure 4 displays one active region lOFC associated with evaluating and contrasting different option choices [45]. From top to bottom, one observes the generated (true) region, the identified regions by the 3DIF method and the FSVD approach. The active regions are highlighted as bright areas. Both methods detect the region, however 3DIF performs better in several aspects. In the NLF case, 3DIF explains more variation for the fMRI signals than FSVD, i.e. 56.3 % against 55.2 %, see Table 1. The variance explained increases when the number of factor increases. Moreover, 3DIF provides more clear-cut results, i.e. if the identified spacial factor corresponds to only one actual region, and simultaneously has less mis-detection, i.e. by wrongly identifying non-active regions. See Table 2 for the average percentage of the true regions detected by each estimated functional factor. More than 60 % of the estimated functional factors correspond to exactly one region in 3DIF. The value drops to 43.33 % in FSVD. As for mis-detection, 3DIF mistakenly detects 28 % and FSVD has more at 36.83 %. More importantly, 3DIF provides contiguous regions instead of discrete voxels thanks to its mathematical properties, see the contour plot of lOFC in Figure 5. On the other hand, FSVD identifies discrete voxels, due to the adoption of SVD in the discrete space, which improves estimation efficiency but at cost of contiguity. The relative good performance applies to the HRF scenario, too. While 3DIF explains 69.5 % variation, FSVD reaches to 55.9 %. When using 3DIF, 70 % of the detected risk regions correspond to exactly one active region, 23.33 % are mis-detected and less than 7 % are mixture of risk regions. The alternative FSVD method has only 54 % of one-to-one match, more than 30 % mis-detection and 15 % of mixture. Again, 3DIF accurately and reasonably detects a contiguous region, while the FSVD gives discrete voxels.

Now we repeat the above two experiments with different designs on the active regions to investigate the robustness of 3DIF. In particular, the five active regions are generated with varying sizes to reflect a more realistic situation. Following the study of [33] on the size of identified brain regions, our spatial moderate assumptions state that the spatial factors are active at location $[51, 54] \times [25, 28] \times [60, 63]$ for Parietal Cortex (64 voxels), $[27, 29] \times [88, 91] \times [38, 41]$ for VLPFC (48 voxels), $[52, 59] \times [92, 99] \times [28, 35]$ for lOFC (512 voxels), $[62, 66] \times [74, 78] \times [37, 39]$ for aINS (75 voxels), and $[64, 70] \times [73, 79] \times [51, 57]$ for

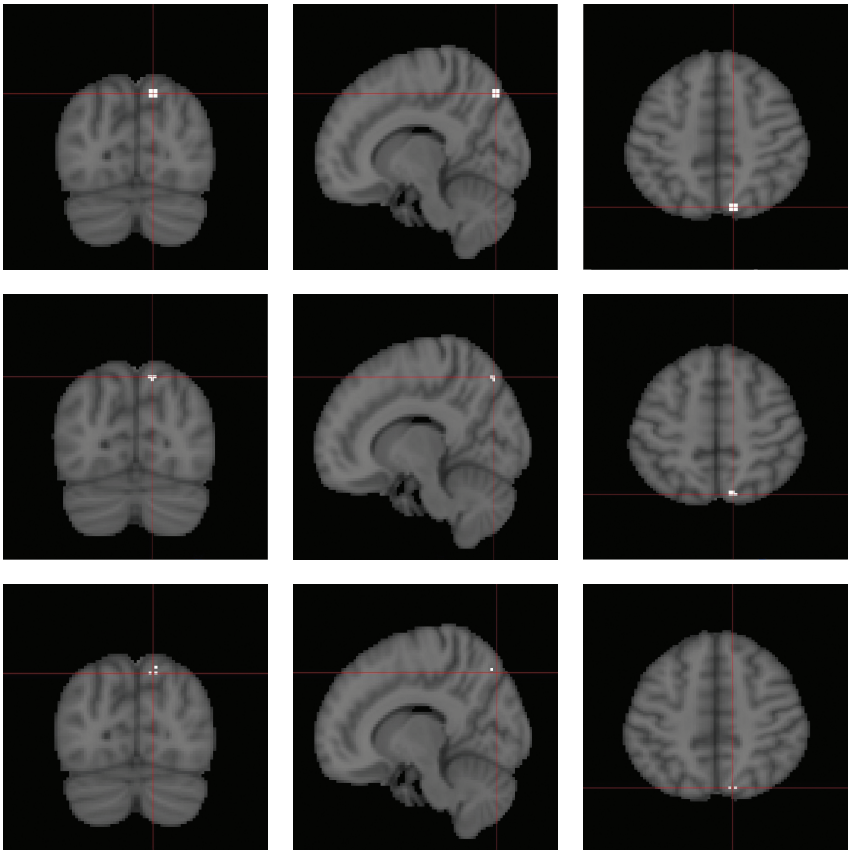


Figure 4: Functional factors on lOFC. From top to bottom are the generated (true) region, the estimated region with 3DIF and the estimated region with the FSVD method.

	Factor						
	1	**2**	**3**	**4**	**5**	**6**	**Total**
NFL: 3DIF	24.2 %	4.5 %	4.2 %	9.9 %	1.7 %	11.7 %	56.3 %
NFL: FSVD	19.2 %	0.7 %	1.6 %	21.5 %	4.8 %	7.4 %	55.2 %
HRF: 3DIF	25.9 %	4.9 %	7.0 %	16.2 %	5.7 %	9.8 %	69.5 %
HRF: FSVD	20.5 %	2.2 %	3.3 %	17.8 %	1.2 %	10.7 %	55.9 %

Table 1: Variance explained by different number of spatial factors for NFL with Gaussian random factor loadings and HRF incorporating HRF in the factor loadings. Two methods have been implemented: 3DIF and FSVD.

	Regions			
	0	**1**	**2**	**≥ 3**
NFL: 3DIF	28.00 %	60.67 %	11.33 %	0.00 %
NFL: FSVD	36.83 %	43.33 %	19.50 %	0.33 %
HRF: 3DIF	23.33 %	70.00 %	6.67 %	0.00 %
HRF: FSVD	31.33 %	54.00 %	14.67 %	0.00 %

Table 2: Average percentage of the estimated functional factors that detect the true regions; "0 region" means no active region and hence a nonzero values indicates mis-detection.

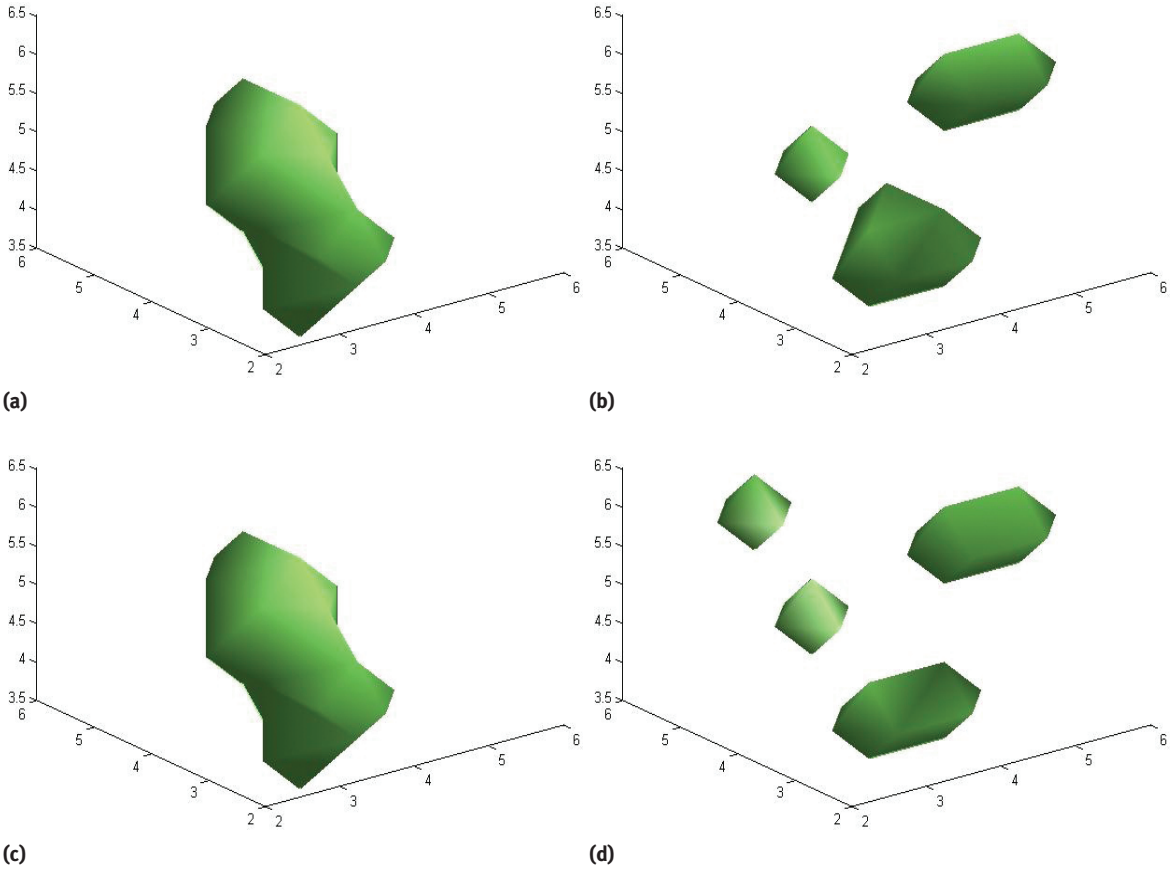


Figure 5: Contour plot of the estimated active region lOFC in NFL (top) and HRF (bottom) cases. On the left is the estimated region with 3DIF and on the right is the estimated region with FSVD.

	Regions			
	0	**1**	**2**	**≥ 3**
NFL: 3DIF	27.00 %	62.67 %	10.33 %	0.00 %
NFL: FSVD	32.17 %	52.33 %	15.50 %	0.00 %
HRF: 3DIF	18.50 %	79.67 %	1.83 %	0.00 %
HRF: FSVD	27.67 %	61.33 %	11.00 %	0.00 %

Table 3: Robust: average percentage of the estimated functional factors that detect the true regions; "0 region" means no active region and hence a nonzero values indicates mis-detection.

DLPFC (343 voxels). The factor loadings and the noise level remain the same as in the previous experiments. Both normal and HRF factor loadings are considered. Each data generation is repeated 100 times.

We still implement the 3DIF and FSVD methods to the generated fMRI data. As the average number of voxels now is about eight times of that in the previous simulations, the active regions are trimmed at extreme quantiles. Results evidence a stable performance. Again, 3DIF provides better identification, see Table 3 for the average percentage of the true regions detected by each estimated factor. In the NFL case, 62.67 % of the estimated functional factors are associated with exactly one region, 27 % are mis-detected and 10.33 % are mixed. On the contrary, the alternative method performs worse with less one-to-one match at 52.33 %, more mis-detection at 32.17 % and mixture at 15.5 %. In the HRF case, 3DIF still outperforms the alternative with 79.67 % one-to-one match, 18.50 % mis-detection and 1.83 % mixture, compared to 61.33 %, 27.67 % and 11.00 % by FSVD. Similarly, the 3DIF method provides realistic contiguous regions, while the alternative FSVD detects discrete voxels, see Figure 6 for the contour plot of the risk region lOFC as illustration.

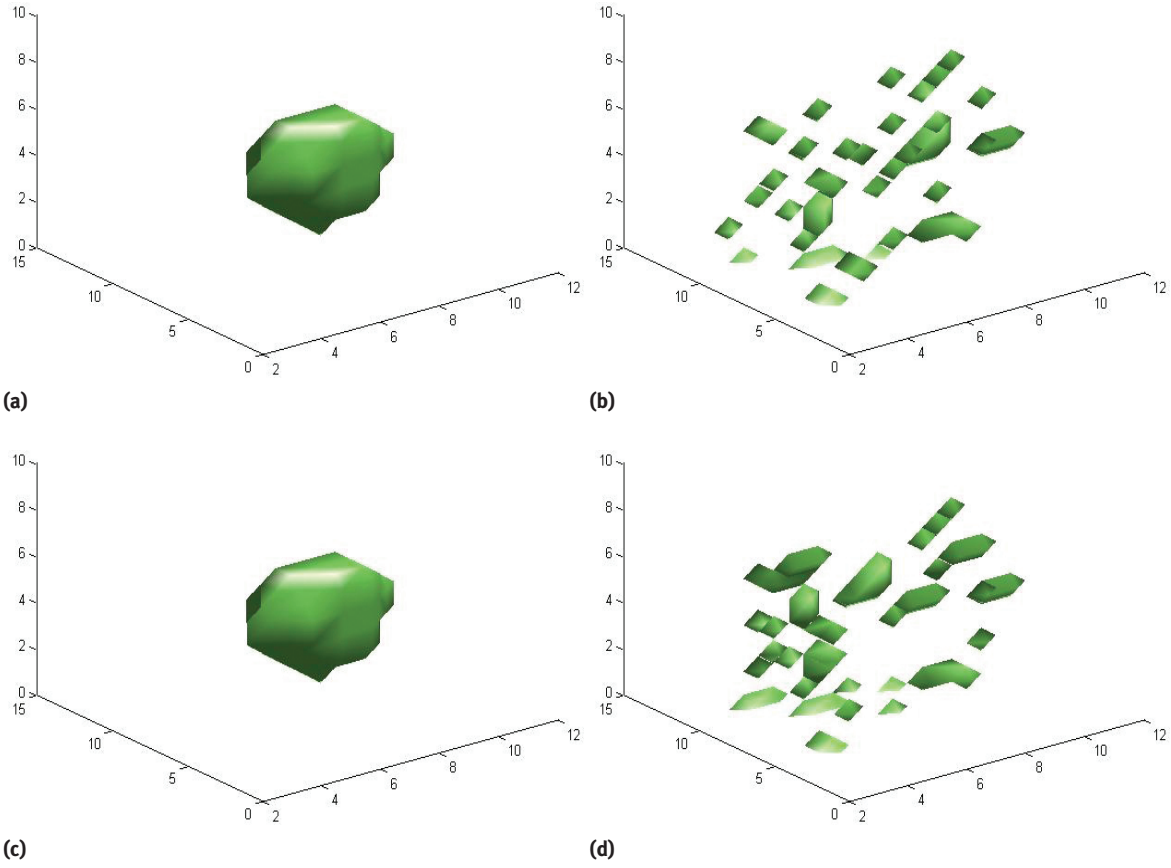


Figure 6: Robust: contour plot of the active region on lOFC. The left column is the estimated region in 3DIF and the right column is the estimated region with FSVD method. The top row is the result for NFL with normal factor loadings and on the bottom is the result for HRF with HRF incorporated in factor loadings.

The simulation study shows that the proposed 3DIF outperforms the alternative functional approach, with better quality of risk related regions detected. The relative good performance is stable for different scenarios with various parameters.

5 Empirical results

We implement the proposed 3DIF method to the fMRI signals data collected in the RPID experiment as described in Section 2, which mimics real-life investment decisions by providing subjects with return streams of investments. We assume that all subjects exhibit homogenous brain structure. In other words, the spatial maps are common for all, while the individual differences are represented by the subject specific scores. We report the detected common risk related regions and compare with several alternative methods. We classify subjects' risk perception based on the extracted subject specific signals, i.e. signature scores, and evaluate the risk classification accuracy with the help of psychological risk-return (PRR) model.

5.1 Computational time

The analyzed fMRI data are high dimensional ($91 \times 109 \times 91 \times 1\,360$ scans = 1,227,575,440) and require large memory (17×1.3 GB). The 3DIF method is implemented on twelve cores ProLiant BL680c G7 server equipped with Intel(R) Xeon(R) CPU E7-4860@2.27 GHz processors and 252 GB memory loading. The main

computation time is spent on computing the tensor integral $\boldsymbol{W} = \iiint \boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{\phi}^\top(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$, which exponentially increases in the number of knots K. Though a large number of knots provides better fit, it extends the computational time. Van Bömmel, Song, Majer, Mohr, Heekeren and Härdle [47] choose the basis function with fourteen knots in the x- and y-axis and twelve knots in the z-axis to balance accuracy and computational time. In our study, we increase the number of knots $K = 16$ in each direction leading to $K^3 = 4096$ basis functions, to further improve the estimation accuracy by utilizing larger computational power. The computation of the triple integral $\boldsymbol{W}$ costs 48 hours. It is worth noting that the value of the triple integral only depends on the B-spline basis functions and hence can be used for other fMRI data analysis. With the value of $\boldsymbol{W}$, the computation of 3DIF only needs 4 hours to complete.

5.2 Alternative methods

For comparison, two alternative methods have been implemented on the same data. Mohr, Biele, Krugel, Li and Heekeren [33] conducted the general linear model (**GLM**) with six design factors on the individual fMRI data. Van Bömmel, Song, Majer, Mohr, Heekeren and Härdle [47] proposed a panel version of the dynamic semiparametric factor model (**PDSFM**) to reanalyze the data. See Section 2 for details of their findings.

In addition, we consider three more methods that have previously been proposed in literature. We implement them to analyze the same data, including singular value decomposition (SVD) – a multivariate statistical technique – in a discrete framework, and two functional data analysis methods functional principal component analysis (FPCA) and functional SVD (FSVD) in a continuous but 1-dimensional space.

SVD: Denote the vectorized fMRI signal data as $Y = [Y_1, Y_2, \ldots, Y_N]$ that has $p \times N$ dimensions with $p = 91 \times 109 \times 91$ and $N = 1360$ in our study, SVD decomposes the discrete data averaged over subjects and constructs common spatial factors of risk-related brain regions $Y = \Gamma\Lambda^{\frac{1}{2}}U^\top$, where Γ is a $p \times N$ orthonormal matrix, Λ is a diagonal matrix and U is an $N \times N$ orthogonal matrix. The ℓ-th spatial factor is constructed with the ℓ-th column of Γ. Compared to the classic principal component analysis (PCA), SVD is computationally efficient and feasible with reduced dimensionality, i.e. decomposing a $p \times N$ sample matrix instead of a $p \times p$ covariance matrix given that $p \gg N$, when dealing with high-dimensional data. It however ignores contiguity nature of the fMRI signals, which leads to discontinued active regions.

FPCA and FSVD: The FPCA method estimates eigenfunctions in a functional framework. Similar to the proposed 3DIF method, the vectorized data is smoothed but using 1D basis functions and one performs eigendecomposition for the covariance operator. Denote the covariance operator by V we have $V\xi = \lambda\xi$, where ξ represents the eigenfunction corresponding the eigenvalue λ, see [39, 40]. The FPCA approach, though guarantees the contiguity of risk related brain regions, is subject to the curse of dimensionality. Zipunnikov, Caffo, Yousem, Davatzikos, Schwartz and Crainiceanu [54] proposed FSVD, which implements SVD to the smoothed functional data instead of the discrete raw data to balance the tradeoff between high dimensionality and computational efficiency. Nevertheless, the two functional data analysis methods requests pre-processing vectorization, which may misrepresent the raw spatial structure of the fMRI data.

5.3 Risk related regions $\widehat{\boldsymbol{\xi}}_\ell$

The 3D Image FPCA (3DIF) technique is utilized to capture the fundamental spatial maps under risk decisions. We identify the common spatial factors and use them to represent the brain regions with significant activity during the RPID experiment. One question remains on how to choose the number of spatial factors, denoted by L. The larger the number of spatial factors, the better the in-sample accuracy of the fitted model. On the other hand, too large L leads to over-fitting and poor out-of-sample performance. The selection of the number of factors may rest on the explained variation for different model specification. Table 4 presents the explained variance averaged over the seventeen subjects for different number of factors. It shows that 86 % variation in the data is attributed to the first spatial factor when using 3DIF, which can be interpreted as the typical

	L				
	1	**2**	**4**	**6**	**20**
3DIF	86.03 %	88.93 %	90.05 %	92.78 %	94.34 %
FSVD	96.50 %	96.57 %	96.65 %	96.74 %	97.07 %
FPCA	70.06 %	81.62 %	87.85 %	92.82 %	95.27 %
SVD	96.67 %	96.73 %	96.80 %	96.89 %	97.21 %

Table 4: Explained variance by different number of spatial factors.

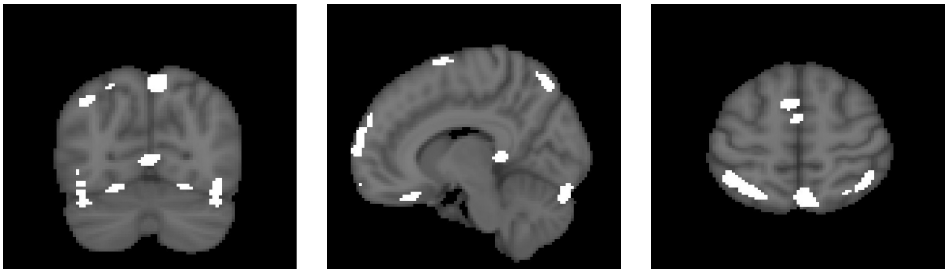
brain activity during the RPID experiment. Alternatively, the dominant component explains 96.50 % variation in FSVD, 70 % in FPCA and 98.67 % in SVD. Though numerically important, the first spatial factor has less psychological meaning and is irrelevant to any important risk related regions documented in literature. On the contrary, the inclusion of subsequent factors allows more useful information captured and simultaneously enables the detection of important risk related regions. For example, aINS is in modest size relative to visual or audial cortex but highly relevant to risk perception and investment decisions. Thus, $L = 20$ is chosen in our study. In this case, 94 % of variation is explained by the 3DIF method, which is lower than the alternatives. However, it is worth mentioning that higher variance is explained by the 3DIF spatial factor associated with important risk related regions. For example, the 3DIF factor for lOFC ($\widehat{\xi}_5$) explains 2.73 % (the difference between 92.78 % for $L = 6$ and 90.05 % for $L = 4$), while FSVD ($\widehat{\xi}_5$) and SVD ($\widehat{\xi}_5$) both contribute 0.09 % and FPCA ($\widehat{\xi}_3$) provide 6.23 %. We will continue the performance comparison of the data-driven methods in the risk classification analysis.

Figure 7 displays the identified risk related brain regions by using the proposed 3DIF method, the alternative 1D functional data analysis methods FSVD and FPCA, and the multivariate technique SVD. All detect the risk related brain regions including parietal cortex (PC), lateral orbifrontal cortex (lOFC) and ventrolateral prefrontal cortex (VLPFC). The three regions have been documented in literature and also by [33] analyzing the same data with GLM. However only the proposed 3DIF method successfully finds anterior insula (aINS) that is associated with value processing, risk and uncertainty. Moreover, the 3DIF method detects the activation of medial orbifrontal cortex (mOFC) as documented in [47] when analyzing the same data using PDSFM. The mOFC has been interpreted to be related to evaluation and contrast of various choices [45]. The FPCA method provides over-smoothed regions, though continuous, due to the extremely high dimensionality larger than 220,000 after vectorization. Table 5 summarizes the region detection for the same data by various methods. The proposed 3DIF method and the GLM [33] both identified five regions, where four of them are consistent. The alternative FSVD, FPCA and SVD found three regions and the PDSFM [47] obtained two.

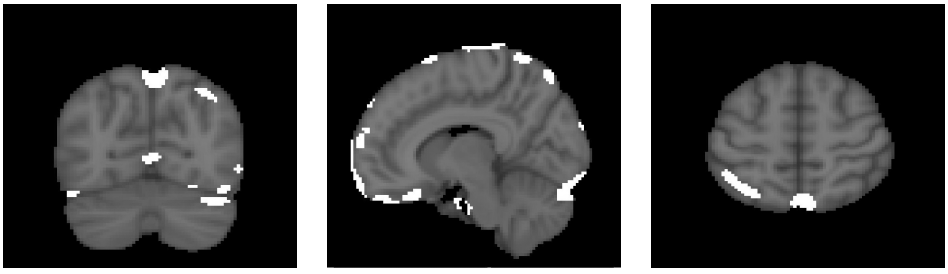
Figure 8 displays details of the detected regions by 3DIF. The relevant spatial factors are $\widehat{\xi}_\ell(x_1, x_2, x_3)$ with $\ell = 3, 4, 5, 12, 18, 19$. In particular, $\widehat{\xi}_3$ and $\widehat{\xi}_{12}$ are located in PC and attributed to risk related processes and selective attention (see [6, 41]); $\widehat{\xi}_4$ is related to the VLPFC region that stands for value processing. The regions mOFC and lOFC picked up by $\widehat{\xi}_5$ that are associated with evaluating and contrasting of different choice options [45]. The aINS region is captured by $\widehat{\xi}_{18}$ and related to risk and uncertainty [18], and the DLPFC area is highlighted by $\widehat{\xi}_{19}$. Figures 9–11 display the detected risk related brain regions by the alternative approaches. The identified regions of lOFC and VLPFC in Figures 9–11are similar due to the nearby coordinates of the regions. The center coordinates of the identified lOFC is (61, 94, 31) and of the VLPFC is (30, 94, 36).

	PC	**VLPFC**	**lOFC**	**aINS**	**DLPFC**	**mOFC**	**MPFC**
3DIF	✓	✓	✓	✓	✓		
GLM	✓	✓		✓	✓		✓
PDSFM	✓					✓	
FSVD	✓	✓	✓				
FPCA	✓	✓	✓				
SVD	✓	✓	✓				

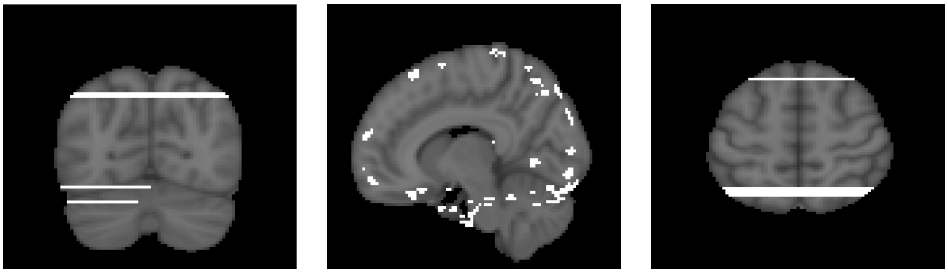
Table 5: Detected risk related brain regions for the same fMRI data of the RPID experiments in [33].



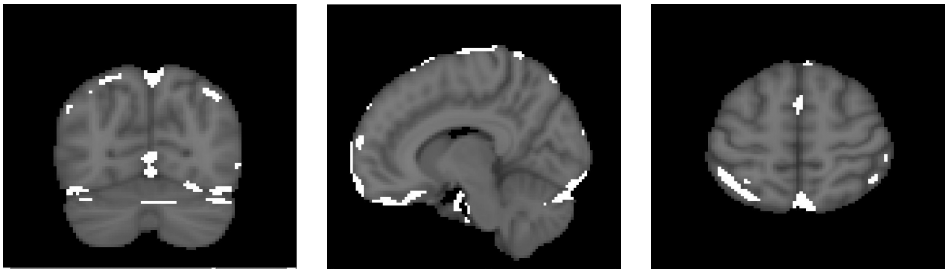
(a) 3DIF.



(b) FSVD.

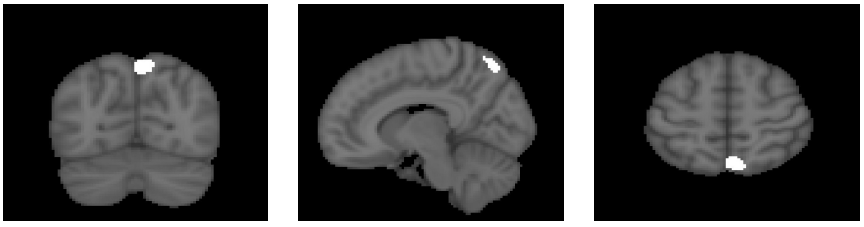


(c) FPCA.

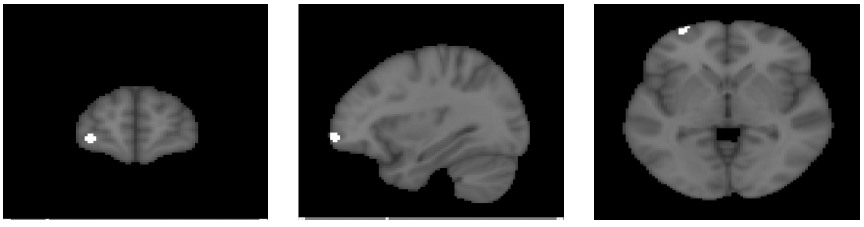


(d) SVD.

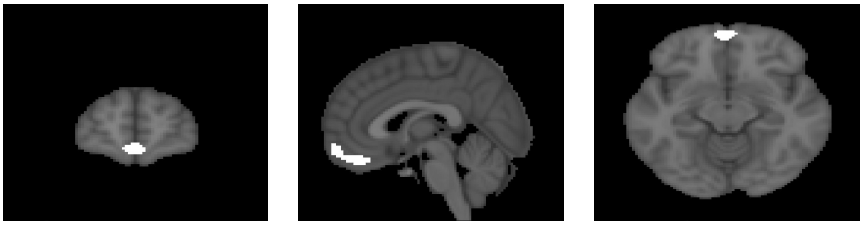
Figure 7: Detected risk-related brain regions by the first twenty eigenfunctions using (a) the 3DIF and alternative methods including (b) FSVD, (c) FPCA and (d) SVD.



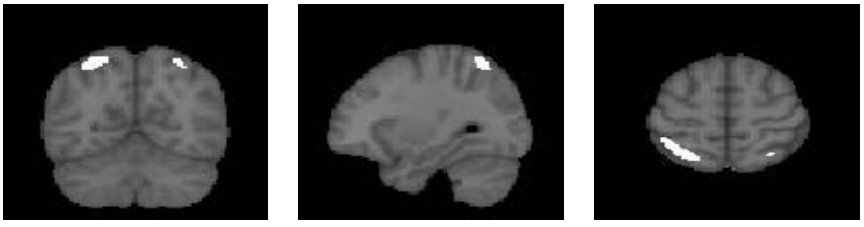
(a) Parietal Cortex.



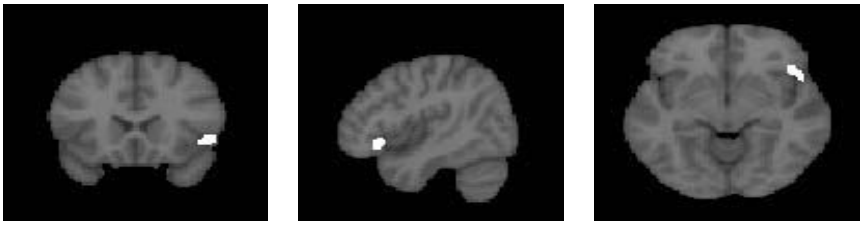
(b) VLPFC.



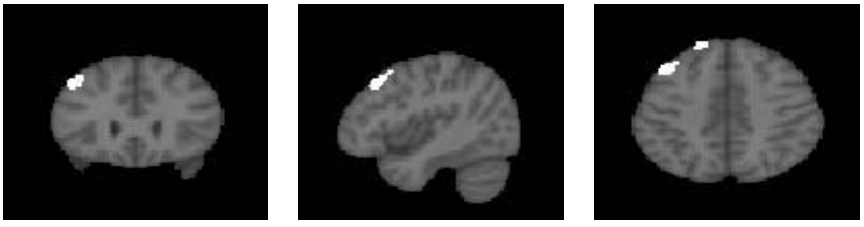
(c) lOFC.



(d) Parietal Cortex.

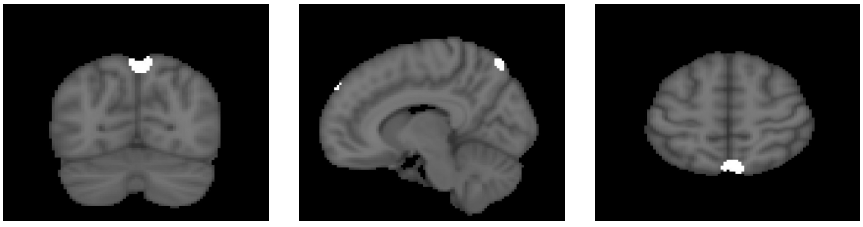


(e) aINS.

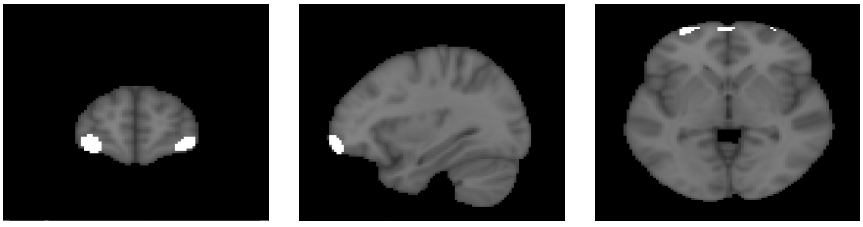


(f) DLPFC

Figure 8: 3DIF: Selected identified risk related regions $\widehat{\xi}_\ell$, $\ell = 3, 4, 5, 12, 18, 19$. (a) Estimated function $\widehat{\xi}_{12}$ in Parietal Cortex; (b) $\widehat{\xi}_4$ in VLPFC; (c) $\widehat{\xi}_5$ in lOFC; (d) $\widehat{\xi}_3$ in Parietal Cortex; (e) $\widehat{\xi}_{18}$ in aINS; (f) $\widehat{\xi}_{19}$ in DLPFC.

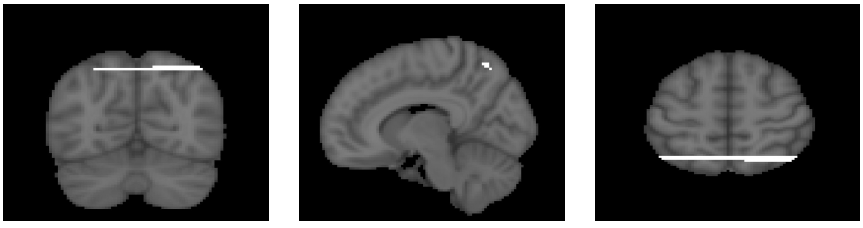


(a) Parietal Cortex.

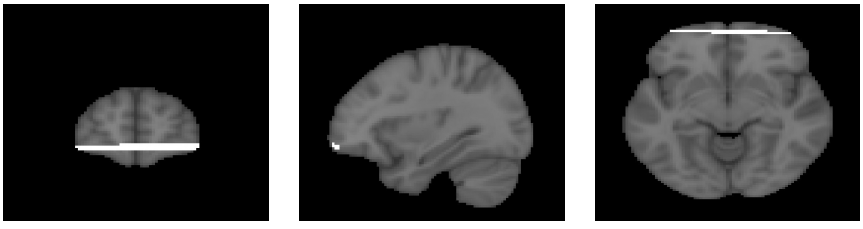


(b) VLPFC & lOFC.

Figure 9: FSVD: Selected identified risk related regions. (a) Estimated function $\widehat{\xi}_{10}$ in Parietal Cortex; (b) $\widehat{\xi}_5$ in VLPFC and lOFC.

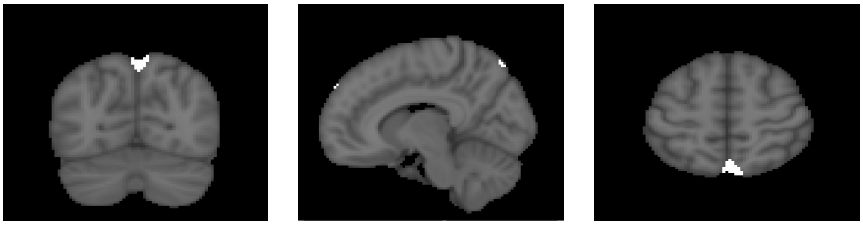


(a) Parietal Cortex.

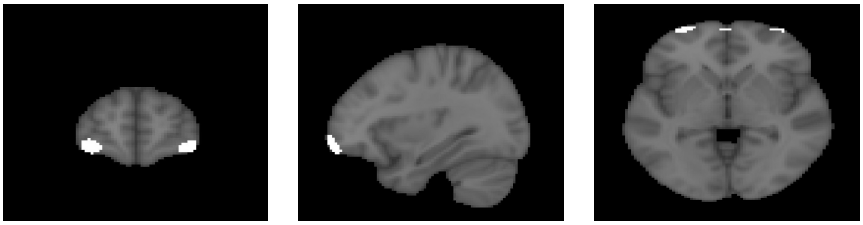


(b) VLPFC & lOFC.

Figure 10: FPCA: Selected identified risk related regions. (a) Estimated function $\widehat{\xi}_2$ in Parietal Cortex; (b) $\widehat{\xi}_3$ in VLPFC and lOFC.



(a) Parietal Cortex.



(b) VLPFC & lOFC.

Figure 11: SVD: Selected identified risk related regions by SVD. (a) Estimated function $\widehat{\xi}_{10}$ in Parietal Cortex; (b) $\widehat{\xi}_5$ in VLPFC and lOFC.

5.4 Subject specific signature scores $Z_{\ell,t}^{(j)}$

The dynamic behaviors of the individual brain activities are represented by the subject specific signature $Z_{\ell,t}^{(j)}$ with $j = 1, \ldots, 17$, $\ell = 1, \ldots, 20$, and $t = 1, \ldots, 1360$. Given the risk related regions common for all subjects, the individual risk perception and attitude during decision making under risk are reflected by the series of the activation. An interesting question is whether the extracted subject specific signature scores properly reflect the risk preference of individual. Among others, for the active brain regions that have been found to be related to risk and uncertainty, the respective signature scores are expected to carry explanatory power for the heterogeneity of individual risk preferences. Understanding those variations requires a careful investigation and is presented in the following risk classification study.

5.4.1 Risk attitudes

Mohr, Biele, Krugel, Li and Heekeren [33] quantify the risk preference of the seventeen subjects in the same experiment with the help of psychological risk-return (PRR) model

$$V_j(x) = \mu_j(x) - \beta_j \sigma_j(x),$$

where $V_j(x)$ is the value of investment x by subject j, $\mu_j(x)$ is the respective expected return, $\sigma_j(x)$ is the perceived risk, and β_j is a subject specific weight coefficient and reflects the risk attitude of subject j. Given the displayed streams of returns in the RPID experiment and the subjects' answers to the two tasks, i.e. subjective expected return and perceived risk, the risk weight β_j is estimated in a logistic regression framework. In total, seven subjects ($j = 2, 5, 6, 8, 10, 11, 17$) are categorized as weakly risk averse with the risk weight $\beta_j < 5$, and the remaining ten subjects are classified as strongly risk averse, with higher risk attitudes. The dichotomization and derived risk attitudes β_j are presented in Figure 12.

5.4.2 Risk classification

The aim of risk classification analysis is to investigate the possible relation between neural processes underlying investment decisions and subjects' risk preferences. A classification method is proposed to predict

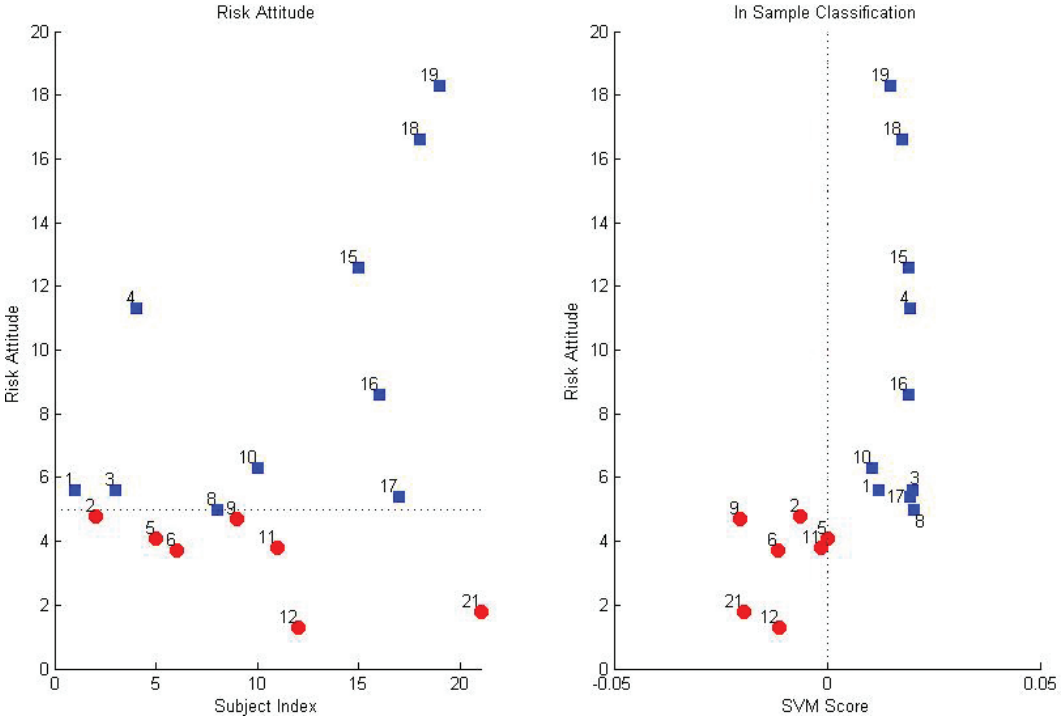


Figure 12: Risk attitudes and SVM scores of seventeen subjects. Subjects with risk attitude ≤ 5 are marked as red circles, otherwise as blue squares.

individual's risk attitude without any information on his or her decision behavior. Instead, the classification is performed solely on the extracted signature scores. The RPID consists of three types of tasks, we here only utilize the decision task, where subject chooses between risky investment return or sure fixed 5 % return, and thus his risk attitude contributes to the perceived value of the displayed return streams and plays a key role in the decision process. The other two tasks, i.e. subjective expected return and perceived risk, have been employed in the PRR model to provide a benchmark and will be used to verify the classification accuracy. Moreover, the analysis is performed for each subject based on six signature scores $Z^{(j)}_{\ell,t}$, $\ell = 3, 4, 5, 12, 18, 19$, of the active brain regions that have been found to be related to risk and uncertainty.

Each subject was exposed to 27 decision tasks and had to make a choice within the next 7 seconds in the RPID experiment. To investigate the brain reactions to the investment decision task of different groups being strongly/weakly risk averse, three consequent observations after the s-th stimulus at scan t_s are considered, covering the decision making period over 7.5 seconds. The three signature scores are demeaned by the score at the stimulus time point Z^{j}_{ℓ,t_s} to capture the peak of the HRF. We compute the average to stand for the average reaction to stimulus s

$$\overline{\Delta}\widehat{Z}^{(j)}_{\ell,t_s} = \frac{1}{3}\sum_{\tau=1}^{3}\widehat{Z}^{(j)}_{\ell,,t_s+\tau} - \widehat{Z}^{(j)}_{\ell,t_s}$$

and the standard deviation of the 27 average reactions as empirical characteristics of subject's risk preference. For each subject, six standard deviations are obtained and will be used in the risk classification analysis. For the alternative FSVD, FPCA and SVD methods, similar procedures are applied to extract the variables for risk classification.

Classification analysis is performed via support vector machines (SVM), see [7, 17]. Subjects are classified based on their six standard deviations of the average reactions to decision task. For the learning part, the strongly risk averse subjects are denoted with 1 and the weakly risk averse subjects with −1. The classification performance is validated by the estimated risk attitudes, see Section 5.4.1.

We first evaluate the in-sample predictive power of the 3DIF method on risk preferences. Figure 12 shows that the seventeen subjects were perfectly classified, with 100 % correction for both strongly and weakly risk

	Overall				Strong				Weak			
k	3DIF	SVD	FSVD	FPCA	3DIF	SVD	FSVD	FPCA	3DIF	SVD	FSVD	FPCA
1	**88 %**	76 %	76 %	76 %	**100 %**	100 %	100 %	90 %	**71 %**	43 %	43 %	57 %
2	**82 %**	76 %	76 %	76 %	**100 %**	100 %	100 %	89 %	55 %	43 %	43 %	**56 %**
3	**79 %**	75 %	75 %	73 %	98 %	**99 %**	**99 %**	87 %	53 %	42 %	42 %	**54 %**
4	**77 %**	74 %	73 %	72 %	95 %	**98 %**	95 %	85 %	51 %	39 %	41 %	**52 %**
5	**74 %**	71 %	70 %	69 %	92 %	**95 %**	91 %	83 %	**50 %**	37 %	39 %	49 %
6	**73 %**	67 %	66 %	66 %	**90 %**	90 %	86 %	81 %	**49 %**	35 %	37 %	46 %

Table 6: SVM classification rate in percentage points by leave-*k*-out for the 3DIF, SVD, FSVD and FPCA methods. The overall refers to the classification rates of all subjects, while strong and weak refer to the classification rates of strongly risk averse subjects and weakly risk averse subjects respectively.

averse groups. The in-sample classification however by utilizing all the information of subjects may involve over-fitting problem. We thus employ the leave-*k*-out cross validation and continue out-of-sample prediction. Samples are iteratively partitioned to two subsets, i.e. training with $N - k$ subjects and validation with k subjects. The prediction for validation is repeatedly performed based on different training sets. The accuracy measurements are averaged among all the predictions. The algorithm can be formulated as follows:

(1) Divide subjects into training set with $N - k$ people and test set with size of k.
(2) Apply the leave-*k*-out cross validation and find the optimal SVM parameters.
(3) Classify the test data.
(4) Repeat (1)–(3) for all different test sets.

Table 6 reports the classification rate (in percentage) by leave-*k*-out cross validation for $k = 1, \ldots, 6$. The classification rate is relatively stable, though it reduces slowly as k increases. The 3DIF method provides consistently better "overall classification" rate than the alternatives, with 73 %–88 % correction using the optimal SVM parameters. The classification accuracy is remarkably improved for the strongly risk averse subjects. The 3DIF and SVD methods are superior in terms of classification accuracy at 90 %–100 %, while 3DIF and FPCA perform better for weakly risk averse individuals at 49 %–71 %. As a comparison, van Bömmel, Song, Majer, Mohr, Heekeren and Härdle [47] have implemented leave-one-out procedure, i.e. $k = 1$, and reached 97 % for strongly risk-averse individuals and 75 % for weakly risk-averse individual. In summary, the analysis implies that the signature scores of the selected risk related regions carry explanatory power for subjects' risk attitudes derived from their choice in the RPID experiment. The risk preferences can be classified by the volatility (standard deviation) of the signature signals with an considerable accuracy. The proposed 3DIF method has consistent reasonable classification power compared to the alternatives.

6 Conclusion

Understanding how people make decisions among risky choices has attracted much attention of researchers in economics, psychology, and neuroscience. While economists evaluate individual's risk preference through mathematical modeling, neuroscientists answer the question by exploring the neural activities in brain. The existing literature has documented the brain regions of PCC, lOFC, mOFC, VLPFC, VMPFC and aNIS to be associated with decision making process under risk. Our study implements a model-free method to further investigate the links between active risk related brain region detection and individual's risk preference.

The proposed 3D Image FPCA (3DIF) methodology is directly applicable to the 3D image data. It avoids spatial information distortion during artificial vectorization or mapping and simultaneously analyzes brain data in the continuous functional domain. Thus, the anatomical brain structure is preserved and efficiently embraced in the estimation procedure. Moreover, it guarantees the contiguity of brain regions rather than discrete voxels. The 3DIF decomposes the fMRI BOLD signals into spatial factors, representing the common spatial maps for all subjects, and the heterogeneity of individual risk preference is explained by subject spe-

cific signature scores. The spatial factors capture the brain regions with the highest variability throughout experiment and consequently represent the activation pattern with a reduced number of factors. The representation precision is controlled by the number of factors L and even subtle effects can be detected. The signature scores mimic activation patterns on subject's risk attitude and correspond to the neural activity of a particular region of interest. As a result, the 3DIF addresses the key limitations of the GLM and the other conventional model-free methods such as PDSFM, FSVD, FPCA and SVD.

The performance is evidenced by our extensive simulation study, where in different setups, region detections and modeling performance were reasonably achieved. Furthermore, our technique outperforms the alternative competitor as the preservation of the spatial brain structure really pays off. In real data analysis, 3DIF detected five risk related regions, which is consistent to the study in [33]. The alternative methods on the other hand only identified limited risk related regions.

Investment decision may be described as a process of evaluating and contrasting of various choices with uncertain outcomes. In this framework the risk preferences are the crucial factor which affects the subjective value of investment. To improve our understanding of the underlying neural activities, we provided the statistical analysis of the extracted signature scores selected in the decision making context. The focus is on the variability in the HRF after the decision stimulus, captured by the score series. The standard deviations derived from the subject-specific responses served as an input in the SVM classifier. We perform both in-sample and out-of-sample risk classifications. In addition to perfect correction for in-sample, the 3DIF provides nice and stable performance for out-of-sample with leave-k-out cross validation, with the best overall classification rate at 73 %–88 %, the 90 %–100 % for strongly risk averse and 49 %–71 % for weakly risk averse. One can conclude that the 3DIF method exhibits better explanatory power for subjects' risk preferences than the alternatives.

Funding: This research was supported by the FRC grant and IDS grant at the National University of Singapore. The authors also acknowledge the support of the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk" and the International Research Training Group (IRTG) 1792 "High-Dimensional Non-Stationary Time Series".

References

[1] C. Amiez, J.-P. Joseph and E. Procyk, Reward encoding in the monkey anterior cingulate cortex, *Cerebral Cortex* **16** (2006), no. 7, 1040–1055.
[2] A. H. Andersen, D. M. Gash and M. J. Avison, Principal component analysis of the dynamic response measured by fMRI: A generalized linear systems framework, *Magn. Resonance Imag.* **17** (1999), no. 6, 795–815.
[3] R. B. Barsky, M. S. Kimball, F. T. Juster and M. D. Shapiro, Preference parameters and behavioral heterogeneity: An experimental approach in the health and retirement survey, Technical report, National Bureau of Economic Research, 1995.
[4] R. Baumgartner, L. Ryner, W. Richter, R. Summers, M. Jarmasz and R. Somorjai, Comparison of two exploratory data analysis methods for fMRI: Fuzzy clustering vs. principal component analysis, *Magn. Resonance Imag.* **18** (2000), no. 1, 89–94.
[5] R. M. W. J. Beetsma and P. C. Schotman, Measuring risk attitudes in a natural experiment: Data from the television game show lingo, *Econom. J.* **111** (2001), no. 474, 821–848.
[6] M. Behrmann, J. J. Geng and S. Shomstein, Parietal cortex and attention, *Current Opinion Neurobiol.* **14** (2004), 212–217.
[7] C. Cortes and V. Vapnik, The nature of statistical learning theory, *Mach. Learn.* **20** (2005), 273–297.
[8] H. D. Critchley, R. N. Melmed, E. Featherstone, C. J. Mathias and R. J. Dolan, Brain activity during biofeedback relaxation, *Brain* **124** (2001), no. 5, 1003–1012.
[9] D. Degras and M. A. Lindquist, A hierarchical model for simultaneous detection and estimation in multi-subject fMRI studies, *NeuroImage* **98** (2014), 61–72.
[10] D. Fetherstonhaugh, P. Slovic, S. Johnson and J. Friedrich, Insensitivity to the value of human life: A study of psychophysical numbing, *J. Risk Uncertainty* **14** (1997), no. 3, 283–300.
[11] K. J. Friston, A. P. Holmes, K. J. Worsley, J. B. Poline, C. Frith and R. S. J. Frackowiak, Statistical parametric maps in functional imaging: A general linear approach, *Human Brain Mapping* **2** (1995), 189–210.
[12] G. H. Glover, Deconvolution of impulse response in event-related bold fMRI, *Neuroimage* **9** (1999), no. 4, 416–429.

[13] G. H. Golub and C. Reinsch, Handbook Series Linear Algebra: Singular value decomposition and least squares solutions, *Numer. Math.* **14** (1970), no. 5, 403–420.
[14] J. Grinband, J. Hirsch and V. P. Ferrera, A neural representation of categorization uncertainty in the human brain, *Neuron* **49** (2006), no. 5, 757–763.
[15] J. Grinband, T. D. Wager, M. Lindquist, V. P. Ferrera and J. Hirsch, Detection of time-varying signals in event-related fMRI designs, *Neuroimage* **43** (2008), no. 3, 509–520.
[16] D. J. Hand and W. E. Henley, Statistical classification methods in consumer credit scoring: A review, *J. Roy. Statist. Soc. Ser. A* **160** (1997), no. 3, 523–541.
[17] W. K. Härdle and L. Simar, *Applied Multivariate Statistical Analysis*, 4th ed., Springer, Heidelberg, 2015.
[18] H. R. Heekeren, S. Marrett and L. G. Ungerleider, The neural systems that mediate human perceptual decision making, *Nat. Rev. Neurosci.* **9** (2008), 467–479.
[19] R. Heller, D. Stanley, D. Yekutieli, N. Rubin and Y. Benjamini, Cluster-based analysis of FMRI data, *NeuroImage* **33** (2006), no. 2, 599–608.
[20] S. A. Huettel, A. W. Song and G. McCarthy, Decisions under uncertainty: Probabilistic context influences activation of prefrontal and parietal cortices, *J. Neurosci.* **25** (2005), no. 13, 3304–3311.
[21] S. A. Huettel, C. J. Stowe, E. M. Gordon, B. T. Warner and M. L. Platt, Neural signatures of economic preferences for risk and ambiguity, *Neuron* **49** (2006), no. 5, 765–775.
[22] J. W. Kable and P. W. Glimcher, The neural correlates of subjective value during intertemporal choice, *Nature Neurosci.* **10** (2007), no. 12, 1625–1633.
[23] D. Kahneman and A. Tversky, Prospect theory: An analysis of decision under risk, *Econometrica* **47** (1979), 263–291.
[24] S. W. Kennerley, A. F. Dahmubed, A. H. Lara and J. D. Wallis, Neurons in the frontal lobe encode the value of multiple decision variables, *J. Cognitive Neurosci.* **21** (2009), no. 6, 1162–1178.
[25] B. Knutson, J. Taylor, M. Kaufman, R. Peterson and G. Glover, Distributed neural representation of expected value, *J. Neurosci.* **25** (2005), no. 19, 4806–4812.
[26] C. M. Kuhnen and B. Knutson, The neural basis of financial risk taking, *Neuron* **47** (2005), no. 5, 763–770.
[27] S.-H. Lai and M. Fang, A novel local pca-based method for detecting activation signals in fMRI, *Magn. Resonance Imag.* **17** (1999), no. 6, 827–836.
[28] G. F. Loewenstein, E. U. Weber, C. K. Hsee and N. Welch, Risk as feelings, *Psychol. Bull.* **127** (2001), no. 2, 267–286.
[29] C. J. Long, E. N. Brown, C. Triantafyllou, I. Aharon, L. L. Wald and V. Solo, Nonstationary noise estimation in functional MRI, *NeuroImage* **28** (2005), no. 4, 890–903.
[30] H. Markowitz, Portfolio selection, *J. Finance* **7** (1952), no. 1, 77–91.
[31] B. A. Mellers, Choice and the relative pleasure of consequences, *Psychol. Bull.* **126** (2000), no. 6, 910–924.
[32] P. N. C. Mohr, G. Biele and H. R. Heekeren, Neural processing of risk, *J. Neurosci.* **30** (2010), no. 19, 6613–6619.
[33] P. N. C. Mohr, G. Biele, L. K. Krugel, S.-C. Li and H. R. Heekeren, Neural foundations of risk-return trade-off in investment decisions, *NeuroImage* **49** (2010), 2556–2563.
[34] M. P. Paulus, C. Rogalsky, A. Simmons, J. S. Feinstein and M. B. Stein, Increased activation in the right insula during risk-taking decision making is related to harm avoidance and neuroticism, *NeuroImage* **19** (2003), no. 4, 1439–1448.
[35] H. Plassmann, J. O'Doherty and A. Rangel, Orbitofrontal cortex encodes willingness to pay in everyday economic transactions, *J. Neurosci.* **27** (2007), no. 37, 9984–9988.
[36] J. W. Pratt, Risk aversion in the small and in the large, *Econometrica* **44** (1964), 122–136.
[37] K. Preuschoff, P. Bossaerts and S. R. Quartz, Neural differentiation of expected reward and risk in human subcortical structures, *Neuron* **51** (2006), no. 3, 381–390.
[38] K. Preuschoff, S. R. Quartz and Peter Bossaerts, Human insula activation reflects risk prediction errors as well as risk, *J. Neurosci.* **28** (2008), no. 11, 2745–2752.
[39] C. Radhakrishna Rao, Some statistical methods for comparison of growth curves, *Biometrics* **14** (1958), no. 1, 1–17.
[40] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*, 2nd ed., Springer, New York, 2005.
[41] A. Rangel, C. Camerer and P. R. Montague, A framework for studying the neurobiology of value-based decision making, *Nat. Rev. Neurosci.* **9** (2008), 545–556.
[42] E. T. Rolls, C. McCabe and J. Redoute, Expected value, reward outcome, and temporal difference error representations in a probabilistic decision task, *Cerebral Cortex* **18** (2008), no. 3, 652–663.
[43] W. F. Sharpe, Capital asset prices: A theory of market equilibrium under conditions of risk, *J. Finance* **19** (1964), no. 3, 425–442.
[44] D. Schunk and C. Betsch, Explaining heterogeneity in utility functions by individual differences in decision modes, *J. Econom. Psychol.* **27** (2006), no. 3, 386–401.
[45] P. N. Tobler, J. P. O'Doherty, R. J. Dolan and W. Schultz, Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems, *J. Neurophysiol.* **97** (2007), 1621–1632.
[46] S. M. Tom, C. R. Fox, C. Trepel and R. A. Poldrack, The neural basis of loss aversion in decision-making under risk, *Science* **315** (2007), no. 5811, 515–518.

[47] A. van Bömmel, S. Song, P. Majer, P. N. C. Mohr, H. R. Heekeren and W. K. Härdle, Risk patterns and correlated brain activities. Multidimensional statistical analysis of fMRI data in economic decision making study, *Psychometrika* **79** (2014), no. 3, 489–514.
[48] T. Vincent, L. Risser and P. Ciuciu, Spatially adaptive mixture modeling for analysis of fMRI time series, *IEEE Trans. Medical Imag.* **29** (2010), no. 4, 1059–1074.
[49] R. Viviani, G. Gron and M. Spitzer, Functional principal component analysis of fMRI data, *Human Brain Mapping* **24** (2005), 109–129.
[50] J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, 1953.
[51] E. U. Weber, The utility of measuring and modeling perceived risk, in: *Choice, Decision, and Measurement: Essays in Honor of R. Duncan Luce*, Lawrence Erlbaum Associates, Mawah (1997), 45–56.
[52] E. U. Weber and E. J. Johnson, Decisions under uncertainty: Psychological, economic, and neuroeconomic explanations of risk preference, in: *Neuroeconomics: Decision Making and the Brain*, Academic Press, New York (2008), 127–144.
[53] K. J. Worsley, C. H. Liao, J. Aston, V. Petre, G. H. Duncan, F. Morales and A. C. Evans, A general statistical analysis for fMRI data, *Neuroimage* **15** (2002), no. 1, 1–15.
[54] V. Zipunnikov, B. Caffo, D. M. Yousem, C. Davatzikos, B. S. Schwartz and C. Crainiceanu, Functional principal component model for high-dimensional brain imaging, *NeuroImage* **58** (2011), no. 3, 772–784.

Journal of Banking and Finance 93 (2018) 21–32

Contents lists available at ScienceDirect

Journal of Banking and Finance

journal homepage: www.elsevier.com/locate/jbf

Downside risk and stock returns in the G7 countries: An empirical analysis of their long-run and short-run dynamics☆

Cathy Yi-Hsuan Chen[a,*], Thomas C. Chiang[b], Wolfgang Karl Härdle[a,c]

[a] Ladislaus von Bortkiewicz Chair of Statistics, C.A.S.E. – Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany
[b] LeBow College of Business, Drexel University, Gerri LeBow Hall, 3220 Market Street, Philadelphia, PA 19104, USA
[c] Sim Kee Boon Institute for Financial Economics, Singapore Management University, 50 Stamford Road, 178899 Singapore, Singapore

ARTICLE INFO

Article history:
Received 8 July 2015
Accepted 15 May 2018
Available online 26 May 2018

JEL classification:
G11
G12
G15
C24
F30

Keywords:
Downside risk
Value-at-risk
Long memory
Fractional integration
Risk-return
Market integration

ABSTRACT

Any risk-return tradeoff analysis in aggregate equity markets relies on appropriate measures of risk, in most studies based on (co-)variance relations. Consequently, in integrated global markets, country-specific expected return is priced with a world price of covariance risk. This study relates domestic excess stock returns to the world downside risk. Evidence shows that downside tail risk (as a multiplier of volatility) has long memory cointegration properties; hence, the underlying risk aversion behavior in an integrated market is associated with the conditional quantile ratio, the correlation of stock returns, and the cointegrating coefficient of downside risk. Our empirical results based on G7 countries indicate that investors are averse to downside risk, which via Cornish–Fisher expansions is related to higher moment risk and interpretable in a utility-based decision framework.

1. Introduction

Merton's (1973) intertemporal capital asset pricing model (ICAPM) postulates a positive risk-return tradeoff relation in aggregate equity markets. Empirical results, however, are inconclusive, being unable to reach strong support for this hypothesis. While a positive tradeoff phenomenon has been empirically documented by, for example, French et al. (1987), Scruggs (1998), Whitelaw (2000), Ghysels et al. (2005), Bali and Peng (2006), and Lundblad (2007), others, including Nelson (1991) and Glosten et al. (1993), find a weak or even a negative relation. Most importantly, the higher moment-return relation has been studied by Harvey and Siddique (2000), Jarrow and Zhao (2006), Harvey et al. (2010), and Lambert and Hübner (2013).

We re-examine the risk-return tradeoff relation in an international context but deviate from the conventional framework in two ways. Firstly, we use Value-at-Risk (VaR) as the risk measure. If investors exhibit loss aversion toward this type of downside risk, it should be reflected in higher moments of the return distribution as well. Indeed, in a Gaussian context, VaR would be tightly linked to variance, but the problem is that the interpretation of the risk aversion coefficient, later called lambda, then changes. In a non-Gaussian world, it is useful that VaR is linked to all higher moments, including skewness and kurtosis. If they matter, they would matter 'separately' in the expected return equation. The VaR measure compresses all of these into one number and is therefore an advantageous risk measure.

Secondly, following Harrison and Zhang (1999), Bandi and Perron (2008), and Ferson et al. (2013), we focus on long-run risk-return relations. In particular, given the persistence of VaR measures and their cointegration properties in an integrated global market, we find that the downside risk series for all G7 markets exhibit a long memory phenomenon. In addition, the cointegrated (long-run) downside risk underlines the significance of downside risk aversion. For countries with a higher return correlation and a higher world cointegrated downside risk, the investors' aversion is mainly based on the global risk. From this perspective, a local market is likely to be exposed to a higher global risk due to the co-

☆ The authors gratefully acknowledge financial support from the Deutsche Forschungsgemeinschaft through SFB 649 "Economic Risk" and IRTG 1792 "High Dimensional Non Stationary Time Series".

* Corresponding author.
E-mail addresses: chencath@hu-berlin.de (C.Y.-H. Chen), chiangtc@drexel.edu (T.C. Chiang), haerdle@hu-berlin.de (W.K. Härdle).

https://doi.org/10.1016/j.jbankfin.2018.05.012
0378-4266/

movement of downside risk in the long run. The predictive power of cointegrated relations between global and local downside risk on future return is confirmed in comparison with the pure short-run system.

A few explanations on VaR as risk measure seem to be appropriate. In a Gaussian context, VaR is simply a multiplier of volatility; the corresponding lambda turns out to be the product of lambda w.r.t variance risk and a fixed multiplier. In a non-Gaussian world, VaR can be expressed by higher-order moments (Cornish–Fisher expansion). Thus, each of them governs the tradeoff equation separately. The calculation of moments is tedious, however, and therefore the link between these and VaR provides a useful parsimonious representation of a risk measure. Although higher moments-based risk measures are consistent with the risk aversion of investors' behavior (Harvey and Siddique, 2000 Eq. (7)), they are more outlier sensitive when compared to VaR (Cont et al., 2010). Our approach of using VaR allows another look at nonstandard risk preferences in addition to earlier studies on an asymmetric attitude toward gains versus losses (Gul, 1991; Ang et al., 2005; Routledge and Zin, 2010).

Another important finding of our research is that in countries with a higher return correlation and a higher world cointegrated downside risk, the investors' aversion is mainly based on the global risk. The methodology behind this result is the FCVAR (fractionally cointegrated vector autoregression) model. Here, the fractional component is a necessity due to the presence of a long memory property in the VaR series. The employed quantitative FCVAR technique disentangles the long-run tradeoff effect from the short-run one. In a nutshell, the FCVAR model, in combination with the economic framework in Section 2, specifies the long-run and short-run downside risk aversion in a comprehensive way, theoretically and empirically.

As a result, a cointegration between any G7 country and the world market portfolio is confirmed. Such a cointegrated (long-run) downside risk underlines the significant and positive sign w.r.t downside risk aversion. In contrast, the short-run risk aversion is less satisfactory. Our results confirm the importance of the long-run effect in the downside risk-return relation. However, the leverage effect in the dynamic system has vanished.

The concerns about employing moments motivate us to innovate a calculation for extreme tail (e.g., 1%) VaR. We propose a nonparametric kernel density estimation procedure that smooths the empirical distribution of return given a small sample size. It yields the desired VaR level as an average of neighboring order statistics. Confidence intervals of these may be obtained by bootstrap. In this way, one can mitigate outlier distortion that may happen in moment-based investigation or in the approximation of VaR via Cornish–Fisher expansion. We believe this approach to be a novel but computer-intensive way to calculate the VaR risk measure, even in moderate data sample situations.

The remainder of this paper is organized as follows. Section 2 presents a theoretical framework for the downside risk-return relation in an integrated system, illustrating investors' risk aversion to the local downside risk and world downside risk. Section 3 describes the data, estimates the VaR as a measure of downside risk, and tests for the long memory process. Section 4 introduces the fractional integrated dynamic system used to empirically analyze the risk-return relation and reports the empirical evidence. Section 5 interrogates the robustness of the model, and Section 6 summarizes our conclusions.

2. Economic framework

2.1. Downside risk in a segmented market

An analysis of international equity markets shows that expected local returns can be seen either as a function of the conditional variance of the country returns or as a function of conditional covariance with a world market portfolio subject to the degree of market integration (see Bekaert and Harvey, 1995; Karolyi and Stulz, 1996). If the market is segmented, the expected return will simply depend upon the local risk. Expressing excess stock return for country i as a linear function of its conditional variance upon the information set at time $t-1$, I_{t-1} yields:

$$E_{t-1}(r_{i,t}) = \lambda_{i,t-1}\sigma^2_{i,t-1} \tag{1}$$

where $E_{t-1}(r_{it})$ is the expected excess return, $\sigma^2_{i,t-1}$ its variance, all conditional on the information set I_{t-1}. Note that lambda $\lambda_{i,t-1} = E_{t-1}(r_{it})/\sigma^2_{i,t-1}$ is a measure of the relative risk aversion of investors in country i, reflecting the ratio of risk premium to the conditional variance (French et al., 1987; Whitelaw, 1994; Scruggs, 1998). This measure is limited in its ability to reflect asymmetric risk aversion, however (Glosten et al., 1993); furthermore, it cannot effectively capture the investor's aversion to higher moments. If one takes a further step from variance to higher moments, one may look at asymmetric elements of risk. From a statistical perspective, a single outlier in the left tail can cause skewness to become negative, whereas an outlier in the right tail can unduly increase the skewness coefficient. This motivated Cont et al. (2010) to propose nonparametric VaR as a robust risk measure. Note that VaR not only provides information about the attributes of investor risk aversion but also mitigates any concern about outliers.

To provide more insight into this higher moment-VaR relation we employ the Cornish-Fisher expansion (Cornish and Fisher, 1937) to link the α-quantile, q_α, of the probability distribution of standardized return to its corresponding skewness, S, and excess kurtosis, k, as:

$$q_\alpha = z_\alpha + \left(z_\alpha^2 - 1\right)\frac{S}{6} + \left(z_\alpha^3 - 3z_\alpha\right)\frac{k}{24} - \left(2z_\alpha^3 - 5z_\alpha\right)\frac{S^2}{36} \tag{2}$$

where z_α is the α-quantile value of a standard normal distribution. V_α, the α-percentile of VaR, is simply the product of q_α and its standard deviation, σ—that is, $V_\alpha = q_\alpha\sigma$. By simple algebra, Eq. (2) with $V_\alpha = q_\alpha\sigma$ can be written as:

$$V_\alpha = \sum\nolimits_{j=2}^{6} w_j \frac{R^j}{\sigma^{j-1}}$$

where R is the demeaned return and w_j is the weight that corresponds to the jth standardized moment, subject to the choice of α-level. Hence, this VaR expansion is a weighted sum of standardized moments and a parsimonious representation since it compresses all of them into one number.

Given I_{t-1} and equity i, the above statistics can be estimated by the daily returns within the month $t-1$:

$$V_{i,t-1} = q_{i,t-1}\sigma_{i,t-1} \tag{3}$$

where $V_{i,t-1}$ is the conditional VaR at month $t-1$ and $q_{i,t-1}$ is the conditional quantile at month $t-1$.[1] Therefore, $\sigma^2_{i,t-1} = V^2_{i,t-1}/q^2_{i,t-1}$. Substituting Eq. (3) into Eq. (1) yields Eq. (4):

$$E_{t-1}(r_{it}) = \lambda^*_{i,t-1}V^2_{i,t-1} \tag{4}$$

where

$$\lambda^*_{i,t-1} = \lambda_{i,t-1}\frac{1}{q^2_{i,t-1}} = \frac{E_{t-1}(r_{it})}{\sigma^2_{i,t-1}} \times \frac{1}{q^2_{i,t-1}} = \frac{E_{t-1}(r_{it})}{V^2_{i,t-1}}$$

$\lambda^*_{i,t-1}$ is the measure of relative downside risk aversion in country i. Eq. (4) posits a positive relationship between expected return and downside risk if $\lambda^*_{i,t-1} > 0$. Invoking expected utility theory (Bali et al., 2009) concludes that agents are averse to variance

[1] α in the subscript has been suppressed.

risk—i.e., $\lambda_{i,t-1} > 0$. Hence, $\lambda^*_{i,t-1}$, as the product of $\lambda_{i,t-1}$and $\frac{1}{q^2_{i,t-1}}$, should be positive as well.

By differentiating Eq. (2) with respect to skewness conditional on $S < 0$, and given small enough α —e.g., $\alpha = 1\%$—we obtain $\frac{\partial q_i}{\partial S_i} > 0$ and $\frac{\partial V_i}{\partial S_i} > 0$; analogous calculation applied to k_i yields $\frac{\partial q_i}{\partial k_i} > 0$ and $\frac{\partial V_i}{\partial k_i} > 0$. These derivations imply that VaR increases with a negative skewness and with a positive excess kurtosis at interesting tail levels. Given σ_{it}, and using a chain rule, yields $\frac{\partial r_i}{\partial k_i} = \frac{\partial r_i}{\partial V_i^2} \times \frac{\partial V_i^2}{\partial k_i} > 0$, where $\frac{\partial r_i}{\partial V_i^2} > 0$ corresponds to a positive local relative risk aversion to the downside risk, say λ^*_i. Note that $\frac{\partial q_i^2}{\partial k_i} = 0.11$ at, for example, $\alpha = 1\%$ derived from Eq. (2), and the second moment σ_i does not demonstrate a relationship with respect to the fourth moment, we obtain a positive value—that is, $\frac{\partial V_i^2}{\partial k_i} = \frac{\sigma_i^2 \partial q_i^2}{\partial k_i} = 0.11 \times \sigma_i^2$ —from this differentiation. In summary, the VaR is monotonically increases with kurtosis and negative skewness.

By the same token, a value-weighted expected world portfolio return, $E_{t-1}(r_{wt})$, in relation to its downside risk-return can be written as:

$$E_{t-1}(r_{w,t}) = \lambda^*_{w,t-1} V^2_{w,t-1} \tag{5}$$

where $V^2_{w,t-1}$ is the conditional downside risk of the world portfolio and $\lambda^*_{w,t-1} = \frac{E_{t-1}(r_{w,t})}{V^2_{w,t-1}}$ is the measure of relative risk aversion of the world downside risk.

2.2. Downside risk in an integrated market

In integrated global markets and in the absence of exchange risk,[2] the expected stock return in Eq. (1) can be written as a function of covariance risk between country i and the world portfolio w, and expressed by:

$$E_{t-1}(r_{it}) = \beta_{i,t-1} E_{t-1}(r_{wt}) = cov_{iw,t-1} \frac{E_{t-1}(r_{wt})}{\sigma^2_{w.t-1}} = \lambda_{w,t-1} cov_{iw,t-1} \tag{6}$$

where $\beta_{i,t-1} = \frac{cov_{iw,t-1}}{\sigma^2_{w,t-1}}$ is the conditional beta at time t-1, $cov_{iw,t-1}$ is the conditional covariance between the return of country i and that of the world portfolio, and $\lambda_{w,t-1} = \frac{E_{t-1}(r_{wt})}{\sigma^2_{w.t-1}}$ is the conditionally expected world price of covariance risk. Transforming $cov_{iw,t-1}$ into the conditional correlation $\rho_{iw,t-1}$, we write:

$$E_{t-1}(r_{it}) = \rho_{iw,t-1} \frac{\sigma_{i,t-1}}{\sigma_{w,t-1}} E_{t-1}(r_{wt}) \tag{7}$$

Following the rationale of Eq. (3), the conditional VaR of the world portfolio, $V_{w,t-1}$, is the product of the conditional quantile value of the world portfolio return distribution and its conditional standard deviation—that is, $V_{w,t-1} = q_{w,t-1}\sigma_{w,t-1}$. Using this relationship, Eq. (7) reads as:

$$E_{t-1}(r_{it}) = \rho_{iw,t-1} \frac{q_{w,t-1}}{q_{i,t-1}} \frac{V_{i,t-1}}{V_{w,t-1}} E_{t-1}(r_{wt}) \tag{8}$$

Given this formula, one may now ask in addition whether a cointegration relationship may exist between downside risks. Recall that if two time series form a comovement in the long-run perspective, they are linearly dependent. Hence for the downside risk case we have:

$$V_{i,t-1} \approx b_i V_{w,t-1} \tag{9}$$

where b_i is the cointegrating coefficient. Eq. (9) states that the series of $V_{i,t-1} - b_i V_{w,t-1}$ is stationary around mean zero. Substituting this expression into Eq. (8) yields:

$$E_{t-1}(r_{it}) = \rho_{iw,t-1} \theta_{iw,t-1} b_i E_{t-1}(r_{wt}) \tag{10}$$

where $\theta_{iw,t-1} = \frac{q_{w,t-1}}{q_{i,t-1}}$ is the conditional quantile ratio between the world portfolio and country i. With the econometric methods below, we will be able to check such cointegration properties and even look out for long memory features.

Eq. (10) links $E_{t-1}(r_{wt})$ as a relevant factor to the expected return $E_{t-1}(r_{it})$ if: (i) $b_i \neq 0$, where the downside risk of country i is cointegrated with that of the world portfolio (see Eq. (9)); (ii) $\rho_{iw,t-1} \neq 0$, describing a nontrivial correlation between $E_{t-1}(r_{it})$and $E_{t-1}(r_{wt})$; and (iii) the ratio $\theta_{iw,t-1} \neq 0$ measures the magnitude of higher moment risk of i relative to the world market. This ratio is greater than one if the world market portfolio has fatter tails than country i. In a nutshell, Eq. (10) displays a variety of sources used to link $E_{t-1}(r_{wt})$ to $E_{t-1}(r_{it})$: the correlation ρ_{iw}, the quantile ratio θ_{iw}, and the long-run cointegrating parameter b_i.

A perfect market integration describes the situation in which i-specific securities will be priced by the same stochastic factor. Since $E_{t-1}(r_{wt})$ is contingent on $V^2_{w,t-1}$, the conditionally local expected returns are consequently correlated with the world downside risk, as can be drawn from the combination of Eqs. (5) and (10):

$$E_{t-1}(r_{it}) = \lambda_{iw,t-1} V^2_{w,t-1} \tag{11}$$

where

$$\lambda_{iw,t-1} = (\rho_{iw,t-1} \theta_{iw,t-1} b_i) \lambda^*_{w,t-1}, \tag{12}$$

$\lambda_{iw,t-1}$, being a measure of risk aversion towards the cointegrated downside risk between i and w, is the world price of downside risk in country i. Eq. (12) links $\lambda_{iw,t-1}$ neatly with $\lambda^*_{w,t-1}$. Indeed, if $\rho_{iw,t-1}$ and b_i are different from zero and if the world downside risk aversion, $\lambda^*_{w,t-1}$, rises, the $\lambda_{iw,t-1}$ of country i will increase accordingly.

The positive value of the estimated $\lambda_{iw,t-1}$ in Eq. (11) is consistent with the market integration hypothesis and confirms that the world downside risk, $V^2_{w,t-1}$, will be priced. The underlying structure in $\lambda_{iw,t-1}$ suggests that the risk aversions are different from country to country, depending on the correlation of individual countries' stock return with that of the world, the quantile ratio, and the cointegrating coefficient.

2.3. Empirical framework and long-run vis-à-vis short-run downside risk aversion

Eq. (11) relates the world downside risk and country i's excess stock return; however, an empirical investigation based on it is confronted with two challenges. Firstly, the full conditional setting is infeasible to examine empirically. We therefore consider, as in Ang et al. (2006) and Bali and Cakici (2010), the empirical version of Eq. (11):

$$r_{i,t+1} = \mu_i + \lambda_{iw} V^2_{w,t} + \varepsilon_{i,t+1} \tag{13}$$

Secondly, estimation of Eq. (11) using ordinary least square regression is biased by the fact that the downside risk exhibits a long memory feature (see Caporin, 2008; Kinateder and Wagner, 2014). In econometric terms, this is denoted as an unbalanced regression with persistent $I(d)$ variables (see Maynard et al., 2013), and we therefore modify Eq. (13) to:

$$r_{i,t+1} = \mu_i + \lambda_{iw} \Delta^d V^2_{w,t} + \varepsilon_{i,t+1} \tag{14}$$

where $\Delta^d V^2_{w,t}$ is the fractionally filtered world downside risk series with an order of integration d. This filter operation, Δ^d, removes the long-memory component and converts $V^2_{w,t}$ from an $I(d)$ series

[2] Like Bali and Cakici (2010), we use the US dollar denominated series in order to eliminate the effect of exchange rate risk on expected returns in our empirical analyses.

to an $I(0)$ process. If in the case of no long memory—that is, $d = 0$—then Eq. (14) boils down to Eq. (13).

Eq. (11) displays the risk premium as a function of the world downside risk series. It does not reflect a possible long-run versus short-run risk premium, however. The quantitative approach taken here offers the opportunity to document these effects. To be more specific, we employ the FCVAR model for the joint dynamics of downside risk that are most likely to be cointegrated (Johansen, 2008; Johansen and Nielsen, 2012; Bollerslev et al., 2013):

$$\begin{pmatrix}\Delta^d V_{i,t}^2\\ \Delta^d V_{w,t}^2\end{pmatrix} = \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix} + \begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix}\begin{pmatrix}-\tilde{\beta}_i & 1\end{pmatrix}\begin{pmatrix}L_d V_{i,t}^2\\ L_d V_{w,t}^2\end{pmatrix} + \begin{pmatrix}\Gamma_{11} & \Gamma_{12}\\ \Gamma_{21} & \Gamma_{22}\end{pmatrix}\begin{pmatrix}L_d \Delta^d V_{i,t}^2\\ L_d \Delta^d V_{w,t}^2\end{pmatrix} + \begin{pmatrix}\varepsilon_{i,t}\\ \varepsilon_{w,t}\end{pmatrix} \quad (15)$$

where $L_d = (1 - \Delta^d)$ is the fractional lag operator, $\tilde{\beta}_i$ is the cointegrating coefficient, while α_1 and α_2 are the adjustment parameters that determine the speed of adjustment to the equilibrium. Γ matrix represents the short-run dynamics. Expanding Eq. (15), the world downside risk dynamics is then expressed as:

$$\Delta^d V_{w,t}^2 = \mu_2 + \alpha_2\left(L_d V_{w,t}^2 - \tilde{\beta}_i L_d V_{i,t}^2\right) + \left(\Gamma_{22} L_d \Delta^d V_{w,t}^2 + \Gamma_{21} L_d \Delta^d V_{i,t}^2\right) + \varepsilon_{w,t} \quad (16)$$

The second term on the right-hand side of the above equation captures the long-run behavior of $\Delta^d V_{w,t}^2$, while the third component represents the short-run dynamics. Finally, recalling Eq. (14) yields the time series regression as:

$$r_{i,t+1} = \mu_i + \lambda_{iw,\ell}\left(L_d V_{w,t}^2 - \tilde{\beta}_i L_d V_{i,t}^2\right) + \lambda_{iw,s} L_d \Delta^d V_{w,t}^2 + \gamma_i L_d \Delta^d V_{i,t}^2 + \varepsilon_{i,t+1} \quad (17)$$

This representation enables us to test the downside risk aversion in an integrated market, separating the long-run risk aversion, $\lambda_{iw,\ell}$, from the short-run risk aversion, $\lambda_{iw,s}$ γ_i stands for the short-run country-specific downside risk aversion.

3. Downside risk estimation and long-memory property

3.1. Data description and estimating downside risk

We collected daily and monthly US dollar-denominated returns on stock market indices for G7 countries and the MSCI world index from Datastream. The stock indices are the total return indices adjusted for dividends. The data cover the MSCI world stock index and seven major stock indices for advanced markets: Canada (CA), France (FR), Germany (GM), the United Kingdom (UK), the United States (US), Italy (IT), and Japan (JP) for the sample period from January 1973 through November 2016. As the risk-free rate, the local three-month T-bill is selected for the US, the UK, and Canada; for the Eurozone countries, the three-month interbank rate is used; and for Japan, we use the three-month gensaki repo rate.[3]

Following the rationale of Dowd (2001) and Cont et al. (2010), we estimate the monthly VaR by using a non-overlapping period of one-month length (usually $n = 22$ trading days). However, the 99% VaR is an extreme quantile ($\alpha = 1\%$) situated in the tail region of the empirical distribution. For this reason, the quantile at the $\alpha = 1\%$ level cannot be calculated precisely based on 22 observations. Accordingly, a nonparametric kernel density estimation (KDE) is employed to smooth the empirical distribution F_t, which yields, in fact, a VaR estimator as a weighted average of the order statistics around the $(n\alpha + 1)$th order statistics. For each month t, we estimate an integrated KDE, $\hat{F}_{t,h}$, with a corresponding bandwidth h, and then perform bootstrap calculations to obtain the desired quantile/VaR estimator for each month (see Appendix A).

It should be pointed out that different kernels will create different KDE tail behaviors and therefore result in (slightly) different VaR estimates. To incorporate more realistic tails, we also consider a double exponential (Laplace) kernel. A modification of the bandwidth can be achieved by using the canonical kernel transformation, resulting in an adjusted bandwidth after multiplying it by an adjustment factor.[4] This procedure enables us to obtain 526 monthly and non-overlapping estimates at $\alpha = 1\%$, and avoids possible statistical problems due to overlapping data (Lettau and Ludvigson, 2010, p. 638).[5]

Table 1 summarizes the statistics of stock returns and downside risk for the G7 and the MSCI world index representing the world portfolio. The monthly excess returns are in the range of 0.30% (Italy) to 0.57% (France), and the VaRs from the Gaussian kernel lie between 3.56% (Italy) and 1.82% (World). Likewise, the VaRs from a double exponential kernel and expected shortfall show their variations across the G7 countries. The data suggest that the VaRs, regardless of the types of kernel, present an AR(1) process. As shown in Fig. 1, the downside risks in the G7 countries and the world market co-move closely, implying that they may have a potential long-run relationship and share a common stochastic trend.

3.2. Long-memory examination for VaR

Most studies of the downside risk-return relation do not pay sufficient attention to the importance of the long-memory process of downside risk series, even though a long-memory property of downside risk is considered to be more appropriate to reflect the persistent fear. Caporin (2008) and Kinateder and Wagner (2014) are among the exceptions, but they keep silent about incorporating the long-memory feature into testing the VaR-return relation. It has been observed that the fear of big financial losses will take a longer period of time for investors to regain their confidence. Therefore, their risk aversion behavior is more likely to exhibit long-term risk aversion (Chen and Chiang, 2016).

To quantify the degree of long memory, we estimate the fractional integration parameter d using both the log-periodogram estimator developed by Geweke and Porter-Hudak (1983, hereafter GPH), and the local Whittle likelihood procedure of Künsch (1986). Table 2 reports the estimated values of GPH d, which range from 0.347 (Japan) to 0.429 (France), indicating that the downside risk in advanced countries entails a stationary long-memory property.[6] The same conclusion can be found in the estimates of the local Whittle likelihood procedure. We also check the spurious long memory caused by occasional structural breaks. The test proposal by Qu (2011) rejects the spurious long memory in all of the VaR series.[7]

[3] Please refer to the web page of Thomson Reuters Datastream. http://extranet.datastream.com/data/Exchange%20&%20Interest%20Rates/RiskFreeInterestRates.htm.

[4] 0.582 is an adjustment factor between the bandwidth of a Gaussian kernel and a double exponential kernel (see Härdle et al., 2004, p.57). The paper's subsequent empirical results for VaR estimates show that the choice of the kernel function is not so relevant for the efficiency of the estimates, which is consistent with the discussion in Härdle et al. (2004, p.57).

[5] Boudoukh et al. (2008), Bali et al. (2009), and Lettau and Ludvigson (2010) found that long-horizon returns become more predictable as the horizon extends. These findings imply that long-horizon returns could, in principle, lead to higher predictability than short-horizon returns. Boudoukh et al. (2008) argue that this evidence is not attributable to small sample bias but rather to the use of overlapping return data.

[6] A suitable value of d usually lies in $|d| < 0.5$. A fractional (non-integer) number with values less than 0 would indicate a weak or memory-less process; if d lies in the interval of (0, 0.5), the series is characterized by a stationary process with long memory; if d lies in $(0.5, \infty)$, the series is a long-memory non-stationary process.

[7] The results will be provided upon request.

Table 1
Descriptive statistics of stock returns, Value-at-Risk (VaR), and ES (Expected Shortfall).

Country	Series	Mean (%)	Std (%)	Skewness	Kurtosis	AR(1)	ADF test
UK	Return	0.45	6.07	0.74	11.63	0.10	−15.93*
	VaR(GaussianKDE)	2.90	1.66	3.14	20.15	0.48	−10.14*
	VaR(exponentialKDE)	2.70	1.57	3.20	21.06	0.47	−10.29*
	ES(GaussianKDE)	3.21	1.80	3.03	19.12	0.49	−9.99*
US	Return	0.45	4.29	−0.40	5.78	0.01	−16.44*
	VaR(GaussianKDE)	2.36	1.61	4.83	45.43	0.45	−9.75*
	VaR(exponentialKDE)	2.21	1.55	5.17	51.79	0.43	−10.04*
	ES(GaussianKDE)	2.61	1.75	4.71	43.52	0.47	−9.70*
GM	Return	0.38	5.73	−0.29	4.14	0.02	−16.39*
	VaR(GaussianKDE)	2.83	1.56	2.29	10.95	0.42	−10.18*
	VaR(exponentialKDE)	2.66	1.50	2.40	11.89	0.40	−10.35*
	ES(GaussianKDE)	3.14	1.70	2.22	10.27	0.44	−10.01*
FR	Return	0.57	6.49	−0.49	4.88	0.05	−16.50*
	VaR(GaussianKDE)	3.06	1.63	2.37	12.24	0.46	−11.11*
	VaR(exponentialKDE)	2.85	1.54	2.35	11.76	0.44	−11.23*
	ES(GaussianKDE)	3.39	1.78	2.33	12.07	0.47	−10.96*
IT	Return	0.30	7.28	−0.11	3.64	0.05	−16.50*
	VaR(GaussianKDE)	3.56	1.94	2.04	9.38	0.40	−11.21*
	VaR(exponentialKDE)	3.33	1.83	2.06	9.59	0.42	−11.06*
	ES(GaussianKDE)	3.95	2.11	2.07	9.55	0.39	−11.37*
CA	Return	0.39	5.27	−0.57	5.27	0.02	−15.58*
	VaR(GaussianKDE)	2.36	1.62	3.37	22.21	0.54	−9.52*
	VaR(exponentialKDE)	2.22	1.54	3.38	21.98	0.52	−9.68*
	ES(GaussianKDE)	2.62	1.77	3.40	22.83	0.56	−9.30*
JP	Return	0.36	5.81	0.28	4.04	0.06	−15.92*
	VaR(GaussianKDE)	2.94	1.66	2.82	19.43	0.37	−10.98*
	VaR(exponentialKDE)	2.75	1.58	2.98	21.58	0.36	−11.19*
	ES(GaussianKDE)	3.26	1.79	2.63	16.99	0.39	−10.79*
World	Return	0.43	4.26	−0.37	4.62	0.07	−15.71*
	VaR(GaussianKDE)	1.82	1.14	3.19	20.05	0.52	−9.75*
	VaR(exponentialKDE)	1.70	1.07	3.21	20.39	0.50	−9.92*
	ES(GaussianKDE)	2.02	1.23	3.14	19.53	0.53	−9.65*

The data cover seven major advanced markets: the United States (US), the United Kingdom (UK), Germany (GM), France (FR), Italy (IT), Canada (CA), Japan (JP) and an additional MSCI world index (World) for the sample period from January 1973 through November 2016. Both excess return and the VaR are calculated on a monthly basis with 526 estimated values. Two kernel densities, a Gaussian and a double exponential kernel, are applied to generate the VaR estimates. ADF is the augmented Dickey–Fuller test. * indicates that the coefficient is significant at the 1% level.

Table 2
Semiparametric analysis for risk measures and spurious check.

	GPH estimate of d	Local Whittle estimate of d
UK	0.360	0.330
US	0.412	0.334
GM	0.420	0.294
FR	0.429	0.310
IT	0.415	0.286
CA	0.410	0.335
JP	0.347	0.322
World	0.374	0.334

The GPH estimate is the log-periodogram estimator by Geweke and Porter-Hudak (1983), and the local Whittle likelihood procedure is from Künsch (1986). A suitable value of d usually lies in $|d| < 0.5$. A fractional (non-integer) number with values less than 0 would indicate a weak or memory-less process; if d falls within the interval of (0, 0.5), the series is characterised by a stationary process with a long memory; if d lies in (0.5,∞) the series shows a non-stationary long-memory process.

4. Fractionally integrated dynamic system

4.1. The empirical model

As discussed in Section 2, the concern of unbalanced regression and the necessity of separating the risk aversion behavior in the long run from the short run motivate us to examine the downside risk-return relation as expressed by Eq. (17). However, it can be expressed in a more general FCVAR$_d(p)$ specification that consists of variables in a vector of downside risks and returns denoted by z_{it} with integration of order d and lag length p in a dynamic system, as:

$$\Delta^d z_i t = \alpha_i(\mu' + \beta_i' L_d z_i t) + \sum\nolimits_{(s=1)}^{p} \Gamma_s L_d^s \Delta^d z_i t + \varepsilon_i t, t = 1, \ldots, T, \tag{18}$$

where $z_{it} \equiv (V_{i,t}^2, V_{w,t}^2, r_{i,t}, r_{w,t})'$ is a 4×1 vector, Δ^d is the fractional difference operator used to remove the long-memory component, μ is interpreted as the mean level of the long-run equilibrium, and ε_{it} is n-dimensional *i.i.d.* $N(0, \Sigma_\varepsilon)$ where $n = 4$. This dynamic FCVAR representation will reduce to the classical error-correction-type representation if $d = 1$, and to VAR if $d = 0$. Note that if $z_{it} \equiv (V_{i,t}^2, V_{w,t}^2)'$, Eq. (18) will boil down to Eq. (15). By adding excess returns to the vector of z_{it} such that $z_{it} \equiv (V_{i,t}^2, V_{w,t}^2, r_{i,t}, r_{w,t})'$, we undertake an investigation of the risk-return relation presented in Eq. (17), although we may particularly focus on the expansion of $r_{i,t}$ in z_{it}.

The coefficient matrix $\Pi_i = \alpha_i \beta_i'$ is an $n \times n$ matrix where α_i and β_i are an $n \times m$ matrix with $m \leq n$. The columns of β_i are the cointegrating vectors, which represent the long-run equilibrium relations; the coefficients in α_i are the adjustment parameters that determine the speed of adjustment to the equilibrium. The second term on the right-hand side of Eq. (18), which specifies the fractional distributed lag matrix Γ_s and powers of L_d^s applied to $\Delta^d z_{it}$, directly mirrors the distributed lag matrix in standard error-correction models. The parameters in Γ_s govern the short-run dynamics of the variables. The selection of lag length p in the fractional distributed lag matrix Γ_s is based on the Bayesian-Schwarz

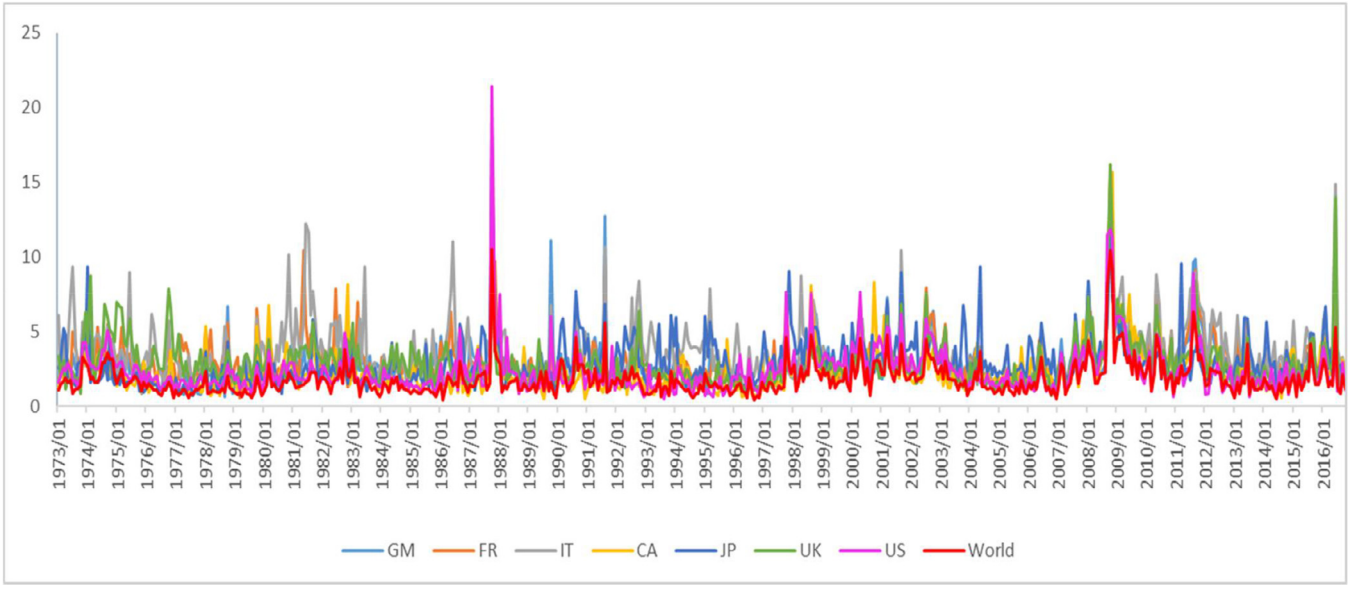


Fig. 1. Time variations of G7 downside risk.
Downside risk is measured by the 99% VaR, which is obtained by using kernel smoothing for the empirical distribution and then bootstrapping from the kernel density estimator. The Gaussian kernel density has been applied in this figure.

information criterion (BIC), the Ljung-Box Q-test for each residual series, and the likelihood ratio test for testing the significance of the Γ_s.

4.2. The economic interpretation of the model and cointegration rank test

To elucidate the parametric relations of the system, we expand Eq. (18) by setting the lag length $p=1$ as[8]

$$\begin{pmatrix}\Delta^d V_{i,t}^2\\ \Delta^d V_{w,t}^2\\ \Delta^d r_{i,t}\\ \Delta^d r_{w,t}\end{pmatrix} = \begin{pmatrix}\alpha_{11}&\alpha_{12}&\alpha_{13}\\ \alpha_{21}&\alpha_{22}&\alpha_{23}\\ \alpha_{31}&\alpha_{32}&\alpha_{33}\\ \alpha_{41}&\alpha_{42}&\alpha_{43}\end{pmatrix}\mu' + \begin{pmatrix}\alpha_{11}&\alpha_{12}&\alpha_{13}\\ \alpha_{21}&\alpha_{22}&\alpha_{23}\\ \alpha_{31}&\alpha_{32}&\alpha_{33}\\ \alpha_{41}&\alpha_{42}&\alpha_{43}\end{pmatrix} \begin{pmatrix}-\tilde{\beta}_i&1&0&0\\ 0&0&1&0\\ 0&0&0&1\end{pmatrix} \begin{pmatrix}L_d V_{i,t}^2\\ L_d V_{w,t}^2\\ L_d r_{i,t}\\ L_d r_{w,t}\end{pmatrix} + \begin{pmatrix}\Gamma_{11}&\Gamma_{12}&\Gamma_{13}&\Gamma_{14}\\ \Gamma_{21}&\Gamma_{22}&\Gamma_{23}&\Gamma_{24}\\ \Gamma_{31}&\Gamma_{32}&\Gamma_{33}&\Gamma_{34}\\ \Gamma_{41}&\Gamma_{42}&\Gamma_{43}&\Gamma_{44}\end{pmatrix} \begin{pmatrix}L_d \Delta^d V_{i,t}^2\\ L_d \Delta^d V_{w,t}^2\\ L_d \Delta^d r_{i,t}\\ L_d \Delta^d r_{w,t}\end{pmatrix} + \begin{pmatrix}\varepsilon_{i1,t}\\ \varepsilon_{i2,t}\\ \varepsilon_{i3,t}\\ \varepsilon_{i4,t}\end{pmatrix}. \tag{19}$$

This model specification encompasses a variety of financial market theories, including testing the restrictions or causal relations in the long run through elements of α_i,[9] and in the short run through elements of Γ_s for: (i) risk spillovers or financial contagion: $V_{w,t}^2 \to V_{i,t}^2$ (King and Wadhwani, 1990; Diebold and Yilmaz, 2009); (ii) comovement of stock returns: $r_{w,t} \to r_{i,t}$ (Karolyi and Stulz, 1996; Forbes and Rigobon, 2002; Caporale et al., 2005; Chiang et al., 2007); (iii) risk-return or volatility feedback: $V_{i,t}^2 \to r_{i,t}$ (French et al., 1987; Bali and Peng, 2006; Bali et al., 2009); and (iv) the leverage effect: $r_{i,t} \to V_{i,t}^2$ (Bekaert and Wu, 2000; Bollerslev et al., 2006).

Since the existence of a cointegration relationship is a necessary condition in the FCVAR model, one has to examine the number of cointegration ranks. In this study, the rank of Π_i could be three ($m=3$) or four ($n=4$), requiring us to perform the cointegration rank test with respect to the following two hypotheses, $H_0:\ \text{rank}(\Pi_i)=m$ against $H_1:\ \text{rank}(\Pi_i)=n$. The likelihood ratio (LR) test statistic is given by:

$$\text{LR}_T(n-m) = 2\log\left(L\left(\hat{d}_n, n\right)/L\left(\hat{d}_m, m\right)\right) \tag{20}$$

where $L(\hat{d}_n, n)$ represents the profile likelihood function given rank n, and other parameters have been concentrated out (see Johansen and Nielsen, 2012, p. 2698). The asymptotic distribution of test statistics in Eq. (20) is highly dependent on the parameter of d. In the case of $0<d<0.5$, it has a standard asymptotic distribution to χ^2 with degree of freedom$(n-m)^2$. For $d \geq 0.5$, asymptotic theory is non-standard and involves fractional Brownian motion of type II. Table 2 shows that all estimated fractional parameters $\hat{d}$ are below 0.5. The results in Table 3 show rank $(\Pi_i)=3$— that is, one cointegrating vector. It confirms that the downside risk for each G7 market is cointegrated with that of the world market.

4.3. Estimations and inferences from the FCVAR model

The estimation of the FCVAR model can be arrived at by using the maximum likelihood method. An asymptotic analysis shows that the maximum likelihood estimators are asymptotically normal conditional on the initial values. The log-likelihood function corresponding to Eq. (18) is:

$$\log L_T(\Theta_i) = -\frac{T}{2}\log\det\left(T^{-1}\sum_{t=1}^{T}\varepsilon_{it}(\Theta_i)\varepsilon_{it}(\Theta_i)'\right) \tag{21}$$

where $\varepsilon_{it}(\Theta_i) = \Delta^d z_{it} - \alpha_i(\beta'_i L_d z_{it} + \mu') - \sum_{s=1}^{p}\Gamma_s L_d^s \Delta^d z_{it}$, and $\Theta_i = (d, \alpha_i, \beta_i, \mu', \Gamma_s)$. Under *i.i.d.* errors with suitable mo-

[8] It is possible that the fractional difference operator applied to the returns potentially leads to over-difference. However, according to theorem 8 in Johansen (2008) and the illustration in Appendix A1 in Bollerslev et al. (2013), if the conditions for inversion of the FCVAR$_d(p)$ are satisfied and $d<0.5$, the resulting return series remains stationary.

[9] The long run here refers to the analytical long run in Bollerslev et al. (2013) instead of the one documented by Lundblad (2007), and the time series statistical framework (Bollerslev et al., 2013) instead of the historical long run documented by Lundblad (2007).

Table 3
Cointegration rank tests.

Rank	UK	US	GM	FR	IT	CA	JP
0	64.163*	34.513*	37.641*	58.599*	57.248*	56.009*	49.380*
1	35.676*	18.055*	15.848*	32.254*	32.492*	23.964*	14.861*
2	4.093*	12.157*	8.908*	21.795*	15.468*	10.813*	4.339*
3	1.225	1.682	2.899	3.828	3.044	2.512	2.606

This table shows the results of the likelihood ratio test statistics shown in Eq. (20).
* denotes the significance at the 5% level.

ment conditions, the conditional maximum likelihood estimates are asymptotically Gaussian.[10]

As shown in Table 4, the estimated long-memory parameters $\hat{d}$ for the FCVAR dynamic system range from 0.168 (JP) to 0.451 (UK). The estimated $\hat{d}$ values relative to respective standard errors (in parentheses) are significant, suggesting that neither the simple VAR models nor the VECM are adequate to describe the dynamic system between downside risk series and return series. The estimates of $\tilde{\beta}_i$ vary among G7 countries, revealing different specifications of cointegration—e.g., $1.084V_{it}^2 \approx V_{wt}^2$ for the case of the UK and $0.750V_{it}^2 \approx V_{wt}^2$ for the case of the US. The cointegration between the UK and the world market translates into the result that the series $1.084V_{it}^2 - V_{wt}^2$ is stationary around mean zero. From this analysis, it can be stated that the long-run downside risk for the UK is roughly 92% of the world downside risk compared with the figure for the US, which is roughly 30% higher than that of world market.[11]

4.3.1. The long-run tradeoff

The long-run tradeoff relation is captured by the parameter α_{31}, which directly maps to $\lambda_{iw,\ell}$ in Eq. (17) as a long-run downside risk aversion parameter. As shown in Table 4, the estimated coefficients, 23.140 (UK), 0.148 (US), 1.190 (GM), 1.527 (FR), 0.421 (IT), and 0.602 (CA), are all positive and statistically significant, confirming a long-run downside risk-return tradeoff. The exception is the Japanese market with a negative sign but statistically insignificant, suggesting that Japanese investors and their markets behave quite differently from those in other advanced markets.

To gain more insight, we recall the definition of λ_{iw}, the measure of risk aversion toward the cointegrated downside risk between country i and world market, in Eq. (12). Table 5 presents the estimates of various components in connecting λ_{iw} given λ_w^*. Inspecting the estimated figures for Japan, we see that stock return correlation ($\rho_{iw} = 0.690$), the quantile ratio of the world portfolio to domestic country ($\theta_{iw} = 0.618$) and its cointegrating coefficient ($b_i = 0.728$) are lower for the Japanese market compared with results from other markets. As a consequence, using Eq. (12), the risk aversion coefficient toward cointegrated downside risk $\lambda_{iw} = \rho_{iw}\theta_{iw}b_i\lambda_w^* = 0.311\lambda_w^*$ for Japan is the lowest among the G7 countries, implying a relatively low risk aversion in connection with the world compared with other advanced markets.

The estimated results α_{31} yield three economic insights. Firstly, the estimated α_{31} sheds some light on $\lambda_{iw,t-1}$, especially from a long-run cointegration perspective, which is linked to $\lambda_{i,t-1}^*$ in Eq. (4) through b_i using Eq. (9).[12] The significance of the long-run downside risk aversion parameter further implies that the downside risk-return relation can be inferred not only in the local market ($\lambda_{i,t-1}^*$) but also in an integrated market ($\lambda_{iw,t-1}$). Secondly, investors are averse to long-run cointegrated downside risk, $\lambda_{iw,\ell}$, as discussed in the economic framework and empirically reflected in the estimated α_{31}. Therefore, one may not obtain satisfying results from a direct estimation for λ_{iw}, which is the mixture of long-run and short-run risk aversion. Thirdly, the finding confirms that risk aversion is driven by tail risk and, in turn, higher moment risk implied via Cornish-Fisher expansions. Thus, the investor behavior is tied closely to its aversion to skewness and kurtosis, as in the expected utility framework.

The overwhelming significance of the long-run downside risk aversion parameter suggests the need to incorporate a long-memory feature and to separate the long-run from the short-run effect in testing the risk-return relation. This approach allows us to potentially tease out the empirically inconclusive risk-return relation. Due to the long-run nature of cointegrated downside risk, investors' aversion to the long-lasting risk commands a long-run risk reward.

4.3.2. The short-run tradeoff and leverage effects

The lag length in the short-run dynamics is jointly determined by the Bayesian-Schwarz information criterion (BIC) and the Ljung-Box Q-statistic, which test for each residual series without entailing serial correlations. We then examine the significance of the likelihood ratio test for the significance of the Γ_s. The overall performance of tests suggests that $p = 1$ is appropriate.[13]

The short-run tradeoff hypothesis can be examined by checking the estimated coefficient of Γ_{31} (and Γ_{41}), which measures the short-term impact of increased domestic (world) risk on domestic expected return. Likewise, Γ_{41} is analogous to $\lambda_{iw,s}$ in Eq. (17), representing the short-run downside risk aversion at the worldwide level, while Γ_{31}, which is analogous to γ_i, represents the short-run downside risk aversion at the country level. For instance, when the US interacts with the world market, the evidence in Table 4 shows that $\Gamma_{31} = -1.066$ and $\Gamma_{41} = -1.363$, indicating a negative short-run risk-return relation that may be attributed to the fear of big capital losses caused by downside risk and an eventual selloff of stocks from investors' portfolios (Chen and Chiang, 2016). This behavior consequently deviates from the tradeoff hypothesis. Bearing this risk, however, will be compensated in the long run as the downside shock gradually dies out, which is consistent with our assertion of a long-run tradeoff hypothesis. Pairings of countries in the G7 with the world market did not show a significant short-run tradeoff effect indicated by Γ_{31}. However, for the Γ_{41}, the results for Germany and Canada show that their short-run downside risk contributes to future world returns. Based on these findings, the tradeoff hypothesis in the short run does not perform as promisingly as that in the long run.

[10] In estimating the FCVAR model represented by Eq. (18), the VaR is defined as the downside risk derived from a Gaussian kernel.

[11] By using the sample period from September 1990 to July 2013, the statistical results here are comparable with our earlier finding, suggesting the empirical results are robust even using a shorter subsample.

[12] $\lambda_i^* = \frac{E_{t-1}(r_{it})}{V_{i,t-1}^2}$ and $\lambda_{iw} = \frac{E_{t-1}(r_{it})}{V_{w,t-1}^2} = \frac{E_{t-1}(r_{it})}{\tilde{\beta}_i V_{i,t-1}^2} = \frac{1}{\tilde{\beta}_i}\lambda_i^* = b_i^2\lambda_i^*$. Using Eq. (9), $V_{i,t} \approx b_i V_{w,t}$, we find that the λ_i^* links to λ_{iw} by the information of b_i^2. Note that $b_i^2 \cong 1/\tilde{\beta}_i$ because $V_{it} \approx b_i V_{wt}$ and $\tilde{\beta}_i V_{it}^2 \approx V_{wt}^2$. In Eq. (12), we know $\lambda_{iw} = (\rho_{iw}\theta_{iw}b_i)\lambda_w^*$. Putting this together, we have $\lambda_{iw} = b_i^2\lambda_i^* = (\rho_{i,t-1}\theta_{i,t-1}b_i)\lambda_w^*$.

[13] When allowing for fractional cointegration in the long-run equilibrium relations, fewer lags are required in the autoregressive system (Dolatabadi et al., 2015).

Table 4
The estimates of FCVAR model for G7 and world indices using a Gaussian kernel.

	d	$-\tilde{\beta}_i$	$\mu'_{3\times1}$	$\alpha_{i(4\times3)}$			$\Gamma_{s=1,(4\times4)}$				*BIC*
UK	0.451	−1.084	$\begin{pmatrix}-0.009\\-0.009\\-0.004\end{pmatrix}$	5.524	−45.641	−16.861	−5.856	55.824	−17.299	−14.467	718
	(0.02)			(4.46)	(14.53)	(6.45)	(5.54)	(16.23)	(6.56)	(7.03)	
				−3.847	−10.573	−4.307	3.671	16.617	−4.641	−7.645	
				(2.52)	(5.29)	(2.34)	(2.84)	(6.25)	(2.41)	(3.28)	
				23.140	−8.202	−30.641	−20.331	−2.777	31.041	−17.876	
				(10.81)	(35.08)	(18.04)	(18.24)	40.85)	(18.57)	(18.17)	
				14.619	1.455	−4.911	−12.006	−8.191	5.278	4.670	
				(12.70)	(9.77)	(13.02)	(14.84)	(34.94)	(13.43)	(4.56)	
US	0.284	−0.750	$\begin{pmatrix}-0.018\\-0.053\\-0.097\end{pmatrix}$	−0.264	−0.541	−0.091	−1.129	3.556	0.095	−0.293	848
	(0.04)			(0.19)	(0.57)	(2.01)	(0.32)	(1.08)	(0.22)	(0.23)	
				−0.074	−0.389	−0.038	−0.177	1.267	0.048	−0.111	
				(0.07)	(0.20)	(0.07)	(0.11)	(0.41)	(0.08)	(0.08)	
				0.148	1.421	−0.749	−1.066	−0.685	−0.539	0.230	
				(0.05)	(1.09)	(0.38)	(0.64)	(1.92)	(0.41)	(0.42)	
				0.751	−0.011	0.034	−1.363	0.845	0.021	−0.035	
				(0.39)	(1.24)	(0.43)	(0.70)	(2.16)	(0.48)	(0.47)	
GM	0.247	−0.721	$\begin{pmatrix}-0.276\\-0.102\\-0.227\end{pmatrix}$	−0.443	0.322	0.153	−0.567	2.141	−0.189	0.139	782
	(0.05)			(0.34)	(0.70)	(0.12)	(0.41)	(1.05)	(0.13)	(0.11)	
				−0.006	−0.353	0.055	−0.225	1.035	−0.058	−0.081	
				(0.20)	(0.42)	(0.07)	(0.25)	(0.68)	(0.08)	(0.07)	
				1.190	−0.830	−0.893	1.526	−3.380	−0.121	0.034	
				(0.46)	(2.90)	(0.49)	(1.89)	(3.86)	(0.55)	(0.46)	
				−2.181	5.271	0.241	4.522	−8.872	−0.222	−0.036	
				(1.48)	(2.97)	(0.48)	(1.96)	(3.86)	(0.52)	(0.46)	
FR	0.394	−1.704	$\begin{pmatrix}-0.925\\-1.514\\-0.183\end{pmatrix}$	−0.649	−0.366	−0.018	−0.443	2.085	−0.022	−0.037	656
	(0.04)			(0.16)	(0.17)	(0.07)	(0.21)	(0.44)	(0.07)	(0.09)	
				−0.291	0.121	−0.039	0.204	−0.073	0.036	−0.071	
				(0.08)	(0.09)	(0.36)	(0.11)	(0.21)	(0.04)	(0.05)	
				1.527	−2.383	−1.072	−0.886	1.445	0.160	−0.304	
				(0.70)	(0.76)	(0.30)	(0.94)	(1.53)	(0.31)	(0.38)	
				0.412	0.144	−0.054	0.260	−1.721	0.170	−0.210	
				(0.52)	(0.56)	(0.22)	(0.70)	(1.15)	(0.23)	(0.28)	
IT	0.398	−1.325	$\begin{pmatrix}-0.269\\0.001\\-0.111\end{pmatrix}$	−0.385	0.154	−0.263	−0.130	0.524	−0.233	0.119	579
	(0.04)			(0.09)	(0.21)	(0.08)	(0.17)	(0.41)	(0.09)	(0.10)	
				−0.106	−0.256	−0.039	0.133	0.174	0.047	−0.076	
				(0.03)	(0.07)	(0.03)	(0.06)	(0.18)	(0.03)	(0.04)	
				0.421	−1.289	−1.008	−0.261	1.556	0.090	−0.102	
				(0.22)	(0.69)	(0.25)	(0.54)	(1.41)	(0.27)	(0.37)	
				0.087	0.931	−0.035	0.422	−2.278	0.233	−0.280	
				(0.21)	(0.45)	(0.16)	(0.35)	(0.93)	(0.18)	(0.24)	
CA	0.328	−1.168	$\begin{pmatrix}-0.367\\-0.142\\-0.301\end{pmatrix}$	−0.163	−0.449	0.290	−0.623	3.273	−0.234	0.129	1027
	(0.03)			(0.07)	(0.04)	(0.06)	(0.23)	(0.51)	(0.07)	(0.09)	
				0.206	−0.768	0.053	−0.387	1.093	−0.067	0.015	
				(0.01)	(0.07)	(0.01)	(0.11)	(0.26)	(0.03)	(0.05)	
				0.602	1.463	−1.009	0.171	−1.544	−0.351	0.436	
				(0.32)	(0.46)	(0.29)	(0.97)	(1.78)	(0.34)	(0.42)	
				1.169	4.021	−0.194	1.993	−1.641	0.050	−0.224	
				(0.32)	(0.28)	(0.23)	(0.77)	(1.41)	(0.27)	(0.34)	
JP	0.168	−1.885	$\begin{pmatrix}-0.209\\-0.426\\0.135\end{pmatrix}$	0.930	−3.603	0.110	−2.496	6.775	−0.196	−0.596	711
	(0.05)			(0.95)	(2.86)	(0.35)	(1.19)	(3.82)	(0.38)	(0.69)	
				0.614	−2.077	−0.045	−1.241	4.141	0.027	−0.329	
				(0.45)	(1.41)	(0.17)	(0.63)	(2.17)	(0.19)	(0.34)	
				−2.184	5.111	−0.395	2.730	−5.158	−1.006	0.413	
				(2.55)	(5.79)	(1.08)	(3.12)	(7.43)	(1.16)	(1.76)	
				−2.398	7.872	0.859	3.399	−10.125	−1.180	1.068	
				(2.21)	(5.66)	(0.93)	(2.75)	(7.05)	(1.03)	(1.59)	

This table presents the estimates of the FCVAR model expressed as: $\Delta^d z_{it} = \alpha_i(\mu' + \beta'_i L_d z_{it}) + \sum_{s=1}^{p} \Gamma_s L_d^s \Delta^d z_{it} + \varepsilon_{it}$, $t = 1, \ldots, T$, where $z_{it} \equiv (V_{i,t}^2, V_{w,t}^2, r_{i,t}, r_{w,t})'$.
The numbers in parentheses are standard errors.

Table 5
The attributes of market integration.

Attribute	UK	US	GM	FR	IT	CA	JP
ρ_{iw}	0.757	0.858	0.741	0.740	0.716	0.756	0.690
θ_{iw}	0.628	0.770	0.642	0.595	0.510	0.769	0.618
$\tilde{\beta}_i$	1.084	0.750	0.721	1.704	1.325	1.168	1.885
b_i	0.960	1.155	1.178	0.766	0.869	0.925	0.728
$\rho_{iw}\theta_{iw}b_i$	0.456	0.763	0.560	0.339	0.317	0.538	0.311

This table reports the parameters in Eq. (12) given by $\lambda_{iw} = \rho_{iw}\theta_{iw}b_i\lambda_w^*$ in an unconditional version where λ_w^* is fixed for each country. λ_{iw} is the expected world price of cointegrated downside risk, depending on the expected correlation between local and world stock returns (ρ_{iw}), and the expected quantile ratio between them (θ_{iw}), and the cointegrating coefficient (b_i). Note that $b_i \cong 1/\sqrt{\tilde{\beta}_i}$ due to $V_{it} \approx b_i V_{wt}$ and $\tilde{\beta}_i V_{it}^2 \approx V_{wt}^2$ implied in Eq. (18). The estimates of $\tilde{\beta}_i$ are reported in Table 4.

To test the leverage effect that the negative return leads to a higher variance in future, we replace variance with downside risk. The estimated parameters of α_{12}, α_{13}, α_{22}, and α_{23} capture the long-run leverage effect, while the parameters of Γ_{13}, Γ_{14}, Γ_{23}, and Γ_{24} represent the short-run leverage effect. The evidence shows that the short-run leverage effect, Γ_{13}, produces a negative sign for the UK (−17.29), Italy (−0.233), and Canada (−0.234), while the long-run leverage effect (α_{12}) is significant for the UK (−45.641), France (−0.366), and Canada (−0.449). Even though the leverage effect lacks robustness, the inclusion of this variable in the model helps to serve as a control variable in testing the risk-return hypothesis.

4.4. Model predictability

Although the evidence presented by the FCVAR model supports a long-run risk-return relation, it is reasonable to confirm superior performance in comparison to a simple vector autoregression (VAR) model in an out-of-sample prediction. We fit Eq. (18) by excluding the long-run component, $\alpha_i(\beta_i' L_d z_{it} + \mu')$, and by setting $d = 0$. This change allows us to evaluate the significance of long-run performance one month ahead in out-of-sample forecasts, using a 35-year rolling window. The first window ranges from January 1973 to December 2007, yielding the first monthly forecast in January 2008. The window keeps rolling one month until the end of the sample period. In total, this procedure yields 106 forecasts based on the FCVAR model and two benchmarked models, a VAR(2) (single-country) model with $z_{it} \equiv (V_{i,t}^2, r_{i,t})'$, and a VAR(4) (cross-country caused by market integration) model with $z_{it} \equiv (V_{i,t}^2, V_{w,t}^2, r_{i,t}, r_{w,t})'$.[14]

Fig. 2 depicts the out-of-sample forecasts against the actual level. Nevertheless, a further out-of-sample predictive ability test is necessary. Here we conduct the conditional predictive ability (CPA) test of Giacomini and White (2006). The Giacomini and White (2006) test essentially provides improvements, in several respects, on the Diebold and Mariano (1995) approach, which has been in widespread use for predictive evaluation. Firstly, the test can exist in an environment in which the sample is finite; secondly, it can handle forecasts based on both nested and non-nested models. In our case, two benchmarked models are nested in the FCVAR model. Thirdly, and more importantly, the model accommodates conditional predictive evaluation in such a way that we can predict which forecast will be more accurate on a specific future day.

An application of this test shows that the FCVAR model has a superior forecasting ability relative to the alternative VAR models. The CPA test statistics are 6.51 and 10.22 for the case of the FCVAR versus a two-country model; versus a one-country model, the corresponding p-values of the test statistics are 0.039 and 0.006. We therefore conclude that the inclusion of a long-run specification is informative in Eq. (18). By the same examination, a two-country model predicts better than a one-country model given that the test statistic is 14.05 and the p-value smaller than 1%. One arrives at the conclusion that investors do ask for a worldwide downside risk premium.

5. Robustness checks

5.1. Evidence from alternative kernel density estimators

As discussed in Section 2.1 and Appendix A, the tail behavior of the KDE is closely tied to the choice of kernel densities, which, in turn, generates (slightly) different estimates of the VaR. These interconnections lead us to ask whether the downside risk-return relation is robust given the choice of kernel densities used to obtain the VaR estimates. This inquiry can be made by fitting a double exponential (Laplace) kernel for a possible fat tail in the stock return distribution. As shown in Table 1, the statistics of the VaR estimates derived from the double exponential kernel are compatible to those from the Gaussian kernel. The downside risk derived from the double exponential kernel is therefore considered in the FCVAR system. By comparing the results in Panel A of Table 6 with those in Table 4,[15] we find that the estimated results are quite compatible. Our results confirm the argument in Härdle et al. (2004) that the choice of the kernel function is not crucial for the efficiency of the estimates and, as a consequence, it has no significant impact on the final results.

5.2. Expected shortfall and expected VaR

As noted by Artzner et al. (1999), one of the weaknesses of the VaR risk measure is that it is not coherent—that is, it may fail to satisfy the sub-additivity property and violates the principle that a merger does not create extra risk. Another weakness of VaR is that it lacks information on the losses beyond the VaR. For these reasons, the expected shortfall, the mean value of losses larger than VaR, may be used to replace VaR as a downside risk measure. The corresponding results, which are shown in Panel B of Table 6, are consistently significant.

The rationale for using the lagged VaR as a proxy for the expected VaR is that VaR behaves extremely persistently. Because of this predictability, we follow Bali et al. (2009) in considering an AR(p) specification as a proxy for expected VaR. Using the BIC, one can select the optimal lag and then use the current and lagged VaR to project the expected VaR. Panel C of Table 6 shows that in using the expected VaR, the estimates do not significantly differ from those we derived previously.

5.3. Predictability of skewness

Because skewness, as a measure of the degree of asymmetry in the distribution, constitutes a downside risk, it naturally begs the question of whether the downside risk-return tradeoff hypothesis can be directly examined by employing skewness as an argument, such as $z_{it} = (Skew_{i,t}, Skew_{w,t}, r_{i,t}, r_{w,t})'$ in the framework of Eq. (18). Empirical skewness can then be calculated based on a monthly interval (from the first trading day until the last trading day of the month t). Using this approach, the coefficients of risk

[14] To save space, we present only the world-UK constellation. Other pairs have consistent forecasting results and are available upon request.

[15] To save space, we report only the case of the UK. The statistics for other countries are available upon request.

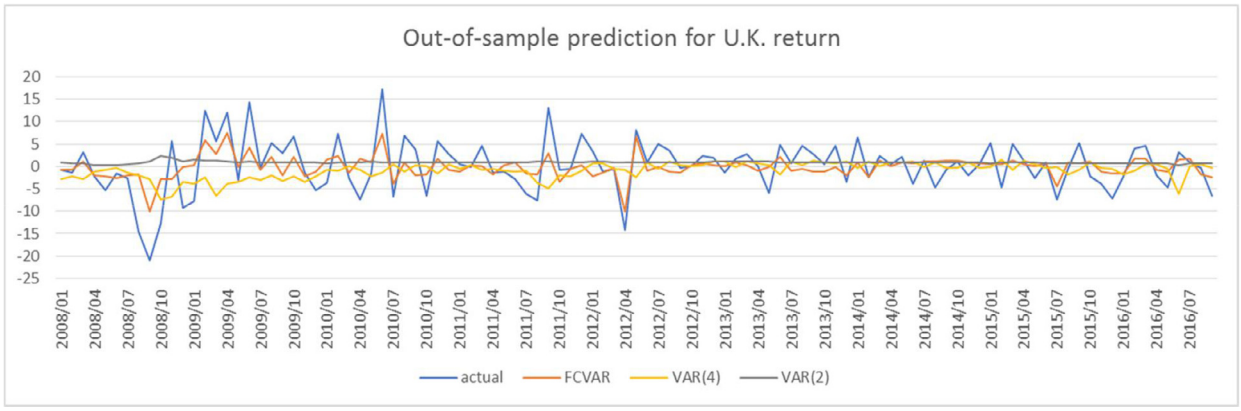


Fig. 2. Out-of-sample predictions for alternative models.
The out-of-sample forecast performances are depicted among the FCVAR, a VAR(2) (single-country model with $z_{it} \equiv (V_{i,t}^2, r_{i,t})'$) and a VAR(4) (cross-country model with $z_{it} \equiv (V_{i,t}^2, V_{w,t}^2, r_{i,t}, r_{w,t})'$) model. The vertical axis is the predictive return (%), in this case for UK, compared with the actual return series.

Table 6
Robustness checks based on different measures: evidence from the UK market.

d	$-\tilde{\beta}_i$	$\mu'_{3\times 1}$	$\alpha_{i(4\times 3)}$			$\Gamma_{s=1,(4\times 4)}$				*BIC*
Panel A. Double exponential kernel										
0.452	−1.399	$\begin{pmatrix}-0.004\\-0.001\\0.001\end{pmatrix}$	1.962	−32.256	−11.234	−1.835	34.735	−11.346	−9.715	957
(0.06)			(2.59)	(13.87)	(4.65)	(2.79)	(14.46)	(4.68)	(5.33)	
			−2.493	−9.380	−3.059	2.429	11.082	−3.140	−5.077	
			(1.52)	(4.50)	(1.86)	(1.56)	(4.83)	(1.88)	(2.48)	
			21.269	−7.989	−22.302	−20.690	5.057	22.560	−13.118	
			(11.77)	(17.99)	(20.89)	(11.78)	(19.27)	(21.18)	(26.86)	
			10.995	4.584	−3.733	−10.251	−6.742	3.823	3.136	
			(5.68)	(13.08)	(14.93)	(5.92)	(14.15)	(15.13)	(18.44)	
Panel B. Expected shortfall										
0.448	−1.273	$\begin{pmatrix}-0.011\\-0.015\\-0.002\end{pmatrix}$	2.457	−30.906	−14.508	−2.378	39.118	−14.982	−12.230	412
(0.06)			(2.74)	(9.48)	(5.33)	(3.47)	(10.88)	(5.34)	(5.79)	
			−2.620	−7.821	−3.548	2.465	13.038	−3.870	−6.462	
			(1.48)	(3.52)	(1.96)	(1.73)	(4.35)	(2.02)	(2.71)	
			15.165	−5.140	−23.137	−13.109	−2.907	23.356	−14.456	
			(8.73)	(21.76)	(13.15)	(10.59)	(26.51)	(13.59)	(14.05)	
			8.194	3.989	−4.304	−6.035	−9.408	4.656	1.749	
			(6.60)	(17.43)	(9.48)	(8.17)	(21.29)	(9.81)	(10.96)	
Panel C. Expected VaR										
0.131	−1.175	$\begin{pmatrix}-0.119\\-0.063\\-0.073\end{pmatrix}$	−2.384	0.065	−0.712	3.895	0.853	−0.796	−0.594	778
(0.01)			(0.67)	(0.55)	(0.26)	(0.82)	(0.83)	(0.27)	(0.26)	
			−1.891	0.802	−0.291	2.345	0.507	−0.366	−0.570	
			(0.51)	(0.38)	(0.18)	(0.58)	(0.59)	(0.19)	(0.19)	
			15.550	−14.847	−5.810	−15.248	10.647	5.315	−2.358	
			(5.63)	(5.02)	(2.30)	(6.67)	(7.40)	(2.43)	(2.33)	
			10.237	−8.825	−1.298	−10.123	8.051	1.431	0.554	
			(4.42)	(3.92)	(1.74)	(5.35)	(5.90)	(1.85)	(1.85)	
Panel D. Skewness										
0.108	−0.150	$\begin{pmatrix}-0.009\\-0.041\\0.125\end{pmatrix}$	−1.140	0.889	−0.424	−0.253	0.177	0.288	−0.122	537
(0.03)			(0.47)	(0.93)	(0.53)	(0.45)	(0.20)	(0.20)	(0.14)	
			0.107	−0.820	−0.021	−0.167	−0.220	0.099	0.155	
			(0.52)	(0.77)	(0.22)	(0.48)	(0.49)	(0.15)	(0.13)	
			3.527	−5.510	−3.687	−8.362	4.693	1.847	−1.742	
			(2.35)	(3.95)	(1.33)	(7.21)	(4.45)	(1.18)	(1.09)	
			5.223	−4.389	−0.041	−3.642	2.966	0.323	−0.689	
			(5.46)	(3.97)	(1.20)	(4.23)	(3.21)	(1.09)	(1.20)	

This table reports the estimates of Eq. (18) by using different measures of risk. The numbers in parentheses are standard errors. Estimates in Panel A are based on the VaR from a double exponential kernel. Estimates in Panel B are based on the risk measure represented by the expected shortfall from a Gaussian kernel. Estimates in Panel C are based on the risk measure of expected VaR. Estimates in Panel D are using the empirical skewness as a measure for the downside risk when testing the risk-return relation. In sum, the evidence for supporting the long-run tradeoff is robust across different model specifications.

aversion toward skewness are insignificant, indicating that skewness, relative to VaR, has a limited return predictability. The findings are consistent with the results documented by Bali et al. (2009) for the case of US data and therefore lead to the conclusion that richer information content is present embedded in VaR.

In summary, the evidence to support the long-run tradeoff on major advanced markets is robust across different model specifications. A small sample bias may appear when the coefficients in a dynamic system with highly persistent regressors are tested. Conducting a biasness test shows that the framework employed in this study is robust relative to the bias.[16]

6. Conclusions

In considering the possible heavy tail property of G7 returns, we extend the variance-based risk measure to a higher moment-based risk measure via kernel density estimators. This study investigates the downside risk-return relation in an integrated market framework. The theoretical model suggests that the expected world price of cointegrated downside risk in a specific country is subject to its return correlation with the world market, the quantile ratio between them, and the long-run cointegration. The downside risk aversion in an integrated market is further decomposed into long-run and short-run components. The relative contribution of each is left to empirical investigation via the FCVAR model with fractional cointegration (long-run) and vector autoregression (short-run) analyses of world and country-specific risk-return relations.

We find evidence that, with the exception of Japan, downside risk forms a cointegration relationship with the world market in the long run, supporting the risk-return tradeoff hypothesis in the long run. Evidence suggests that investors are averse not only to local downside risk but also to world downside risk, as these downside risks are significantly cointegrated in the long run. This indicates that investors command long-run risk premiums with respect to the cointegrated downside risk. We also find that the proposed FCVAR model indeed yields better stock return predictability than a pure short-run-based VAR model.

Appendix A. The nonparametric VaR estimates

The following section presents detailed description of the proposed procedure. In designating the daily excess returns $\{r_{t,i}\}_{i=1}^{n}$ within month t and n as the number of trading days in this specific month, the KDE-based smoothed distribution function is defined as

$$\hat{F}_{t,h}(x) = n^{-1}\sum_{i=1}^{n}\int_{-\infty}^{x} K_h(u - r_{t,i})du, \tag{A.1}$$

where $K_h(s) = h^{-1}K(s/h)$ is the rescaled kernel with bandwidth h. The bandwidth is chosen so that the squared bias and the variance are balanced, which minimizes the tradeoff between a shrinking variance and a rising bias that occurs as h increases. The bandwidth h, therefore, should be optimized to balance this tradeoff. In addition to the choice of bandwidth, the kernel function also governs the degree of smoothness. It is clear that in K_h, the smoother, should be used to replace the indicator function in the formulation of $F_t(x)$. Once we choose the Gaussian kernel $K_h = exp(-u^2/2)/\sqrt{2\pi}$ as the kernel for estimating VaR, Silverman's (1984) rule of thumb can be applied to obtain an optimal bandwidth, $\hat{h}_{rot,t}$, expressed as:

$$\hat{h}_{rot,t} = 1.06min\left\{\hat{\sigma}_t, \frac{Q_t}{1.34}\right\}n^{-\frac{1}{5}}, \tag{A.2}$$

where $\hat{\sigma}_t$ is the sample standard deviation estimated from $\{r_{t,i}\}_{i=1}^{n}$ and $Q_t = r_{t,[0.75n]} - r_{t,[0.25n]}$. (Eq. A.2) indeed takes into account the sensitivity of outliers, since a single outlier may cause too large an estimate of $\hat{\sigma}_t$ and hence may create too large a bandwidth. The interquartile range Q_t is invoked here to compensate for this effect. The constants 1.34 and 1.06 are scaling factors that are related to the choice of kernel (see Härdle et al., 2004).

The question arises of how to bootstrap from the KDE, $\hat{F}_{t,h}(x)$. It turns out that one doesn't have to simulate it via an inversion or rejection technique from (Eq. A.1), since smoothing of the distribution function can be interpreted as a convolution of the empirical distribution function with the kernel K_h. The use of a convolution operator to calculate the convolution of a sum of two random variables also allows us to simply view Eq. (A.1) as the integrated probability density function of the sum of $r_{t,i}$ with a random variable Z having probability density function K_h. To be more explicit, given month t, we bootstrap 1000 times from $\{r_{t,i}\}_{i=1}^{n}$. For each bootstrapped sample from $\{r_{t,i}\}_{i=1}^{n}$, we merely add the product of $\hat{h}_{rot,t}$ in Eq. (A.1.2) and 1000 generated random variables Z from i.i.d $N(0,1)$. This idea can be expressed as:

$$r^*_{t,i} = r_{t,i} + \hat{h}_{rot,t}Z \tag{A.3}$$

The $V_{i,t}$ estimate of country i at month t can now be obtained by calculating the 1% quantile of the simulated distribution $r^*_{t,i}$. The corresponding expected shortfall is the mean loss exceeding the VaR value. The kernel density technique here achieves our goal of a 1%-quantile value from limited observations, which builds on Bali et al. (2009), who regard the minimum daily return within the given month as the VaR. This minimum daily return is actually around a 4%- to 5%-quantile value over 22 daily returns and does not seem to be too extreme. In fact, it is rather unrealistic to produce a 1%-expected shortfall estimator under this condition.

[16] A legitimate argument can be made that a small sample bias might be arrived at via a simulation. It can be demonstrated (though not reported here to save space) that the fractional differentiation framework employed in this study is robust relative to the bias. The report is available upon request.

References

Ang, A., Bekaert, G., Liu, J., 2005. Why stocks may disappoint. J. Financ. Econ. 76, 471–508.

Ang, A., Hordrick, R.J., Xing, Y., Zhang, X., 2006. The cross-section of volatility and expected returns. J. Finance 259–299.

Artzner, P., Delbaen, F., Eber, J.M., Heath, D., 1999. Coherent measures of risk. Math. Finance 9, 203–228.

Bali, T.G., Cakici, N., 2010. World market risk, country-specific risk and expected returns in international stock markets. J. Bank. Finance 34, 1152–1165.

Bali, T.G., Demirtas, K.O., Levy, H., 2009. Is there an intertemporal relation between downside risk and expected returns. J. Financ. Quant. Anal. 44, 883–909.

Bali, T.G., Peng, L., 2006. Is There a Risk-return Tradeoff? Evidence from high frequency data. J. Appl. Econ. 21, 1169–1198.

Bandi, F.M., Perron, B., 2008. Long-run risk-return trade-offs. J. Econometr. 143, 349–374.

Bekaert, G., Harvey, C.R., 1995. Time-varying world market integration. J. Finance 50, 403–444.

Bekaert, G., Wu, G., 2000. Asymmetric volatility and risk in equity markets. Rev. Financ. Stud. 13, 1–42.

Bollerslev, T., Litvinova, J., Tauchen, G., 2006. Leverage and volatility feedback effects in high-frequency data. J. Financ. Econometr. 4, 353–384.

Bollerslev, T., Osterrieder, D., Sizova, N., Tauchen, G., 2013. Risk and return: long-run relations, fractional cointegration and return predictability. J. Financ. Econ. 108, 409–424.

Boudoukh, J., Richardson, M., Whitelaw, R., 2008. The myth of long-horizon predictability. Rev. Financ. Stud. 21, 1577–1605.

Caporale, G.M., Cipollini, A., Spagnolo, N., 2005. Testing for contagion: a conditional correlation analysis. J. Empir. Finance 12, 476–489.

Caporin, M., 2008. Evaluating value-at-risk measures in presence of long memory conditional volatility. J. Risk 10, 79–110.

Chen, C.Y.-H., Chiang, T.C., 2016. Empirical analysis of the intertemporal relation between downside risk and expected returns: evidence from time-varying transition probability models.". Eur. Financ. Manag. 22, 749–796.

Chiang, T.C., Jeon, B.N., Li, H., 2007. Dynamic correlation analysis of financial contagion: evidence from Asian markets. J. Int. Money Finance 26, 1206–1228.

Cont, R., Deguest, R., Scandolo, G., 2010. Robustness and Sensitivity Analysis of Risk Measurement Procedures. Quant. Finance 10, 593–606.

Cornish, E.A., Fisher, R.A., 1937. Moments and cumulants in the specification of distribution. Rev. Int. Stat. Inst. 5, 307–320.
Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. J. Busin. Econ. Stat. 13, 253–263.
Diebold, F.X., Yilmaz, K., 2009. Measuring financial asset return and volatility spillovers, with application to global equity markets. Econ. J. 119, 158–171.
Dolatabadi, S., Nielsen, M.Ø., Xu, K., 2015. A fractionally cointegrated VAR analysis of price discovery in commodity futures markets. J. Futures Markets 35, 339–356.
Dowd, K., 2001. Estimating VaR with order statistics. J. Deriv. 8, 23–30.
Ferson, W., Nallareddy, S., Xie, B., 2013. The "out-of-sample" performance of long run risk models. J. Financ. Econ. 107, 537–566.
Forbes, K., Rigobon, R., 2002. No Contagion, only interdependence: measuring stock market co-movements. J. Finance 57, 2223–2261.
French, K.R., Schwert, G.W., Stambaugh, R.F., 1987. Expected stock returns and volatility. J. Financ. Econ. 19, 3–29.
Geweke, J., Porter-Hudak, S., 1983. The estimation and application of long memory time series models. J. Time Ser. Anal. 4, 221–238.
Giacomini, R., White, H., 2006. Tests of conditional predictive ability. Econometrica 74 (6), 1545–1578.
Ghysels, E., Santa-Clara, R., Valkanov, R., 2005. There is a risk-return tradeoff after all. J. Financ. Econ. 76, 509–548.
Glosten, L., Jagannathan;, R., Runkle, D., 1993. On the relation between the expected value and volatility of the nominal excess return on stocks. J. Finance 48, 1779–1801.
Gul, F., 1991. A theory of disappointment aversion. Econometrica 59, 667–686.
Härdle, K.W., Müller, M., Sperlich, S., Werwatz, A., 2004. Nonparametric and Semiparametric Models. Springer Series in Statistics.
Harrision, P., Zhang, H.H., 1999. An investigation of risk and return relation at long horizons. Rev. Econ. Stat. 81, 399–408.
Harvey, C.R., Liechty, J.C., Liechty, M.W., Muller, P., 2010. Portfolio selection with higher moments. Quant. Finance 10, 469–485.
Harvey, C.R., Siddique, A., 2000. Conditional skewness in asset pricing tests. J. Finance 55, 1263–1295.
Jarrow, R., Zhao, F., 2006. Downside loss aversion and portfolio management. Manag. Sci. 52, 558–566.
Johansen, S., 2008. Representation of cointegrated autoregressive processes with application to fractional processes.". Econometr. Rev. 28, 121–145.
Johansen, S., Nielsen, M.O., 2012. Likelihood inference for a fractionally cointegrated vector autoregressive model. Econometrica 80, 2667–2732.
Kinateder, H., Wagner, N., 2014. Multiple-period market risk prediction under long memory: when VaR is higher than expected. J. Risk Finance 15, 4–32.
King, M., Wadhwani, S., 1990. Transmission of volatility between stock markets. Rev. Financ. Stud. 3, 5–33.
Karolyi, G.A., Stulz, R.M., 1996. Why do markets move together? An investigation of US–Japanese stock return comovements. J. Finance 51, 951–986.
Künsch, H.R., 1986. Discrimination between monotonic trends and long range dependence. J. Appl. Probab. 23, 1025–1030.
Lambert, M., Hübner, G., 2013. Comovement risk and stock returns. J. Empir. Finance 23, 191–205.
Lettau, M., Ludvigson, S.C., 2010. Measuring and modeling variation in the risk-return trade-off. In: Aït-Sahalia, Y., Hansen, L., Scheinkman, J.A. (Eds.), Handbook of Financial Econometrics. Amsterdam, North Holland.
Lundblad, C., 2007. The risk return tradeoff in the long run: 1836–2003. J. Financ. Econ. 85, 123–150.
Maynard, A., Smallwood, A., Wohar, M.E., 2013. Long memory regressors and predictive regressions: a two-stage rebalancing approach. Econometr. Rev. 32, 318–360.
Merton, R.C., 1973. An intertemporal capital asset pricing model. Econometrica 41, 867–887.
Nelson, D., 1991. Conditional heteroskedasticity in asset returns: a new approach. Econometrica 59, 347–370.
Qu, Z.A., 2011. Test against spurious long memory. J. Busin. Econ. Stat. 29, 423–438.
Routledge, B.R., Zin, S.E., 2010. Generalized disappointment aversion and asset prices. J. Finance 65, 1303–1332.
Scruggs, J.F., 1998. Resolving the Puzzling Intertemporal relation between the market risk premium and conditional market variance: a two-factor approach. J. Finance 53, 575–603.
Silverman, B.W., 1984. Spline Smoothing: the equivalent variable kernel method. Annal. Stat. 12, 898–916.
Whitelaw, R.F., 1994. Time variations and covariations in the expectation and volatility of stock market returns. J. Finance 49, 515–541.
Whitelaw, R.F., 2000. Stock market risk and return: an equilibrium approach. Rev. Financ. Stud. 13, 521–547.

Macroeconomic Dynamics, 2018, Page 1 of 21. Printed in the United States of America.
doi:10.1017/S1365100518000482

A NOTE ON THE IMPACT OF NEWS ON US HOUSEHOLD INFLATION EXPECTATIONS

Ben Zhe Wang, Jeffrey Sheen, and Stefan Trück
Macquarie University

Shih-Kang Chao
University of Missouri

Wolfgang Karl Härdle
Humboldt-Universität zu Berlin
and
Singapore Management University

Monthly disaggregated US data from 1978 to 2016 reveal that exposure to news on inflation and monetary policy helps to explain inflation expectations. This remains true when controlling for household personal characteristics, perceptions of government policy effectiveness, expectations of future interest and unemployment rates, and sentiment. We find an asymmetric impact of news on inflation and monetary policy after 1983, with news on rising inflation and easier monetary policy having a stronger effect in comparison to news on lowering inflation and tightening monetary policy. Our results indicate the impact on inflation expectations of monetary policy news manifested through consumer sentiment during the lower bound period.

Keywords: Inflation Expectations, News Impact, Monetary Policy place with Communication

1. INTRODUCTION

Inflation expectations play a major role in modern macroeconomics, with rational expectations ubiquitous as the modeling device for a representative agent. However, the literature provides both theoretical models and empirical observations that can explain how different economic agents form inflation expectations and why they might disagree on their forecasts. For example, Mankiw et al. (2004) document a considerable degree of disagreement in surveys of US inflation expectations. This disagreement is time-varying and exhibits covariation with macroeconomic variables. Mankiw and Reis (2002) construct a formal model and

We thank the editor and the two anonymous referees for their constructive comments. Address correspondence to: Ben Zhe Wang, 4ER 432, Department of Economics, Macquarie University, North Ryde, 2109, NSW, Australia; e-mail: ben.wang@mq.edu.au. Phone: +61 2 98508500.

 1469-8056/18

attribute disagreements to information rigidity. The idea is that the dissemination of new information occurs gradually between people.

One way households acquire information is through media reports, which we refer to as "news" in this paper. News can directly impact on household inflation expectations by directly informing the consumer about the possible future path of inflation (e.g. through expert forecasts), or indirectly through impacting on household perceptions of current inflation. Lamla and Maag (2012) find that the disagreement in household inflation expectations in Europe depends on the reporting intensity and the "tone" of the news about inflation, while Dräger (2015) finds that the media has a small but significant impact on inflation expectations in Sweden. Carroll (2003) uses an epidemiology model and finds that professional forecasts as a proxy for news have predictive power for household forecasts in the USA.

All the aforementioned studies use aggregated news measures obtained from a separate source than that for the measure of inflation expectations. One drawback with this approach is that the news measures do not necessary reflect the news heard by the individual household, and thus may not necessarily be attributable to household inflation expectation formation. In this paper, we use the Michigan Survey of Consumers from 1978 to 2016, which allows us to examine the direct impact of news on individual households.

There is an emerging literature on investigating the effect of perceived news using the Michigan Survey of Consumers data. For example, utilizing the panel structure of the Michigan Survey of Consumer data,[1] Pfajfar and Santoro (2013) test the epidemiology model of Carroll (2003) using an aggregate measure of news and household perceived news, and find at best weak support for the epidemiology model. Although hearing inflation news increases the probability of updating inflation expectations, it enlarges the forecast gap between households' inflation expectation and those of professional forecasts, as well as the gap between households' inflation expectation and actual realized inflation. Similarly, Dräger and Lamla (2017) find the hearing of news on inflation increases the chance of households updating their inflation expectations, irrespective of whether it is favorable or unfavorable news. Pfajfar and Santoro (2009) find households with different socioeconomic background form inflation expectation differently in response to inflation news, and they exhibit different degrees of information stickiness when updating their inflation expectations. In addition, Ehrmann et al. (2017) find households tend to forecast inflation higher if they have financial difficulties or are pessimistic about major purchases, income developments, or the unemployment rate—however, their bias shrinks by more than the average household in response to inflation news. Lahiri and Zhao (2016) also find consumer sentiment responds to perceived news and Zhang et al. (2016) find stock markets react to news through its impact on sentiment.

In this paper, we contribute to the literature by considering monetary policy news along with inflation news and evaluate whether favorable or unfavorable

news has asymmetric impacts on household inflation expectations. Unlike Pfajfar and Santoro (2013) and Dräger and Lamla (2017), we do not restrict ourselves to using the panel structure of the Michigan data to investigate how inflation expectations are updated according to news, but rather examine how inflation expectations are formed in general.

Our results using the Michigan Survey from 1978 to 2016 show that households raise their inflation expectations when they are exposed to news of rising inflation and of contractionary monetary policy. The latter result is an indication that monetary policy acts as a signaling device for the formation of inflation expectations.[2] Our results are robust after controlling for household demographics, their perception of the effectiveness of government policies, their expectations of future interest and unemployment rates, and their sentiment. We also find an asymmetric impact of news on rising inflation (contractionary monetary policy) compared to news on falling inflation (expansionary monetary policy). Our results indicate that this asymmetric impact started to become significantly stronger in the early 1990s. We find that the absolute impact of news on higher inflation became statistically greater than news on lower inflation after 1991, while after 1999 news on easing monetary policy had a significantly greater impact on inflation expectations than contractionary monetary policy. Finally, during the zero lower bound period after 2008, news about monetary policy becomes an imperfect signal for inflation expectations formation. This signal manifested through consumer sentiment, which implies central banks should pay attention to consumer sentiment when communicating monetary policies.

The subsequent paper is organized as follows. Section 2 describes the applied model and the data used. Section 3 examines the impact of news on households inflation expectations. Section 4 tests if the content of news has an asymmetric impact on inflation expectations, while Section 5 examines the news effects during the zero lower bound period. Section 6 concludes.

2. THE MODEL AND DATA

Since 1978, around 500 adults in households have been surveyed each month on their 1-year-ahead inflation expectations by the University of Michigan (Survey of Consumers). The survey asks respondents to provide a numerical answer to the following question:

> By about what percent do you expect prices to go (up/down) on the average, during the next 12 months?

The data exhibit a considerable degree of disagreement among these US households in any month. In addition to inflation expectations, the survey also asks respondents whether they have heard news about current economic conditions, and also for their evaluations of current and expected future paths of the economy as well as their personal financial situation.

We test if news plays a role in explaining household inflation expectations by estimating equation (1) using pooled ordinary least squares:

$$\pi_{it}^{e} = \alpha + TD_t\theta' + \phi^{\pi} N_{it}^{\pi} + \phi^{r} N_{it}^{r} + C_{it}\gamma' + \epsilon_{it}, \quad (1)$$

where π_{it}^{e} is the 1-year ahead inflation expectation of household i at time t, α is a constant, and TD_t collects monthly time dummies that are invariant among households at a given month. Since our focus is on investigating the impact of news on individual inflation expectations, we include these time dummies to account for aggregate developments of the economy in each month that might have an impact on household inflation expectations.

N_{it}^{π} and N_{it}^{r} indicate whether household i has been exposed to any news of inflation and monetary policy, respectively. The survey asks respondents to indicate whether they have heard news of changes in business conditions:

> During the last few months, have you heard of any favorable or unfavorable changes in business conditions? What did you hear?

The respondents may indicate they have heard news on rising or falling prices, which we use to approximate inflation news, and lower or higher interest rates or easier or tighter credit conditions, which we use to approximate monetary policy news.

If no particular news has been heard, the respective variable has a value of 0; N_{it}^{π} is set to a value of 1 if household i has been exposed to news about higher inflation, and -1 in the case of news about lower inflation. In the same manner, N_{it}^{r} takes on a value of 1 if household i has heard news about higher interest rates or tighter credit conditions, mostly associated with tighter monetary policy, and -1 for exposure to news about lower interest rates or easier credit conditions, mostly associated with expansionary monetary policy. ϕ^{π} and ϕ^{r} measure the impact of inflation news and monetary policy news on household inflation expectations, which is a key focus of this paper.

$C_{it} \in [D_{it}, P_{it}, E_{it}, CS_{it}]$ represents control variables for the characteristics of household i. Hereby, D_{it} denotes economic and demographic variables for respondent i, including log income, age, gender (1 for a female), and level of education (measured on a scale between 1 and 6, with 6 indicating the highest level of education).[3] P_{it} denotes household perceptions on the effectiveness of government policies in managing inflation or unemployment, taking a value of 1 if the government is perceived to have done a good job, 0 for a fair job, and -1 for a poor job.[4] E_{it} collects household expectations on the future course of interest rates and unemployment over the next year, with 1 indicating that household i expects the future respective rate will increase, 0 that it stays the same, and -1 indicates that the household expects the rate to be reduced.[5] CS_{it} is the measure from the Michigan Survey for consumer sentiment. It is constructed from five qualitative questions about the household's current and future expected personal financial situation, its current buying attitude regarding large ticket household items, and its expectation of short- and medium-term business conditions.[6] The Appendix

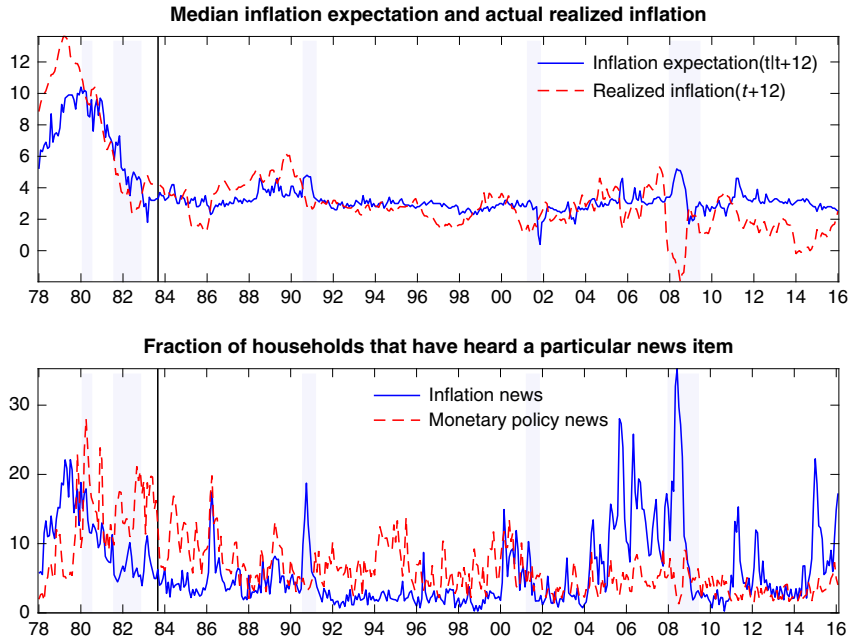


FIGURE 1. The top panel shows the median 12-month ahead household inflation expectation and the realized 12-month ahead inflation. The bottom panel shows the fraction of households that have heard inflation news or monetary policy news.

shows the pairwise correlations among the explanatory variables. Apart from log income with education, the perception of government policies with sentiment and unemployment expectation, and sentiment with unemployment expectation, all explanatory variables are not highly correlated.

Our monthly sample starts from January 1978 and ends in February 2016, containing 208,777 individual records in total. The cross-sectional inflation expectations data range from 0% to (a cap at) 50%. We follow the literature, see, for example, Curtin (1996), and restrict our sample to those respondents who gave inflation expectations below 30%, on the grounds that such outliers are likely to be frivolous. Since our sample covers the high inflation episode in the late 1970s, and there are possible structural breaks in the inflation expectations series, we test for structural breaks in median household inflation expectations following Bai and Perron (2003). The results suggest a structural break in median household inflation expectations in September 1983 so we split our sample into pre- and post-September 1983 periods.

The top panel of Figure 1 shows the 12-month ahead median household inflation expectations (blue solid line) and the actual realized 12-month ahead inflation (red dashed line), with shaded areas indicating NBER-dated recessions and the vertical line showing our structural break date. Both actual and expected inflation were high in the late 1970s but gradually decreased during the two recessions

in the early 1980s. Both remained relatively low throughout the 1990s and early 2000s.[7] It is interesting that households on average also expected higher inflation during and after the financial crisis of 2008. These expectations remained much greater than realized inflation, while deflation was likely more of a concern to policymakers.

The bottom panel shows the fraction of households that have heard inflation news (blue solid line) or monetary policy news (red dashed line). As illustrated by the figure, the fraction of households who have heard inflation news is quite volatile. As proposed by Ehrmann et al. (2017), this fraction is often driven by people who have heard that prices are higher. The authors also find the fraction to be highly correlated with retail gasoline price inflation in general, suggesting that frequently purchased items shape households' inflation (news) perceptions. The spikes in the series could be related to economic recessions or actual low or high inflation rates. For example, the high percentage of households who had heard inflation news in the period March–May 1986 can be explained by the fact that in 1986 inflation rates had reached levels below 2% for the first time in 20 years, a situation that was frequently discussed in the media. The spike in the series in September–November 1990 is most likely a result of the USA entering into recession in July 1990, lasting until March 1991. The recession was at least partially related to the restrictive monetary policy enacted by the Federal Reserve throughout 1989 and 1990, when the stated policy was to reduce inflation. The high fraction of households who had heard inflation news in the early and mid-2000s was typically related to news about higher prices. For example, during the first 3 months of 2004, consumer prices increased at a seasonally adjusted annual rate of over 5%, which was much higher than in previous years. In September 2005, the consumer price index had also risen again by almost 5% in comparison to 12 months earlier. Notably, a significant part of both increases could be related to rising energy costs, an issue often reported in the news around that time. The fraction of households that had heard news about inflation was also extremely high in the second half of 2008. This is most likely related to the subprime mortgage crisis and the fact that both households and the media paid far more attention to news about macroeconomic and financial conditions during that period.

Generally, most of the spikes are due to periods where people had heard news on higher prices, except for three episodes: between March and May 1986, when inflation rates below 2%; between December 2014 and February 2015, when the US economy deflated for the first time since 2009; and from January 2016 to the end of the sample, when a general discussion on the risks of a prolonged deflation period became more prevalent in the media. It is interesting that news of lower inflation were rare and became prevalent only near the end of the sample, even though a fear of deflation had become a widespread concern among policymakers and commentators after 2008.

Interestingly, the fraction of households that heard monetary policy news has remained low since the 2000s, and this is true even during the recent global financial crisis when the Federal Reserve had used extensively unconventional monetary policies.[8]

3. NEWS IMPACT ON HOUSEHOLD INFLATION EXPECTATIONS

We test if exposure to direct news of inflation and monetary policy affects household inflation expectations. Model 1 considers only the impact of news on inflation and Model 2 considers both news on inflation and monetary policy without controlling for characteristics of household i, thus serving as a benchmark. The subsequent five models extend the benchmark specification: Model 3 controls for the additional impact of demographic characteristics D_{it}; Model 4 controls for the additional impact of perceptions of the effectiveness of government policies P_{it}; Model 5 controls for the additional impacts of expectations about interest rates and unemployment; Model 6 controls for the additional impacts of consumer sentiment on inflation expectations ICS_{it}; and Model 7 considers all explanatory variables jointly. We report results in Table 1, with the top panel of the table showing the results for January 1978 to September 1983 and the middle panel focusing on October 1983 to February 2016. Since we have 457 monthly time dummies and their interpretations do not necessarily relate to news effects, we omit results for the time dummies in the table.[9]

3.1. The First Subsample—January 1978 to September 1983

Results for the first subsample show news of inflation and monetary policy having a strong impact on household expectations. In this relatively high inflation period, Model 1 shows that hearing news of higher inflation led an average household to increase their inflation expectations by 1.03%. When jointly considered with news of monetary policy, the impact of news of inflation reduces to 0.98%.

News of monetary policy changes can affect inflation expectations in two opposite ways. One is if households understand the *transmission mechanism* of monetary policy to future inflation, in which case news of tighter monetary policy implies lower expected future inflation. The other is if households do not understand the transmission mechanism but understand that the central bank targets inflation, in which case news of contractionary monetary policy is a *signal* that inflation is higher than previously expected. Model 2 shows that hearing news of contractionary monetary policy induced households to expect 0.49% higher inflation, indicating households understood that higher interest rates are a result of the central bank's concern about higher future inflation[10]. Therefore, the average household expectations *appear* to be informed as a signal by the central bank's response function, for example, a Taylor rule [as suggested by Carvalho and Nechio (2014)], rather than households concerning themselves with the expected future contractionary effect of the interest rate change on inflation.[11] The results from Models 1 and 2 are consistent with earlier findings using *aggregate* data from the media having a role in driving household inflation expectations [Lamla and Maag (2012) and Dräger (2015)]. Our results are also consistent with the information rigidity hypothesis, which suggests that households' private information sets (through news) play a role in explaining disagreements in inflation expectations [Mankiw and Reis (2002) and Madeira and Zafar (2015)].

TABLE 1. Regression results for inflation expectations

	Subsample 1: 1978:01–1983:09						
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Constant	4.45***	4.48***	4.43***	4.56***	4.47***	6.08***	4.04***
News: inflation(ϕ^{π})	1.03***	0.98***	0.94***	0.83***	0.83***	0.87***	0.67***
News: monetary policy(ϕ^{r})	–	0.49***	0.47***	0.37***	0.35***	0.40***	0.22***
Log income	–	–	0.07	–	–	–	0.21***
Age	–	–	−0.03***	–	–	–	−0.03***
Female	–	–	0.26***	–	–	–	0.06
Education	–	–	0.15***	–	–	–	0.19***
Perception: government policy	–	–	–	−1.08***	–	–	−0.73***
Expectation: interest rate	–	–	–	–	0.82***	–	0.72***
Expectation: unemployment rate	–	–	–	–	0.82***	–	0.48***
Consumer sentiment	–	–	–	–	–	−0.02***	−0.01***
Adjusted-R^2	0.623	0.624	0.628	0.629	0.633	0.628	0.642

TABLE 1. Continued

	Subsample 2: 1983:10–2016:02						
Constant	3.21***	3.19***	9.04***	3.08***	2.92***	4.92***	8.87***
News: inflation(ϕ^{π})	0.64***	0.62***	0.61***	0.52***	0.50***	0.43***	0.39***
News: monetary policy(ϕ^{r})	–	0.35***	0.32***	0.26***	0.18***	0.17***	0.08**
Log income	–	–	−0.45***	–	–	–	−0.35***
Age	–	–	−0.01***	–	–	–	−0.01***
Female	–	–	0.66***	–	–	–	0.54***
Education	–	–	−0.14***	–	–	–	−0.11***
Perception: government policy	–	–	–	−0.75***	–	–	−0.33***
Expectation: interest rate	–	–	–	–	0.47***	–	0.40***
Expectation: unemployment rate	–	–	–	–	0.73***	–	0.35***
Consumer sentiment	–	–	–	–	–	−0.02***	−0.01***
Adjusted-R^2	0.520	0.521	0.532	0.529	0.532	0.535	0.549

Notes: 1. The sample size is 43,599 between January 1978 and September 1983 and 165,178 between October 1983 and February 2016.
2. *, **, and *** represent significance at 10%, 5%, and 1% levels of significance.

Model 3 confirms results in the earlier literature that households with different demographic backgrounds disagree on inflation expectations—see, for example, Bryan and Venkatu (2001*a*,b) for the USA, Blanchflower and MacCoille (2009) for the UK, Easaw et al. (2013) for Italy, and Jonung (1981) for Sweden. Households with different demographics may purchase different consumption bundles. In particular, we find that those who are younger, female, and better educated tended to forecast higher inflation levels. Households may form inflation expectations according to their lifetime inflation experience. Therefore, younger people expected much higher inflation following the years of high inflation in the 1970s. Indeed Malmendier and Nagel (2016) find households with members older than 70 years expected lower inflation compared to households with members younger than 40 years in the 1970s. It is well documented that women on average forecast higher inflation than men. One possibility is that women have higher perceived inflation [e.g. Jonung (1981)] due to being likely responsible for grocery shopping and thus exposed more often to prices than men. The impact of education on inflation forecasts is interesting. As seen in Figure 1, median inflation expectations underestimate actual inflation in the higher inflation period of the first subsample (and overestimate inflation in the lower and more stable inflation period of the second subsample). The fact that better educated households forecast higher inflation in the first subsample (and lower inflation in the second subsample) indicates they are better forecasters than the median household. Better educated households thus appear to have better understood the severity of the implications of oil price shocks on inflation in the 1970s, realizing that monetary policy would need to be strategically accommodative to minimize the effects of the rise in the relative oil price. This accommodative monetary policy did not increase the nominal interest rate more than inflation, thus reducing the real interest rate and making monetary policy expansionary, as argued by Clarida et al. (2000). A similar logic may apply to households with higher income, if we assume that these households are more likely to participate in financial markets, and thus tend to be better informed.

Perceiving government policies to be effective in managing the business cycle significantly reduced inflation expectations by 1.08% (Model 4). Households that heard tightening monetary policy news and who believe government policies are effective will form 0.71 (0.37−1.08)% lower inflation expectation compared to the average household, whereas households hearing the same news but who do not believe policies are effective would expect 1.45 (0.37 + 1.08)% higher inflation than the average household. This appears to indicate that households who believe government policies are effective tend to forecast inflation consistent with the transmission mechanism of monetary policy rather than being informed as a signal by monetary policy. This result also implies that if the government wants to lower inflation, it can reduce inflation expectations by influencing household perceptions about the effectiveness of its policies.

Expecting a rise in interest rates over the next year was associated with higher inflation expectations (Model 5). This confirms the findings in Model 2

and suggests that households understood monetary policy responds to inflation now and in the near future, so that higher expected inflation is associated with higher current and expected future interest rates. Also higher expected future unemployment was associated with higher inflation expectations. These results suggest that the average household seemed not to be concerned with the implied negative correlation between expected inflation and expected unemployment of an expectations-augmented Phillips curve, instead associating higher expected inflation with a higher expected future unemployment rate.

The negative and significant consumer sentiment parameter estimate (Model 6) shows that more optimistic households expected slightly lower inflation than the average. This result indicates that households' perception and sentiment—reflecting their interpretation of their private information set—help to explain why they disagreed in their inflation expectations. This parameter estimate is robust to the sample used.

Model 7 includes all regressors and shows that the impact of news about inflation and monetary policy, perceptions on the effectiveness of government policies, expectations of future interest rates and the unemployment rate, and consumer sentiment were all important factors for explaining the heterogeneity of inflation expectations. Adding perceptions on government policies, expectations and sentiment induced a magnification of the impacts of income, while gender had a lessened impact. Therefore, in the first subsample, those who were richer tended to expect higher inflation, owing to their perceptions, expectations, and sentiment being less positive than those of poorer households.

3.2. The Second Subsample—October 1983 to February 2016

Many of the aforementioned results remain true in the second subsample, but there are some notable differences. First, the impacts on inflation expectations of news about both inflation and monetary policy were smaller, although they remain significant. This lower impact of news in the second subsample reflects the fact that inflation had fallen and stabilized during this period, making news of inflation and monetary policy less salient for households, and thus reducing their impact on inflation expectations. Second, the signs of the impact of household income changed to be negative and was much larger in absolute size. Third, the sign on education also reversed so that now better than average educated households expected lower inflation. These two sign reversals mean that households with higher income and better education forecasted lower inflation than the average household in this subperiod. Finally, gender played a much larger role, with the difference between male and female expectations becoming significantly larger.

In summary over the whole sample, our results indicate that exposure to news of higher inflation and contractionary monetary policy significantly increased household inflation expectations. This result is robust across sample periods and holds even after controlling for household demographic characteristics, their perceptions on the effectiveness of government policies, their expectations about interest

rates and unemployment, and their sentiment. Among macroeconomic theories, information rigidity models have been widely used to explain cross-sectional disagreements of inflation expectations—see, for example, Mankiw and Reis (2002) and Mankiw et al. (2004). These models typically assume information is costly to acquire, so people have a different information set when forecasting future paths of the economy. Our results indicate that households who had a larger news exposure expected different inflation rates (*ceteris paribus*), thus supporting the information rigidity theory. The fact that the estimated news effect (ϕ^{π} and ϕ^{r}) between Models 2 and 3 are very similar suggests that household demographics and news almost independently explain inflation expectations. This means that the demographic impacts on inflation expectations were not due to the different demographic groups' exposure to news.

Controlling for the perception of the effectiveness of government policies, for expectations on future interest and unemployment rates and for consumer sentiment reduces the impact of news on household inflation expectations. These findings suggest that news of inflation and monetary policy impacted on inflation expectations partially through these household perceptions about policy effectiveness, their expectations of future interest rates and unemployment, and their sentiments about current economic conditions.

4. THE ASYMMETRIC IMPACT OF NEWS

News on the movements of underlying economic variables may have an asymmetric impact on inflation expectations. This may arise if one particular direction of movement of the variable has a more salient effect on expectations than the other at the time of making the expectation decision. For example, due to diminishing marginal utility, higher inflation can erode household wealth and reduce utility more than it would increase it, if inflation fell by the same amount—households may thus pay more attention to news of higher inflation than of lower inflation. Households may also have experienced the high inflation episodes in the 1970s and understand high inflation may indicate unsuccessful policies and have long-lasting effects on future paths of inflation [Madeira and Zafar (2015)] compared to lower inflation. Thus, it may be reasonable to assume a bigger impact of high inflation on future inflation expectations. Using aggregated data, Lamla and Maag (2012) and Dräger (2015) indeed find the content of media reports have an asymmetric impact on inflation expectations.

Utilizing the cross-sectional nature of the Michigan inflation expectations and news data, we investigate whether news content has an asymmetric impact on household inflation expectations at the disaggregated level. For each of the news variables N_{it}^{π} and N_{it}^{r} considered, we construct two dummy variables according to the content of the news. An upward arrow $\uparrow$ denotes news that corresponds to an increasing value of the underlying variable, while a downward arrow $\downarrow$ relates to news decreasing the value of the underlying variable. For example, news of rising inflation would result in a value of 1 for $N_{it}^{\pi}\uparrow$ and a value 0 for $N_{it}^{\pi}\downarrow$; $N_{it}^{r}\downarrow = 1$

TABLE 2. Regression results for asymmetric news impacts

	Subsample 1 1978:01–1983:09	Subsample 2 1983:10–2016:02
Constant	4.07***	8.84***
News: lower inflation($\phi^{\pi}_{\downarrow}$)	−0.49**	−0.22***
News: higher inflation($\phi^{\pi}_{\uparrow}$)	0.70***	0.46***
News: easing monetary policy($\phi^{r}_{\downarrow}$)	−0.08	−0.15***
News: tightening monetary policy($\phi^{r}_{\uparrow}$)	0.31***	−0.00
Perception: government policy	−0.73***	−0.33***
Expectation: interest rate	0.72***	0.40***
Expectation: unemployment rate	0.48***	0.35***
Consumer sentiment	−0.01***	−0.01***
Log income	0.20***	−0.35***
Age	−0.03***	−0.01***
Female	0.07	0.54***
Education	0.19***	−0.11***
Adjusted-R^2	0.642	0.549
Hypothesis: $\phi^{\pi}_{\downarrow} = -\phi^{\pi}_{\uparrow}$	0.95	2.72***
Hypothesis: $\phi^{r}_{\downarrow} = -\phi^{r}_{\uparrow}$	1.37	−1.96*

Notes: 1. Subsample 1 is from January 1978 to September 1983, with sample size 43,599. Subsample 2 is between October 1983 and February 2016, with sample size 165,178.
2. *, **, and *** represent significance at 10%, 5%, and 1% levels of significance.
3. The estimation results for the monthly time dummy are omitted from the table.

indicates news about easing monetary policy and $N^{r}_{it}\uparrow = 1$ indicates news about contractionary monetary policy. We thus replace N^{π}_{it} and N^{r}_{it} in equation (1) with $N^{\pi}_{it}\downarrow, N^{r}_{it}\downarrow$ and $N^{\pi}_{it}\uparrow, N^{r}_{it}\uparrow$:

$$\pi^{e}_{it} = \alpha + \phi^{\pi}_{\downarrow} N^{\pi}_{it}\downarrow + \phi^{\pi}_{\uparrow} N^{\pi}_{it}\uparrow + \phi^{r}_{\downarrow} N^{r}_{it}\downarrow + \phi^{r}_{\uparrow} N^{r}_{it}\uparrow + C_{it}\gamma' + TD_t\theta' + \epsilon_{it}. \quad (2)$$

For $j \in \{\pi, r\}$, we expect $\phi^{j}_{\downarrow}$ to have the opposite impact on inflation expectations to $\phi^{j}_{\uparrow}$. We are interested in testing whether or not increases and decreases have the same absolute impact on inflation expectations. To test this, we calculate the z score between the two estimated parameters and test whether the null hypothesis $\phi^{j}_{\downarrow} = -\phi^{j}_{\uparrow}$ is rejected:

$$z^{j} = \frac{\phi^{j}_{\downarrow} - (-\phi^{j}_{\uparrow})}{\sqrt{\mathrm{Var}(\phi^{j}_{\downarrow} - (-\phi^{j}_{\uparrow}))}}. \quad (3)$$

Table 2 shows the estimation results of equation (2) with the columns giving the estimation results for the two subsamples. The results are broadly consistent with those presented in Table 1. Consistent with our expectations, the two directions of news content had opposite effects on household inflation expectations. This result is robust across both inflation and monetary policy news and across sample periods.

Both news of rising and declining inflation had significant impacts on household inflation expectations across both subsamples. Hearing news of higher inflation in the first subsample increased inflation expectations by 0.70% on average, and hearing news of lower inflation reduced inflation expectations by 0.49% on average in this high inflation period. Hearing news on higher inflation in the second subsample increased household inflation expectations by 0.46% on average, but being exposed to news on lower inflation only reduced inflation expectations by 0.22%. This result indicates that households respond to news on higher inflation more than to news on lower inflation in general, though the effect of inflation news was much weaker in the second subsample. This is especially true for news on lower inflation, where the impact is more than halved in the second subsample. The second row of the lower panel of Table 2 shows the significance of the z-score test of equation (3) in terms of inflation news ($\phi^{\pi}_{\downarrow} = -\phi^{\pi}_{\uparrow}$). Consistent with previous results, the test indicates that the symmetric effect of news on inflation cannot be rejected for the first subsample period, while the effect became significantly asymmetric in the second subsample, where news of higher inflation had a bigger absolute impact than news of lower inflation.

News on easing monetary policy did not significantly alter household inflation expectations in the high inflation period (Subsample 1), but significantly reduced inflation expectations in the second subsample. On the other hand, news of tightening monetary policy significantly increased household inflation forecasts in the first subsample but was irrelevant in the second subsample.[12] The third row of the lower panel of Table 2 shows the significance of the z-score test of $\phi^{r}_{\downarrow} = -\phi^{r}_{\uparrow}$: we find that the symmetric effect of news on monetary policy could not be rejected for the first subsample, even though only tightening monetary policy was significant. However, news on monetary policy became asymmetric in the second subsample, when easing monetary policy had a much bigger absolute impact on inflation expectations than tightening monetary policy.

Since the asymmetries became significant in the second subsample, we are interested to know how they evolved over time. We do this by conducting an expanding window estimation of z^{j} [equation (3)] starting in October 1983. Figure 2 shows the evolution of z^{j} for both inflation and monetary policy news, with the horizontal black line indicating significance at the 10% level. It is interesting that both news of increase and decrease on inflation (monetary policy) had similar absolute impacts on household inflation expectations for most of the 1980s. However, both news on inflation and monetary policy started to become increasingly asymmetric in the early 1990s, with the absolute impact of news on higher inflation becoming statistically greater than news on lower inflation after 1991 (top panel of Figure 2), and news on easing monetary policy having a greater impact than contractionary monetary policy after 1999 (bottom panel).

One explanation for this interesting evolution of asymmetric news may be rational inattention to information [Sims (2003)]. Since information is costly to process, households may only pay attention to news information that they regarded as relatively important. A general consensus developed in the 1980s and

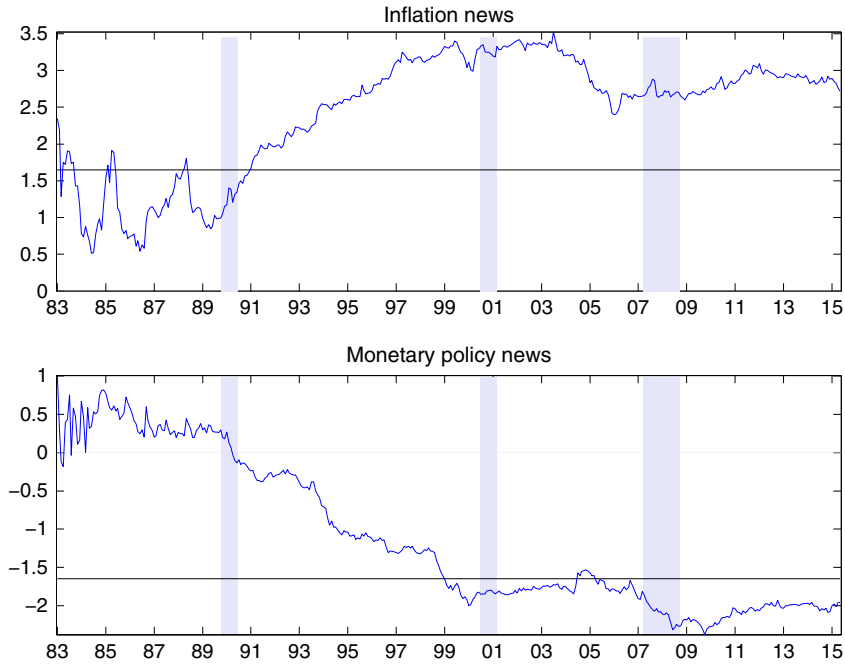


FIGURE 2. Time-varying asymmetries—z-score tests.

1990s was that high inflation was bad and needed to be avoided. Presumably then, high inflation news came to represent unfavorable information for households. As a consequence, low and stable inflation became a norm in the late 1980s and households inflation expectations became firmly anchored around 3%. Even though inflation became a lesser concern, household paid disproportionate attention to news on higher inflation that was regarded as unfavorable. Households may also consider that higher inflation (above the norm) tends to be more persistent compared to lower inflation, thus regarding higher inflation as unfavorable. After 2008, however, there may well have been a growing relative unease about the risks of deflation, but we see no evidence of that in Figure 2, since the z-score tests remained flat. With regards to monetary policy, we find that there is consistently significant evidence since 2007 of a greater impact of news on easing monetary policy in comparison to news on contractionary monetary policy. One interpretation of this result could be that news on upcoming cuts in the federal funds rate (that started to occur in late 2007) as well as on quantitative easing by the Federal Reserve (that started in November 2008) had a noticeably strong impact on inflation expectations.

In summary, we find evidence that rising inflation news and easing monetary policy impacts on household inflation expectations significantly more than does lower inflation and tightening monetary policy. This is true in particular for the relatively lower inflation period (Subsample 2: 1983:10–2016:02). Extending

window estimation shows that the impact of news on higher inflation (easing monetary policy) increasingly became bigger compared to lower inflation (contractionary monetary policy) during the 1990s. These asymmetries on both news persisted through the remaining sample.

5. THE IMPACT OF NEWS UNDER THE ZERO LOWER BOUND

Does the impact of inflation and monetary policy news change under the zero lower bound from 2008? Table 3 shows the estimates of Models 1–7 for the period between June 2008 and February 2016. Compared with the second subsample in Table 1, inflation news has a much bigger impact on household inflation expectations for all models. This may reflect the fact that households in this period realize that the FED had lost the effectiveness of its conventional instrument (the federal funds rate) in managing inflation (deflation). Therefore, in this period households may have reacted more sensitively to any news on inflation. Looking again at Figure 1, note that median inflation expectations were almost always greater than realized inflation from 2008.

During this period, the FED could not cut the *current* federal funds rate any further, though it was able to and did use extensively forward guidance, aiming to influence expectations of *future* interest rates and inflation to try to stimulate the economy. Forward guidance can be either Odyssean—when the Feb publicly commits monetary policy to a future action, or Delphic—when the policy states the likely future policy actions based on the policymaker's potential private information about macroeconomic fundamentals; see, for example, Campbell et al. (2012). In addition, the FED undertook three rounds of large-scale asset purchases from 2008 to 2014, otherwise known as "quantitative easing," leading to a significant expansion of its balance sheet with bank debt, treasury securities, and mortgage-backed securities.

Similar to the earlier results, the effectiveness of such unconventional policies relies crucially on how those economic agents respond on hearing the news of these policies. However, compared with the second subsample of Table 1, there are two noticeable differences. First, the impact of monetary policy news on expected inflation *appears* strengthened (Models 2–3), even when the households' demographic backgrounds are jointly considered, thus seeming to strengthen the signaling role of the FED's policy. Second, jointly considering consumer sentiment (Model 6) makes the impact of monetary policy small and statistically insignificant. The comparison of this to the results in Table 1 is indicative of the different implications of unconventional and conventional monetary policy news. The (unconventional) monetary policy news estimate of Model 2 in Table 3 may be seen as a proxy for consumer sentiment in relation to inflation expectations formation. Regressing consumer sentiment on monetary policy news yields a significant coefficient (at the 1% level) of −7.56. Therefore, hearing news on monetary policy contraction would be associated with a 7.56 reduction in consumer sentiment during the zero lower bound period. Consumer sentiment

TABLE 3. Regression results for inflation expectations

	Sample period: 2008:06–2016:02						
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Constant	4.04***	4.04***	10.30***	3.79***	3.80***	5.84***	10.53***
News: inflation(ϕ^{π})	1.22***	1.21***	1.08***	1.01***	0.95***	0.80***	0.67***
News: monetary policy(ϕ^{r})	–	0.52***	0.52***	0.35***	0.20*	0.14	0.07
Log income	–	–	−0.49***	–	–	–	−0.44***
Age	–	–	−0.00	–	–	–	−0.01***
Female	–	–	0.58***	–	–	–	0.48***
Education	–	–	−0.26***	–	–	–	−0.17***
Perception: government policy	–	–	–	−0.90***	–	–	−0.37***
Expectation: interest rate	–	–	–	–	0.35***	–	0.35***
Expectation: unemployment rate	–	–	–	–	1.14***	–	0.58***
Consumer sentiment	–	–	–	–	–	−0.02***	−0.01***
Adjusted-R^2	0.527	0.527	0.543	0.540	0.549	0.554	0.572

Notes: 1. The sample size is 38128 between June 2008 and February 2016.
2. *, **, and *** represent significance at 10%, 5%, and 1% levels of significance.

fell significantly in 2008–2009, but improved consistently thereafter. The significant negative estimate of consumer sentiment in Model 6 suggests that those households hearing news on monetary policy easing, and thus credit easing, recognized this as a signal to lower their inflation expectations and to expect easier conditions that improved consumer sentiment. Households not hearing this news had no signal and were responsible for maintaining inflation expectations above realized inflation.

In summary, these results suggest in a consistent way that monetary policy news provided a signal about future inflation. This signaling effect manifests through consumer confidence during the zero lower bound period. This implies that central banks should pay particular attention to the impact of their policy communications on consumer sentiment to maximize the impact of asset purchases and forward guidance on inflation expectations.[13]

6. CONCLUSIONS

We have examined the impact of news on household inflation expectations. Using monthly US consumer inflation expectations data between January 1978 and February 2016, we find that, in general, exposure to news on inflation and monetary policy significantly helps to explain household inflation expectations. This remains true even after controlling for households demographic characteristics, their perception of the effectiveness of government policies in managing business cycles, their expectations of future interest and unemployment rates, and their sentiment. This result tells us that the average effect of news is unaffected by the controls. To understand better other distributional aspects of the response, we would need to consider empirical nonlinearities which we leave for future research.

We find evidence that news on inflation and monetary policy had an asymmetric impact on household inflation expectations. In particular, households responded to news of higher inflation and easing monetary policy significantly more than news of lower inflation and tightening monetary policy. This was especially true in the relatively low inflation period after 1983 and probably was a result of the broad persuasion by public figures about the dangers of high inflation. The more unfavorable perception of risks of higher inflation remained valid also after 2008, even though, also, the impact of a deflation threat has likely increased since then.

From 2008, expected inflation became persistently higher than realized inflation. We find news of unconventional monetary policy acts as an imperfect signaling device for household inflation expectations, which may be seen as a proxy for consumer sentiment in relation to inflation expectation formation. Weak consumer sentiment through perceived credit market conditions may have played an important role in understanding the relatively high inflation expectations in this period.

NOTES

1. Each month, about 40% of the households are randomly chosen to be reinterviewed 6 months after their initial interview. This rotating panel feature is useful for analyzing how consumers update their inflation expectations.

2. Our paper is also related to a growing theoretical literature that shows monetary policy could have real effects even in the absence of nominal rigidities, if we are willing not to assume rational expectations. The transmission channels may arise from information rigidities [Woodford (2001)], rational inattention [Adam (2007)], and potential signaling effects [Melosi (2017)].

3. A value of 1 indicates grade 0–8 without high school diploma; 2 indicates grade 9–12 without high school diploma; 3 indicates grade 0–12 with high school diploma; 4 indicates grades 13–17 without a college degree; 5 indicates grade 13–16 with a degree; and 6 indicates grade 17 with a college degree.

4. The survey asks respondents to provide their opinion on the following question:

> As to the economic policy of the government—I mean steps taken to fight inflation or unemployment—would you say the government is doing a good job, only fair, or a poor job?

5. The survey asks respondents to provide their forecast of interest rate and unemployment:

> No one can say for sure, but what do you think will happen to interest rates for borrowing monetary during the next 12 months—will they go up, stay the same, or go down?
>
> How about people out of work during the coming 12 months—do you think that these will be more unemployment than now, about the same or less?

6. More details about the calculation can be found at https://data.sca.isr.umich.edu/fetchdoc.php?docid=24770

7. The median inflation expectation in the first subsample (1978:01–1983:09) was 6%, compared with 3% in the second subsample (1983:10–2016:02). This reduction in the median expectation was accompanied by a reduction in the heterogeneity of inflation expectations, with the variance of the cross-sectional distribution decreasing from 34.3 in the first subsample to 15.2 in the second subsample. This reduction in the heterogeneity of inflation expectations was likely due to the low and stable inflation rate in the second subsample, and a stronger emphasis placed by the FED on maintaining low and stable inflation.

8. The fraction of household that had heard inflation (monetary policy) news was 11.33(12.47)% on average for the first subsample, decreasing to 5.61(5.96)% in the second subsample.

9. The monthly dummies capture the effect of common factors on household inflation expectation. One of these factors may be the objective intensity of news reporting on inflation and monetary policy.

10. A potential endogeneity problem arises for the two types of news used in this paper. For example, in response to tighter monetary policy there may be a perception in this news that inflation prospects will be mitigated. To address this issue, we reran the estimation excluding those households that indicated they have heard news on lower inflation and tightening monetary policy and those households that indicated they have heard news on higher inflation and easing monetary policy. There are only marginal changes to the estimated coefficients and all results remain qualitatively the same. The results are available from the authors upon request.

11. We cannot rule out the possibility that households form higher inflation expectations when hearing of contractionary monetary policy because they may think the monetary policy is too accommodative–see, for example, Clarida et al. (2000) and Gertler et al. (1999). A consensus about accommodative monetary policy contributing to high inflation was achieved much later (in the 1980s), and in the first sample period, it was surely not well understood when households formed their inflation expectation. Therefore, we do not expect the *average* household would form inflation expectation in this sophisticated way.

12. Our credibility interpretation remains valid after distinguishing easing and tightening of monetary policy news in the second subperiod. Though the average household reduces inflation expectations when hearing news on easing monetary policy, those who perceive effective government policies understand the implication of monetary policy and forecast higher inflation. The detailed results on this are available on request.

13. We thank the referee for suggesting this.

REFERENCES

Adam, K. (2007) Optimal monetary policy with imperfect common knowledge. *Journal of monetary Economics* 54(2), 267–301.

Bai, J. and P. Perron (2003) Computation and analysis of multiple structural change models. *Journal of Applied Econometrics* 18(1), 1–22.

Blanchflower, D. and C. MacCoille (2009) The Formation of Inflation Expectations: An Empirical Analysis for the UK. Technical report, National Bureau of Economic Research.

Bryan, M. and G. Venkatu (2001a) The curiously different inflation perspectives of men and women. Federal Reserve Bank of Cleveland, *Economic Commentary*.

Bryan, M. and G. Venkatu (2001b) The demographics of inflation opinion surveys. Federal Reserve Bank of Cleveland, *Economic Commentary*.

Campbell, J., C. Evans, J. Fisher and A. Justiniano (2012) Macroeconomic effects of federal reserve forward guidance. *Brookings Papers on Economic Activity* 2012(1), 1–80.

Carroll, C. D. (2003) Macroeconomic expectations of households and professional forecasters. *The Quarterly Journal of Economics* 118(1), 269–298.

Carvalho, C. and F. Nechio (2014) Do people understand monetary policy? *Journal of Monetary Economics* 66, 108–123.

Clarida, R., J. Gali and M. Gertler (2000) Monetary policy rules and macroeconomic stability: Evidence and some theory. *The Quarterly Journal of Economics* 115(1), 147–180.

Curtin, R. (1996) *Procedure to Estimate Price Expectations*. University of Michigan Survey Research Center.

Dräger, L. (2015) Inflation perceptions and expectations in Sweden–Are media reports the missing link? *Oxford Bulletin of Economics and Statistics* 77(5), 681–700.

Dräger, L. and M. J. Lamla (2017) Imperfect information and consumer inflation expectations: Evidence from microdata. *Oxford Bulletin of Economics and Statistics* 79(6), 933–968.

Easaw, J., R. Golinelli and M. Malgarini (2013) What determines households inflation expectations? Theory and evidence from a household survey. *European Economic Review* 61, 1–13.

Ehrmann, M., D. Pfajfar and E. Santoro (2017) Consumer attitudes and their inflation expectations. *International Journal of Central Banking* 13(1), 225–259.

Gertler, M., J. Gali and R. Clarida (1999) The science of monetary policy: A new Keynesian perspective. *Journal of Economic Literature* 37(4), 1661–1707.

Jonung, L. (1981) Perceived and expected rates of inflation in Sweden. *The American Economic Review* 71(5), 961–968.

Lahiri, K. and Y. Zhao (2016) Determinants of consumer sentiment over business cycles: Evidence from the US surveys of consumers. *Journal of Business Cycle Research* 12(2), 187–215.

Lamla, M. and T. Maag (2012) The role of media for inflation forecast disagreement of households and professional forecasters. *Journal of Money, Credit and Banking* 44(7), 1325–1350.

Madeira, C. and B. Zafar (2015) Heterogeneous inflation expectations and learning. *Journal of Money, Credit and Banking* 47(5), 867–896.

Malmendier, U. and S. Nagel (2016) Learning from inflation experiences. *The Quarterly Journal of Economics* 131(1), 53–87.

Mankiw, N. G. and R. Reis (2002) Sticky information versus sticky prices: A proposal to replace the new Keynesian Phillips curve. *The Quarterly Journal of Economics* 117(4), 1295–1328.

Mankiw, N. G., R. Reis and J. Wolfers (2004) Disagreement about inflation expectations. In: M. Gertler and K. Rogoff (eds.), *NBER Macroeconomics Annual 2003*, Volume 18, NBER Chapters, pp. 209–270. Cambridge, MA: MIT.

Melosi, L. (2017) Signalling effects of monetary policy. *The Review of Economic Studies* 84(2), 853–884.
Pfajfar, D. and E. Santoro (2009) Asymmetries in Inflation Expectations across Sociodemographic Groups. Technical report, mimeo.
Pfajfar, D. and E. Santoro (2013) News on inflation and the epidemiology of inflation expectations. *Journal of Money, Credit and Banking* 45(6), 1045–1067.
Sims, C. (2003) Implications of rational inattention. *Journal of Monetary Economics* 50(3), 665–690.
Woodford, M. (2001) Imperfect Common Knowledge and the Effects of Monetary Policy. Technical report, National Bureau of Economic Research.
Zhang, J., W. K. Härdle, C. Y. Chen and E. Bommes (2016) Distillation of news flow into analysis of stock reactions. *Journal of Business & Economic Statistics* 34(4), 547–563.

APPENDIX: PAIRWISE CORRELATION

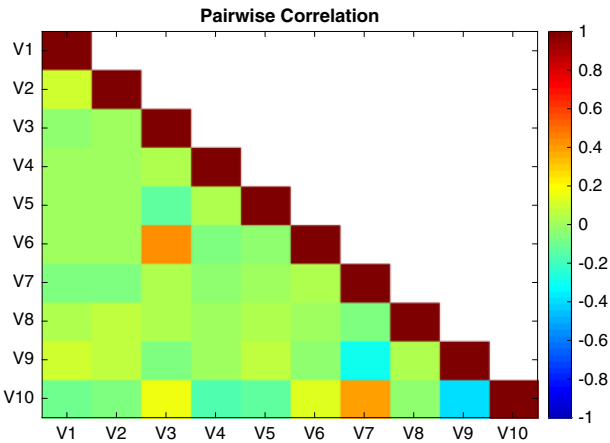


FIGURE A1. V1: News: inflation; V2: News: monetary policy; V3: Log income; V4: Age; V5: Female; V6: Education; V7: Perception: government policy; V8: Expectation: interest rate; V9: Expectation: unemployment rate; V10: Consumer sentiment.

Decision Support Systems 113 (2018) 73–85

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Improving crime count forecasts using Twitter and taxi data

Lara Vomfell[a,*], Wolfgang Karl Härdle[b,c], Stefan Lessmann[b]

[a] Warwick Business School, University of Warwick, Coventry CV4 7AL, UK
[b] Faculty of Business and Economics, Humboldt University of Berlin, Unter den Linden 6, Berlin 10099, Germany
[c] Singapore Management University, 50 Stamford Road, Singapore 178899, Singapore

ARTICLE INFO

Keywords:
Predictive policing
Crime forecasting
Social media data
Spatial econometrics

ABSTRACT

Crime prediction is crucial to criminal justice decision makers and efforts to prevent crime. The paper evaluates the explanatory and predictive value of human activity patterns derived from taxi trip, Twitter and Foursquare data. Analysis of a six-month period of crime data for New York City shows that these data sources improve predictive accuracy for property crime by 19% compared to using only demographic data. This effect is strongest when the novel features are used together, yielding new insights into crime prediction. Notably and in line with social disorganisation theory, the novel features cannot improve predictions for violent crimes.

1. Introduction

Every day, people leave their neighbourhood to commute to work, shop in malls or relax in museums and bars. Such travel creates a social flow of both crime targets and perpetrators that connect areas beyond spatial distance and facilitates criminal activity [33].

Exploitation of location-based data offers new perspectives on the mechanisms of crime emergence and helps predict the occurrence of crime. Government institutions and especially police depend on adaptive, short-term crime predictions to anticipate changes and breaks in crime patterns and allocate scarce resources efficiently [e.g., 35].

The objective of this paper is to establish the performance of data on human dynamics in predicting crime. In pursuing this goal, the paper proposes predictive models that extend conventional crime forecasts by incorporating three sources of data: public venues, social media activity and taxi flows. We suggest alternative ways to extract features from the data sources and examine how their interaction improves predictive performance. This provides concrete guidance to decision makers on how to leverage these new data sources for accurate crime forecasting.

Empirical results using crime data from New York City confirm the relevance of the proposed features. Using a rolling-window prediction approach, we demonstrate that including the novel features significantly improves crime predictions for some types of crime. The results reveal interaction effects: Features from different data sources work best when used in combination.

Our dual approach of prediction and explanatory analysis addresses policy makers' concerns about preventing crime in a predictive policing and a wider prevention context. In line with social disorganisation and opportunity theory, our results add to a better understanding of the link between crime opportunities and human dynamics and highlight new areas for policy design.

The paper is organised as follows: Section 2 discusses related work. Section 3 introduces spatial and non-spatial prediction models. Section 4 outlines the data sources and feature construction methods. Empirical results are presented in Section 5 and discussed in Section 6. Section 7 concludes the paper.

2. Related work

Our study uses online data together with spatial analysis to understand behavioural aspects of the emergence of crime and how this improves crime prediction. In this section, we briefly introduce seminal explanatory studies that use spatial analysis to provide empirical support for prominent crime theories. Then, we elaborate how online data sources have been used in explanatory contexts before presenting forecasting studies that use online data or spatial analysis to predict crime.

The main theories concerned with explaining the spatio-ecological dimension of crime are opportunity theory and social disorganisation theory. The former analyses crime events as opportunities created by the intersection of a suitable target, a motivated offender, and lack of supervision [11]. Social disorganisation theory considers neighbourhood characteristics that influence the likelihood of criminal activity among inhabitants. A lack of social control and social cohesion within a community combined with structural disadvantages gives rise to criminal behaviour [20].

* Corresponding author.
E-mail addresses: l.vomfell@warwick.ac.uk (L. Vomfell), haerdle@hu-berlin.de (W.K. Härdle), stefan.lessmann@hu-berlin.de (S. Lessmann).

https://doi.org/10.1016/j.dss.2018.07.003
Received 28 February 2018; Received in revised form 20 June 2018; Accepted 19 July 2018
Available online 02 August 2018

Classical crime modelling draws on these theories and uses regression analysis to identify socio-economic predictors of criminal behaviours on an aggregated level. The findings emphasise the relevance of demographic characteristics such as residential instability, ethnic heterogeneity and population density [e.g. 29]. These results have been supplemented with spatial analysis to evaluate the relevance of spatial dependence. Spatial proximity to violence has been shown to be more important than demographic data [19,24].

With the availability of online data, crime modelling has shifted to incorporating aggregated, anonymous human behavioural data. Geotagged Twitter data in particular has been used to understand how topics on social media relate to crime, for example through term-frequency analysis [34]. Dynamic data on human activity has also been used to model crime. Traunmueller et al. [30] examine correlations between people activity features, which they derive from mobile phone data, and monthly crime rates.

Paralleling the development in classical crime modelling, online data has been combined with spatial analysis to explore their relationship. Bendler et al. [5] include Twitter and local points of interest (POI) data in a geographically weighted regression to capture human activity and explore spatial dependence between crime locations. They show that only some crimes such as burglary are related to Twitter activity.

Wang et al. [32] also consider POI data, which they integrate with taxi flow data to model yearly crime rates in Chicago. They find a model using both types of information to outperform models using only POI or taxi data. Such synergy hints at an interdependence between the two sources, which has also been observed by Bendler et al. [5].

In contrast to explanatory studies, crime prediction has paid comparatively less attention to the intersection of space and human dynamics and usually uses online data as inputs for crime prediction. For example, Aghababaei & Makrehchi [1] employ temporal topic detection to identify Twitter topics predicting crime. Bogomolov et al. [7] train a Random Forest to predict high-crime areas using features related to visitors volumes based on telecommunication records. Gerber [15] finds that prediction models using Twitter topic modelling outperform Kernel density estimation-based models.

There are few predictive studies using spatial analysis. Most notably, the work by Rosser et al. [28] analyses criminal incidents on a street segment-level instead of a grid- or census unit-level. However, human dynamics are not explicitly taken into consideration since the predictions are not based on any analysis of traffic volume or pedestrian density on those streets.

Xue & Brown [35] model the coordinates of crime as a locally optimal site picked by the offender from a set of spatial alternatives to commit the crime. Similar to Rosser et al. [28], they do not take human dynamics into account.

An interesting approach to synthesising different data sources for crime prediction is proposed by Kang & Kang [16] who train a deep neural network (DNN) to integrate Google Streetview images and temporal features into a joint feature as input layers for DNN-based crime prediction.

Table 1 shows how explanatory studies frequently incorporate behavioural data and spatial dependence, whereas predictive studies focus on only one of the two aspects. Therefore, a contribution of this paper is the joint consideration of data on spatial structure and human dynamics. A second contribution stems from combining explanatory and predictive analysis as it is crucial to understand the underlying process of crime generation to not only successfully predict crime incidents but also prevent crime [9].

3. Methodology

Crime rates depend on the underlying population at risk, which need not correspond to the residential population in a geographic unit [e.g. 23]. Therefore, a common modelling approach, which we adopt in this study, is to use counts of crime incidents. Our data forms a panel of crime counts and covariates for 1974 census tracts for 26 weeks, indexed by i and t, respectively.

Our analysis approach is two-fold: our main focus is crime prediction, which we supplement with explanatory analysis. We use spatial econometric models and machine learning techniques to fit models and predict crime. In the following section, we first describe the econometric models in Subsections 3.1 and 3.2. Then, we describe the machine learning methods used in Subsection 3.3. We use a rolling window prediction approach which we explain in Subsection 3.4 where we also present the linear predictors.

Before detailing the models, we introduce some notation. Modelling crime counts in a city begins with a specific, bounded two-dimensional area $D \subset \mathbb{R}^2$, where D denotes the surface area of the city. D can be partitioned into a finite number N of well-defined, non-overlapping areal units, e.g. census tracts.

Crime events are modelled as realisations of a point process on D. The locations of k_t crime events at time t are denoted by $S_t = \{s_{1t}, ..., s_{k_t t}\}$. This allows modelling the number of realised events in an areal unit as a time-dependent count variable. Let this count variable be defined as $m(i, t) = \sum_{l=1}^{k_t} \mathbb{1}(s_{lt} \in i)$, $i = 1, ..., N$ such that $m(i,t)$ gives the number of crimes in unit i at time t. Let y denote the vector of NT count variables observed at the N areal units in T periods such that $m(i,t) \equiv y_{it}$.

Spatial dependence between areas can take the form of a Markov random field, which defines a neighbourhood for each element in y. An areal unit j is a neighbour of areal unit i if the conditional distribution of y_i depends on y_j [12]. Let $A_i = \{j : j \text{ is a neighbour of } i\}$ be the neighbourhood of unit i. Note that A_i excludes unit i.

3.1. Linear models

Consider the simple pooled linear panel regression model:

$$y = X\beta + e, \quad e \sim N(0, \sigma^2 I_{NT}), \tag{1}$$

where X is a $NT \times K$ matrix of K regressors. In the presence of spatial dependence, the error terms in Eq. (1) are no longer uncorrelated. Approaches to account for such error correlation include the simultaneous autoregressive (SAR) and the conditional autoregressive (CAR) model.

The SAR model introduces spatial structure through a spatial lag [12, p. 406]:

$$y = (I_T \otimes \rho W)y + X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_{NT}), \tag{2}$$

where $\otimes$ denotes the Kronecker product, I_T denotes the identity matrix of order T, and W is a $N \times N$ binary matrix specifying which areas are spatially adjacent with $w_{ii} = 0\ \forall i$. ρ is the parameter that specifies the magnitude of spatial dependence.

The inclusion of a spatial lag of the dependent variable accounts for spatial spillovers and a mismatch of the spatial scale with the spatial event. Both effects occur in crime modelling since the contagion effect of crimes leads to a diffusion through space. In addition, economic and criminal features do not match perfectly with the spatial units. A spatial lag SAR model is a convenient choice to account for these characteristics [3].

The CAR model introduces a spatial dependence parameter in the error term which accounts for small-scale spatial variation [12, p. 407]. This yields the following model:

$$\begin{aligned} y &= X\beta + \varepsilon, \\ \varepsilon &\sim N(0, \sigma^2\{I_T \otimes (I_N - \delta W)^{-1}\}), \end{aligned} \tag{3}$$

where W is again a $N \times N$ spatial adjacency matrix and δ denotes the magnitude of spatial dependence between neighbouring regions.

The CAR model introduces spatial structure as a Markov random field, such that the conditional distribution of each area depends on the neighbourhood. The distribution of y_{it} conditional on all y_{jt} can be

Table 1
Literature overview.

Study	Explanatory/ predictive	Spatial	Human dynamics	Machine learning	Crime type	City	Time frame
Wang et al. [32]	E	✓	✓		All crime	Chicago	Yearly
Bendler et al. [5]	E	✓	✓		Assault, burglary, homicide, theft, …	San Francisco	Hourly
Traunmueller et al. [30]	E		✓		Street vs. indoor	London	Monthly
Williams et al. [34]	E		✓		Burglary, theft, drugs, violent crime, …	London	Monthly
Gerber [15]	P		✓		Theft, battery, drugs, burglary, …	Chicago	Daily
Xue & Brown [35]	P	✓		✓	Burglary	Richmond, VA	Monthly
Rosser et al. [28]	P	✓		✓	Residential burglary	Anonymous UK city	Daily
Bogomolov et al. [7]	P		✓	✓	(Hotspot classification)	London	Monthly
Kang & Kang [16]	P	✓		✓	All crime	Chicago	Daily
Aghababaei & Makrehchi [1]	P		✓	✓	Theft, drugs, burglary, …	Chicago	Daily
This study	E and P	✓	✓	✓	violent and property crime	New York	Weekly

shown to be

$$y_{it}|y_{jt} \sim N\left(X_{it}^{\top}\beta + \sum_{j} \delta W_{ij}(y_{jt} - X_{jt}^{\top}\beta),\ \sigma_i^2\right), \tag{4}$$

for $i \neq j$, where σ_i^2 denotes the conditional variance [12, p. 407]. This conditional dependence structure is different from the structure modelled in a SAR model. There, the inclusion of the spatial lag means that values in unit i do not only depend on values in the direct neighbourhood A_i but also on higher-order neighbours, i.e. neighbours of neighbours. Therefore, the SAR model implies a global dependence structure compared to the CAR model [3].

3.2. Count models

Linear models offer a broad framework to include spatial structure but fail to accommodate the integer-valued and non-negative nature of crime counts. Small counts are better modelled by a Poisson Generalised Linear Model. In the case of crime counts, the Poisson parameter λ represents the expected incident count:

$$\lambda = E(y \mid X) = e^{X^{\top}\beta}. \tag{5}$$

Similar to the linear model, the errors of the Poisson model in Eq. (5) are no longer uncorrelated under spatial dependence. Poisson Generalised Linear Mixed Models (GLMMs) account for this dependence by incorporating a random effect in the GLM predictor. GLMMs model $E(y \mid X)$ as a linear combination of fixed effects X and random effects Z with a logarithmic link function [2]:

$$\log \lambda_{it} = X_{it}^{\top}\beta + Z_{it}\eta_i. \tag{6}$$

Here, $Z\eta$ are location-specific random effects. At each cross-section t, Z is a $N \times N$ indicator matrix of the spatial units, which means that the random effect is simply a random intercept added to the conditional mean. The distribution of the random vector η is assumed to be multivariate normal:

$$\eta \sim N(0, D), \quad D = \sigma^2 Q^{-1}. \tag{7}$$

Q is a symmetric spatial dependency matrix different from the adjacency matrix W used before. Its entries are as follows:

$$Q_{ij} = \begin{cases} |A_i| & \text{if } i = j, \\ -1 & \text{if } j \in A_i \text{ and } i \neq j, \\ 0 & \text{if } j \notin A_i \text{ and } i \neq j, \end{cases} \tag{8}$$

where the $|A_i|$ entries on the diagonal denote the size of the neighbour set and neighbours are indicated by -1 [22, p. 186]. In the non-spatial Poisson GLM in Eq. (6), the variance is equal to the expectation [2]. In the model in Eq. (6), this is not the case. Here, σ accounts for both the variance and spatial dependence. The parameters in Eqs. (6) and (7) are estimated using restricted maximum likelihood (REML) and Fisher Scoring [18].

3.3. Machine learning models

Previous models make assumptions about the data-generating process and consider a linear additive relationship between crime counts and covariates. Machine learning techniques are more flexible and account for non-linearity in a data-driven manner [21]. We concentrate on random forest (RF), gradient boosting machines (GBMs), and feed-forward artificial neural networks (ANNs), all of which have shown promising results in previous studies [e.g. 6,13].

RF develops an ensemble of size k through drawing k bootstrap samples from the training data. The base models in RF consist of individual decision trees, which are grown from the bootstrap samples. To increase randomness among the base models, RF determines the best split during tree growing among a randomly sampled subset of covariates [8]. The model prediction consists of the simple average calculated across the k base models.

GBMs embody the idea of additive modelling. The algorithm incrementally develops an ensemble through adding base models. In our paper, we use regression trees as base models. These are fitted to the residuals via the negative gradient of the loss function of the current ensemble. GBM predictions are obtained by calculating a weighted average over base model forecasts, whereby the weights are determined during gradient descent [14].

An ANN model consists of interconnected layers of processing units (neurons) with connection weights representing the model parameters. Estimating an ANN model involves minimising loss functions with respect to connection weights using gradient-based methods. ANNs calculate the output of a neuron as a non-linear transformation of the weighted sum over its input neurons. The transformations are called activation functions and allow an ANN to capture non-linear patterns in data [17]. We use a Rectified Linear Unit (ReLU) activation function.

3.4. Rolling window prediction

We use a rolling window prediction approach where we use all $y_{1:t} = (y_{1,1:t},\ldots,y_{N,1:t})^{\top}$ to estimate our models and produce forecasts $\hat{y}_{t+1} = (\hat{y}_{1,t+1},\ldots,\hat{y}_{N,t+1})^{\top}$ for the next week. We compute the prediction errors $e_{t+1} = y_{t+1} - \hat{y}_{t+1}$. We repeat this step for $t = h,\ldots,T-1$ where h is the smallest number of observations used for estimating the model. We set $h = T/2$. We then calculate the total mean squared error based on the obtained errors $MSE = \frac{1}{Nh}\sum_{i=1}^{N}\sum_{t=h+1}^{T} e_{it}^2$.

For the linear model, the predictions for weekly crime counts are

Table 2
Predictors for the spatial linear regression models considered in the study.

Model	Predictor
LR	$\hat{y}_{t+1} = X_{t+1}\hat{\beta}$
SAR	$\hat{y}_{t+1} = (I_N - \rho W)^{-1} X_{t+1}\hat{\beta} + (I_N - \rho W)^{-1}\hat{\varepsilon}$
CAR	$\hat{y}_{i,t+1} = X_{i,t+1}^{\top}\hat{\beta} + \sum_j \delta w_{ij}((1/t)\sum_{k=1}^{t}(y_{jk} - X_{jk}^{\top}\hat{\beta}))$
GLM	$\hat{y}_{t+1} = \exp(X_{t+1}\hat{\beta})$
GLMM	$\hat{y}_{t+1} = \exp(X_{t+1}\hat{\beta} + Z_{t+1}\hat{\eta})$

obtained by using the best linear unbiased predictor or its panel equivalent [4]. Table 2 gives the predictors for the time period $t + 1$ for the regression models. The SAR predictor is obtained by spatially lagging the linear predictor and adding the spatially lagged error vector of the model. The CAR predictor is obtained by taking a time-averaged conditional expectation.

Machine learning models require auxiliary data for hyperparameter tuning to enable adaption of a learning algorithm to a given task [e.g., 10]. For such models, we use the first 1,…,$t - 2$ weeks in the window of length t as training set and the last two weeks as validation set for parameter tuning. This way, we still produce an out-of-sample one-step ahead forecast for $t + 1$. We report the models with the lowest prediction errors on the test set. We tune the hyperparameters using grid search (see Appendix A for details) at each window. Since we include lagged crime counts as predictors, the rolling window approach corresponds to cross-validation for time-dependent data.

4. Data integration and feature construction

Since the data sources we use (census, POI, Twitter and taxi flow data) have different time coverages, we use the most recent complete overlap from June 1, 2015 to November 29, 2015. We aggregate the temporal data to weekly intervals which begin uniformly on Monday. The final data set covers 26 weeks. As discussed in Section 3.4, we set $h = T/2 = 13$ weeks. This choice results in 13 windows of length 1 : t,$t = 13$,…,25, on which we train our models. We then produce 13 separate one-step ahead forecasts for week $t + 1$. We use the human dynamics features at time t to predict crime counts at time $t + 1$.

The short time frame of the data makes explicit modelling of temporal effects infeasible since 26 weeks are not sufficient to reliably estimate weekly or monthly seasonality. We also do not include a dummy for the week of the first of a month to account for a potential "pay day effect": While one might expect that criminal behaviour associated with drinking increases after receiving the monthly salary, this implied human activity is already captured by our novel data sources.

The following subsections introduce the data sources. For each source, we elaborate on alternative options for feature engineering since different formulations may differ in their predictive power. The definitions always include a general definition using raw counts and additional versions similar to data standardisation or variance reduction such a log-transformed features. We do not consider further feature transformations such as Principal Component Analysis due to their non-interpretability. Section 4.6 details how the final set of features has been selected.

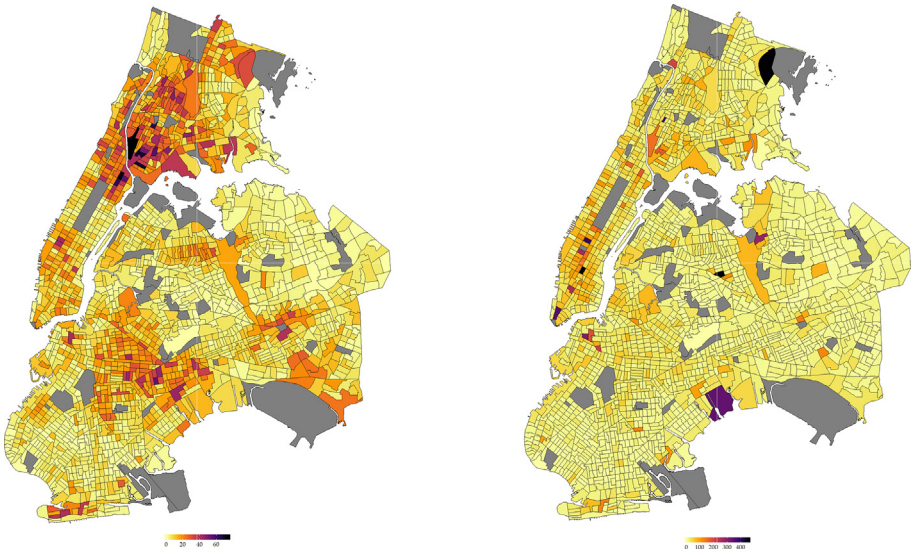
4.1. Census

The spatial units of analysis are census tracts as defined by the US Census Bureau. We use the coordinates of point-referenced data to match them to the corresponding census tract. We select the following eight demographic variable from Summary File 1 of the 2010 census data [31] based on previous studies [e.g., 32]: the total population in the census tract, the median age of the population, the share of males, the share of the Black, Asian, and Hispanic population, respectively, the rate of female-headed family households, and the rate of vacant accommodation.

4.2. New York City crime data

Data on criminal incidents is provided by the New York City Police Department [25]. We focus on violent and property crime because their spatial distribution differs, which facilitates examining the proposed features in a context of varying spatial dependence. Violent crime encompasses murder and non-negligent manslaughter, robbery, and aggravated assault. Since rape incidences are not geo-located in the NYPD dataset, we exclude them from the analysis. Property crime comprises burglary, larceny-theft, motor vehicle theft, and arson.

Fig. 1a and b shows the spatial distribution of crime for the analysis period of June to November 2015. Property crime exhibits a more even distribution than violent crime. The strength of spatial correlation between areas is tested using Moran's I [3]. For both crime types and



(a) Violent Crime (b) Property Crime

Fig. 1. Number of crime incidents between June and November 2015. In the property crime map, the area around Penn Station (largest outlier with 2002 incidents) is excluded for more consistent colour scaling. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

every time period, the null hypothesis of no spatial dependence is rejected with $p < 0.000$.

4.3. Foursquare

We gather POI data from Foursquare, a mobile recommendation app. We consider POI data a characterisation of the census tract since POI categories attract specific groups of people. For example, one can expect that more nightlife venues attract drunken behaviour. Prior work has evidenced a connection between criminal activity and local points of interest in a geographic area [5]. Foursquare categorises all venues along nine main dimensions: nightlife, food, arts & entertainment, residence, shops, travel, outdoors & recreation, college & education, and professional. In total, we obtain 47,113 POI in the geographic area of interest.

Two different ways of constructing the feature from POI data are considered:

1. the total counts of venues per category,
2. the share of categories on the total number of venues in the census tract.

4.4. Taxi

The NYC Taxi & Limousine Commission [26] provides taxi flow data. We argue that taxi flows illustrate connections between different neighbourhoods beyond what is already covered through spatial proximity. Around 25% of all taxi trips end in a census tract that is not a neighbour of the tract they started in, suggesting that the taxi feature captures connections between census tracts that go beyond spatial proximity. Fig. 2 supports this view and, in agreement with Wang et al. [32], confirms taxi data as a valuable source for crime modelling.

We consider all trips within New York City in the analysis time frame but exclude trips that start or end outside the analysis area. This gives 70,288,218 trips in the 26 weeks. We aggregate individual trips to a weekly connection flow matrix F, with rows (columns) of F referring to the census tract where the trip started (ended). Hence, f_{ij} denotes the number of trips made from tract i to j for each time interval. Note that $f_{ii} = 0\ \forall i$ as otherwise, crime rate of census tract i would be used as its own predictor.

The taxi flow feature is then constructed as $c_t = F_t y_{t-1}$ such that neighbouring crime rates are weighted by the magnitude of flow F. It is crucial to note that the crime vector y is lagged by a week to prevent unintended implicit simultaneity of the response y_t and its predictors. The week index t is dropped for ease of notation.

We propose three different ways to construct c and demonstrate the calculation for c_1, the feature of example tract 1:

1. Raw multiplication: One can define t as the simple matrix multiplication of the flow matrix F and the crime count vector y:

$$c_1 = f_{12}y_2 + \ldots + f_{1N}y_N.$$

2. Normalised by source: The taxi flow arriving in each census tract is normalised by the total number of flows leaving the source census tract. For example, the flow leaving the second census tract towards the first tract is normalised by all flows leaving from the second tract:

$$c_1 = \frac{f_{21}}{f_{21} + f_{23} + \ldots + f_{2N}}y_2 + \ldots + \frac{f_{N1}}{\sum_{i=1}^{N} f_{Ni}}y_N.$$

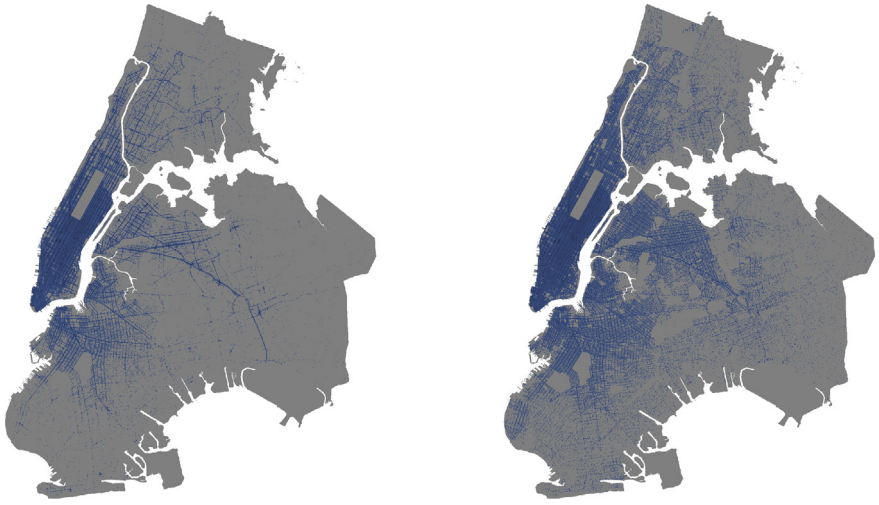
3. Normalised by destination: The taxi flow arriving in each census tract is normalised by the total number of flows arriving in the destination census tract:

$$c_1 = \frac{f_{21}}{f_{21} + f_{31} + \ldots + f_{N1}}y_2 + \ldots + \frac{f_{N1}}{\sum_{i=1}^{N} f_{i1}}y_N.$$

4.5. Twitter

We use Twitter data as a proxy for day-to-day population density through tourists or visitors. Accordingly, we focus on the number of Tweets in an area but do not attempt to extract their topical content. While Foursquare data covers venues as potential destinations of human activity and taxi flow data records where people move to, some of the overall activity is not captured. For example, we observe high numbers of tweets in the census tract containing the 9/11 Memorial site, unaccounted for by any other feature, whether novel or demographic.

We source Twitter data from Pfeffer & Morstatter [27] who provide IDs to tweets published in the United States between June 1, 2015 and



(a) Pickups (b) Dropoffs

Fig. 2. Coordinates of complete taxi trips in New York City in week 46 in 2015.

November 30, 2015. We aggregate the number of tweets per week and census tract, and implement four versions of the Twitter feature:

1. Using the full activity,
2. counting night-time tweets only,
3. using log-transformed full activity,
4. using log-transformed night-time activity.

Any tweet sent out between 22 pm and 6 am contributed to the night-time feature. Taking the logarithm of the number of tweets serves to reduce variation between census tracts.

4.6. Evaluation and feature selection

We proposed multiple variable definitions for each novel data source. Since we are interested in interactions, we select the best combination of all feature types using a variable selection procedure where we estimate CAR models for all possible combinations of the definitions. We then produce one-step ahead forecasts for 13 weeks in total using the procedure described in Section 3.4 and pick the combination with the overall lowest MSE. In comparison with in-sample goodness-of-fit statistics such as R^2, the MSE-based selection strategy emphasises the predictive value of a feature on out-of-sample data. We suggest that a prediction-centric feature selection strategy is better aligned with the goal of forecasting crime accurately.

While some machine learning techniques such as Random Forests entail variable importance rankings that can guide variable selection, they may pick up non-linear relationships that linear models cannot accommodate. This would give machine learning models an advantage in subsequent comparisons. To counterbalance this, we select feature definitions through optimising predictions of a linear model. Out of the linear models, we choose the CAR model because it models the outcome variable as a linear combination directly (rather than on the log scale) and because the implied spatial dependence structure is local rather than global. Therefore, the selected feature definition combination is expected to suit the wide range of spatial and non-spatial models we consider.

Table 3a and b shows MSE values for property and violent crime. We present results for alternative definitions of the Twitter and taxi features. Total counts of venues for Foursquare category produces uniformly better results than the venue share. Overall, we observe the best results with the non-normalised POI feature, log-transformed nightly tweet activity, and taxi data normalised by destination. For the POI feature, however, using total counts outperforms normalisation. The counts preserve differences in the POI distribution across New York City, which results in better predictions than the shares of categories.

We provide a short data overview in Table 4. We find that the new features have low correlations with the demographic variables (all Pearson's $r < 0.35$) but higher correlations with crime of up to 0.63. This makes them valuable predictors in addition to the demographic variables which capture characteristics of the residential population only.

Table 3
MSE values for crime predictions from CAR models including POI data in the form of the total counts of venues per Foursquare category together with alternative definitions of the Twitter and taxi features.

Twitter	Taxi		
	Raw	Destination	Source
(a) Property crime			
All	4.5415	4.5221	4.8355
Night	4.5301	4.5272	4.8033
log All	4.5259	4.5329	4.8744
log Night	4.5193	**4.5051**	4.8626
(b) Violent crime			
All	0.5402	0.5396	0.5400
Night	0.5411	0.5404	0.5400
log All	0.5418	0.5405	0.5414
log Night	0.5417	**0.5392**	0.5398

The lowest MSE obtained is printed in bold.

Table 4
Summary statistics for the data set.

Variable	Mean	Std. deviation	Median	Min	Max
Property crime	1.45	2.34	1.00	0.00	56
Violent crime	0.37	0.74	0.00	0.00	11
Population	3829.61	2118.97	3431.50	56	26,588
Median age	35.92	6.01	35.40	13.40	80.90
Male	0.48	0.03	0.48	0.32	0.94
Black	0.28	0.31	0.12	0.00	0.96
Asian	0.13	0.16	0.06	0.00	0.88
Hispanic	0.27	0.23	0.18	0.00	0.91
Vacancy rate	0.08	0.06	0.07	0.00	0.65
Female-headed HH	0.20	0.12	0.17	0.00	0.58
log night tweets	1.22	1.42	0.69	0.00	7.87
Entertainment POI	2.90	3.39	2.00	0	64
Uni POI	2.51	3.43	2.00	0	61
Food POI	3.01	3.01	2.00	0	28
Professional POI	2.61	2.47	2.00	0	20
Nightlife POI	2.82	2.81	2.00	0	27
Outdoors POI	2.29	2.33	2.00	0	19
Shops POI	2.75	2.73	2.00	0	26
Travel POI	2.52	2.64	2.00	0	26
Residential POI	2.76	2.51	2.00	0	22
Taxi (property)	1.45	3.21	0.28	0.00	58.26
Taxi (violent)	0.37	0.77	0.08	0.00	22.92

$N = 1974$ census units observed over $T = 26$ weeks: 51,324 observations

5. Results

We consider eight different combinations of the features to investigate interactions. The census data serves as baseline and is included in all settings. The other groups are added in all possible combinations which we number from 1 to 8 (Table 5).

We begin with examining the explanatory power of the individual features and their interactions. In view of the large number of fitted models (2 types of crime × 5 model specifications × 8 settings over 13 windows), we do not reproduce all results. Instead, Tables 6 and 7 show the regression coefficients only for the largest possible window of 25 weeks and for setting 8, which includes all feature groups. As detailed in Appendix B, the coefficients are stable over different fitting windows.

Since the significance levels vary across models, we do not discuss each model individually. Instead, we focus on effects identified as significant by all models and refer to the average effect over models in the text. As the coefficients for GLM and GLMM are on the log-scale, we present the effects for linear and exponential models separately.

For property crime, the largest effect size across all non-exponential models is observed for the vacancy rate, which is significantly positively associated with property crime counts. The new features are significantly associated with property crime. In particular, a 1 unit

Table 5
Definition of experimental settings in terms of different groups of crime predictors.

Features	Settings							
	1	2	3	4	5	6	7	8
Census	✓	✓	✓	✓	✓	✓	✓	✓
POI		✓	✓	✓				✓
Taxi			✓		✓	✓		✓
Twitter				✓	✓		✓	✓

Table 6
Estimates and standard errors for property crime in the full setting (setting 8).

Variable	CAR	SAR	LR	GLM[a]	GLMM[a]
Intercept	−0.4255	−0.9090***	−0.9500	−0.2687***	−1.5419***
	(0.2448)	(0.2220)	(0.2246)	(0.0794)	(0.0001)
Population	0.0001*	0.0001***	0.0001	0.0001***	0.0001***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Median age	0.0124***	0.0011	−0.0003	−0.0076***	−0.0012***
	(0.0024)	(0.0018)	(0.0018)	(0.0008)	(0.0003)
Male	−1.3028***	0.4165	0.8414	−0.5642***	0.2675***
	(0.3929)	(0.3708)	(0.3752)	(0.1331)	(0.0000)
Black	0.7076***	0.1839**	0.2690	0.3999***	0.7297***
	(0.0920)	(0.0644)	(0.0652)	(0.0299)	(0.0001)
Asian	0.5009***	0.1802**	0.2511	0.1389***	0.2387***
	(0.1066)	(0.0690)	(0.0698)	(0.0322)	(0.0000)
Hispanic	0.9776***	0.2430**	0.3017	0.4787***	0.6541***
	(0.1029)	(0.0749)	(0.0758)	(0.0339)	(0.0001)
Vacancy rate	2.3155***	2.0637***	2.3373	0.6015***	0.8929***
	(0.2095)	(0.1777)	(0.1799)	(0.0519)	(0.0000)
Female-headed HH	−0.1474	1.3941***	1.5020	−0.0366	−0.6935***
	(0.2418)	(0.2054)	(0.2078)	(0.0887)	(0.0001)
log night tweets	0.0987***	0.1221***	0.2034	0.2344***	0.0682***
	(0.0099)	(0.0087)	(0.0089)	(0.0031)	(0.0032)
Entertainment POI	0.0151***	0.0157***	0.0171	−0.0055***	−0.0013
	(0.0036)	(0.0036)	(0.0036)	(0.0013)	(0.0013)
Uni POI	−0.0025	0.0012	0.0023	0.0000	0.0015
	(0.0032)	(0.0032)	(0.0032)	(0.0013)	(0.0016)
Food POI	0.0543***	0.0512***	0.0441	0.0218***	0.0293***
	(0.0045)	(0.0045)	(0.0046)	(0.0017)	(0.0019)
Professional POI	0.0185***	0.0162**	0.0221	0.0241***	0.0240***
	(0.0055)	(0.0055)	(0.0056)	(0.0020)	(0.0023)
Nightlife POI	−0.0599***	−0.0704***	−0.0761	−0.0334***	−0.0141***
	(0.0049)	(0.0048)	(0.0049)	(0.0017)	(0.0019)
Outdoors POI	0.0222***	0.0114*	0.0157	0.0138***	0.0127***
	(0.0058)	(0.0058)	(0.0058)	(0.0021)	(0.0023)
Shops POI	0.1433***	0.1236***	0.1209	0.0581***	0.0552***
	(0.0049)	(0.0049)	(0.0049)	(0.0017)	(0.0012)
Travel POI	0.0335***	0.0349***	0.0316	−0.0021	0.0124***
	(0.0051)	(0.0048)	(0.0049)	(0.0017)	(0.0019)
Residential POI	−0.0639***	−0.0474***	−0.0468	−0.0323***	−0.0212***
	(0.0052)	(0.0050)	(0.0050)	(0.0020)	(0.0021)
Taxi	0.1757***	0.2060***	0.2549	0.0480***	0.0250***
	(0.0043)	(0.0037)	(0.0037)	(0.0007)	(0.0011)

Standard errors in parentheses.
[a] Coefficients are on the log scale.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

increase in the weekly taxi flow is associated with an increase of 0.21 property crime counts. Similarly, an increment of one venue in the shops category results in a 0.13 increase of crime counts. Interestingly, a single additional residential venue, often elderly homes, is associated with a 0.05 decrease of property crime. This is an intuitive result when considering the higher presence of watchful neighbours. A similar result is observed for nightlife venues, which are associated with a 0.07 decrease. While the Twitter feature is significant, its effect of property crime is comparatively small as a 1 percent increase in night tweets yields a 0.14/100 = 0.0014 increase in crime counts. For the exponential models, we observe very similar results. The largest effect is, again, observed for the vacancy rate, followed by the taxi feature. The same POI venues are identified as influencing property crime counts.

For violent crime, the effect of social cohesion is pronounced. A 10% increase in the male share predicts a 0.1 increase of violent crime counts. Similarly, 10% increases in the rates of female-headed households and vacant homes are associated with increases of 0.1 and 0.04 in counts. The relevance of ethnic heterogeneity is less pronounced compared to property crime.

Regarding the new features, the effect of the Twitter feature is even smaller than for property crime. This is contrasted with the taxi feature where a 1 unit increase yields a 0.08 increase in violent crime. As with property crime, the food category has the largest effect size. Even then, an additional food venue is associated with a relatively small increase of 0.007 in violent crime. Again, the results for the exponential models are similar. The largest effects over both models are observed for demographic variables such as the male share of female-headed households.

With respect to spatial dependence, we find that estimates of the corresponding parameter in the CAR model are considerably larger than in the SAR model. For the CAR model, the average estimate for δ is 0.1357. For the SAR model, we obtain an average ρ estimate of 0.0629. Since the CAR model implies stronger local autocorrelation we find evidence for substantial dependence on direct neighbours.

We complete the explanatory analysis by inspecting the variable importance for the machine learning models. Since we estimate models over 13 windows, we average the importance rank for each variable over 13 windows. We present the five variables with the overall highest mean ranks in Tables 8 and 9. If the mean rank equals the importance rank, the variable has that rank across all 13 windows. We find that for both crime types, the taxi feature is highly important. Furthermore, the Twitter feature, which does not have a large effect size in the econometric models, is highly ranked for both crime types and machine learning models. Overall, we find that the regression results and the

Table 7
Estimates and standard errors for violent crime in the full setting (setting 8).

Variable	CAR	SAR	LR	GLM[a]	GLMM[a]
Intercept	−0.6394***	−0.8085***	−0.8674***	−3.2569***	−3.4166***
	(0.0863)	(0.0780)	(0.0785)	(0.1781)	(0.0001)
Population	0.0000**	0.0000***	0.0001***	0.0001***	0.0001***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Median age	−0.0009	−0.0013*	−0.0019**	−0.0258***	−0.0063***
	(0.0008)	(0.0006)	(0.0006)	(0.0020)	(0.0004)
Male	0.7738***	1.0597***	1.1766***	2.0673***	1.2500***
	(0.1386)	(0.1302)	(0.1311)	(0.2818)	(0.0001)
Black	0.2058***	0.0412	0.0991***	1.3403***	1.5275***
	(0.0325)	(0.0227)	(0.0228)	(0.0594)	(0.0003)
Asian	0.0773*	−0.0142	−0.0094	0.6724***	1.1663***
	(0.0376)	(0.0242)	(0.0244)	(0.0760)	(0.0000)
Hispanic	0.3434***	0.1154***	0.1931***	1.2196***	1.7608***
	(0.0363)	(0.0264)	(0.0266)	(0.0646)	(0.0003)
Vacancy rate	0.3272***	0.4673***	0.5500***	1.3155***	0.2863***
	(0.0735)	(0.0617)	(0.0621)	(0.1385)	(0.0001)
Female-headed HH	0.8420***	1.4462***	1.6264***	1.8172***	0.3705***
	(0.0853)	(0.0721)	(0.0726)	(0.1605)	(0.0002)
log night tweets	0.0107**	0.0104***	0.0158***	0.1099***	0.0207***
	(0.0035)	(0.0029)	(0.0030)	(0.0065)	(0.0053)
Entertainment POI	−0.0005	0.0008	0.0027*	−0.0056	−0.0072**
	(0.0013)	(0.0013)	(0.0013)	(0.0031)	(0.0024)
Uni POI	0.0007	0.0021	0.0027*	0.0086**	0.0055*
	(0.0011)	(0.0011)	(0.0011)	(0.0028)	(0.0026)
Food POI	0.0059***	0.0073***	0.0070***	0.0154***	0.0245***
	(0.0016)	(0.0016)	(0.0016)	(0.0036)	(0.0029)
Professional POI	0.0068***	0.0046*	0.0045*	0.0189***	0.0181***
	(0.0019)	(0.0019)	(0.0019)	(0.0045)	(0.0035)
Nightlife POI	0.0035*	0.0026	0.0030	0.0149***	0.0081**
	(0.0017)	(0.0017)	(0.0017)	(0.0037)	(0.0030)
Outdoors POI	0.0016	−0.0028	−0.0028	−0.0053	0.0001
	(0.0021)	(0.0020)	(0.0020)	(0.0047)	(0.0040)
Shops POI	0.0041*	0.0006	0.0001	−0.0056	0.0129***
	(0.0017)	(0.0017)	(0.0017)	(0.0039)	(0.0033)
Travel POI	0.0034	0.0014	−0.0004	−0.0118**	0.0280***
	(0.0018)	(0.0017)	(0.0017)	(0.0039)	(0.0031)
Residential POI	−0.0058**	−0.0029	−0.0027	−0.0026	−0.0091**
	(0.0018)	(0.0017)	(0.0018)	(0.0042)	(0.0033)
Taxi	0.0530***	0.0877***	0.1094***	0.1205***	0.0754***
	(0.0054)	(0.0049)	(0.0049)	(0.0060)	(0.0073)

Standard errors in parentheses.
[a] Coefficients are on the log scale.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Table 8
Variable importance for property crime in Setting 8 over 13 windows.

Rank	RF	Mean rank	GBM	Mean rank
1.	Taxi	1.00	Taxi	1.00
2.	log night tweets	2.00	Hispanic	2.00
3.	Hispanic	3.00	Entertainment POI	3.00
4.	Population	4.25	Median age	3.50
5.	Shops POI	4.50	log night tweets	4.08

Table 9
Variable importance for violent crime in Setting 8 over 13 windows.

Rank	RF	Mean rank	GBM	Mean rank
1.	Taxi	1.00	Taxi	1.00
2.	log night tweets	2.00	Female-headed HH	2.00
3.	Female-headed HH	3.08	Population	3.08
4.	Population	4.25	log night tweets	3.92
5.	Black	4.64	Median age	5.00

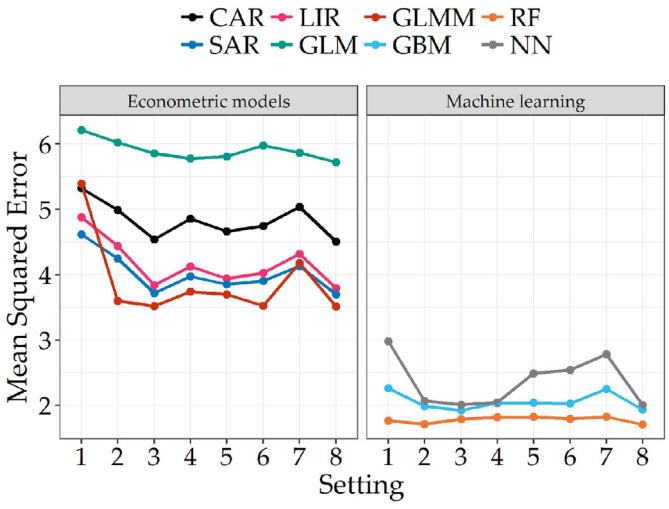


Fig. 3. MSE values of different models for property crime predictions.

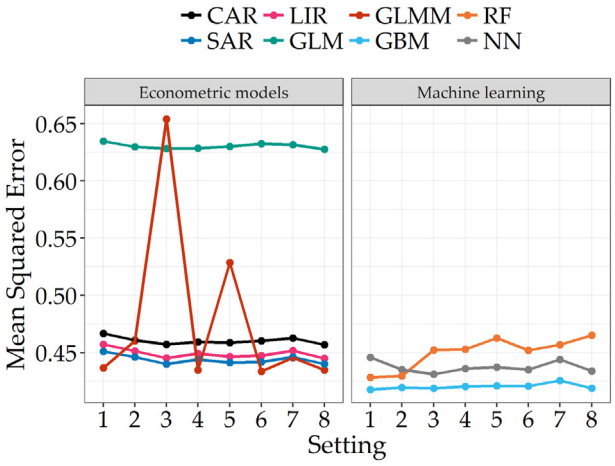


Fig. 4. MSE values of different models for violent crime predictions.

variable importance ranking are in agreement.

We now focus on the predictive results. Figs. 3 and 4 plot the MSE over 13 periods. For property crime, we observe a clear pattern: the MSE is largest across all models for setting 1, which uses demographic variables only, and it decreases upon adding novel features. This provides strong evidence in favour of using novel data sources for property crime prediction. In addition, we observe that some features perform better when used in combination. In particular, settings 3 and 5 use the taxi feature together with POI data (setting 3) or Twitter (setting 5). These settings perform better than the combination of POI data and Twitter data alone. Adding only one feature already improves the predictive accuracy but to a lesser degree compared with adding a combination. Setting 8 using all features together produces the best result. Over all models considered, the MSE in setting 8 is on average 19% lower compared to the baseline setting. This is the largest improvement compared to all other settings, which result in a MSE that is on average 11% lower than the baseline. Clearly, the machine learning models outperform the econometric models across all settings. A Random Forest in setting 8 produces the smallest prediction error. We suggest that the superior performance is driven by non-linear relationships between the features and property crime.

For violent crime, there is a very different trend. As before, the machine learning models perform better than the econometric models but the margin is smaller. With respect to novel data sources, the econometric models slightly improve on their predictions in the baseline setting (setting 1) when having access to the full set of features (setting 8). The machine learning techniques, however, benefit very little from the new features. We observe the lowest prediction error with a GBM in Setting 1. Over all models, the MSE in the other settings is 1% higher than the MSE in setting 1. We conclude that using data on human dynamics and POI offers little advantage for violent crime prediction.

We investigate the robustness of our results for the two best performing models: a RF using setting 8 for property crime and a GBM for violent crime in setting 1. In Fig. 5, we plot the MSE obtained for individual windows hyperparameter configurations during grid search. Each point on the x-axis corresponds to a MSE obtained for a single window and hyperparameter configuration. The vertical line corresponds to the lowest average MSE obtained over all windows as reported in Figs. 3 and 4. Especially for property crime, there is a clear peak regarding the mean MSE for each individual window which means that the predictions are relatively robust against specific hyperparameter settings as they all yield similar results. This provides strong evidence for the superiority of the new features since a wide range of RF produce competitive property crime predictions.

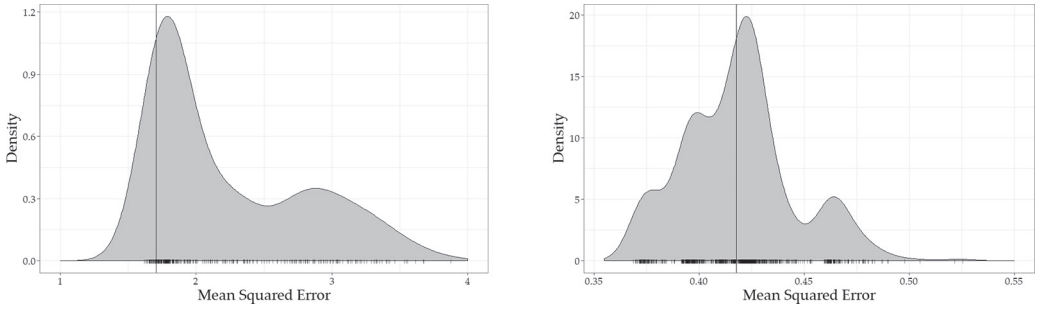
The results for the GBM predicting violent crime are different: the prediction errors are more variable as a function of hyperparameters and windows and the best-performing hyperparameter combinations at each window are more dissimilar than for property crime. Given that these results are obtained with Census data only, the sensitivity to the window choice is not surprising.

6. Discussion

Our mixed approach of explanatory analysis and prediction reflects the dual objective of police and policy makers. We can not only show that crime forecasting benefits from including the novel feature, we also shed light on the emergence of urban crime and find clear support for well-known crime theories. This provides clear guidance on how to conceptualise and address crime in a predictive policing context.

The forecasting results show that using the new features significantly improves the prediction accuracy for property crime. We find that adding static data such as POI venues does not suffice to forecast crime counts accurately. Instead, dynamic Twitter or taxi data and in particular their interaction greatly reduce the prediction error. These results are in line with prior work by Wang et al. [32] and Bendler et al. [5]. We suggest that a combination of node-specific data on the demographic make up as well as the visitor make up through Twitter and POI data in combination with edge-specific data on social taxi flow is the best combination of different data sources to predict property crime counts. The taxi feature proxies human dynamics between areas and how people proliferate crime through space. The spatial dependence matrix models only first-order dependence of immediate neighbours. Many taxi trips traverse multiple areas such that the taxi feature accounts for social connection and crime proliferation beyond just neighbouring sites.

For violent crime, however, the spatio-temporal dimension of the new features adds very little. Our explanatory analysis reveals the origins of this result. Violent crime is taking place in neighbourhoods



(a) Property crime: RF in setting 8. (b) Violent crime: GBM in setting 1

Fig. 5. MSE distribution over hyperparameters and windows.

with poor social cohesion as evident by the positive association with vacant homes and female-headed family households. In line with disorganisation theory, social deprivation provides the context for delinquent, violent behaviour [20]. Support for social disorganisation theory is supplemented by the fact that violent crime counts are not particularly sensitive to POI venues. That long-term structural conditions are more important for violent crime is further emphasised by the poor explanatory and predictive performance of short-term human activity as captured by the novel features.

In contrast, property crime is far less related to the residential make up of the census tract where the crime takes place. Rather than local deprivation, local opportunities through anonymity and vacant homes matter. The coefficients and variable importance rankings capture a trade-off between more opportunities and targets through high human activity on the one hand and more watchful eyes, deterring crime on the other hand. This is for instance illustrated in the negative association of property crime with nightlife and residential venues and the positive association with shopping venues. The notion that different circumstances drive property and violent crime differently is further supported by the rather low correlation between the crime types (Pearson's $r = 0.17$), indicating that the two crime types take place in areas with very different characteristics.

Police react to crime with temporary resource allocations as well as with long-term policy decisions on funding, intervention programmes and task forces. In order to tackle crime, public decision makers depend on both immediate, accurate crime volume forecasts and insight into the underlying crime generating process.

Our explanatory analysis reveals that violent crime emerges from long-standing social environments where short-run movement dynamics do not matter. Based on these results, crime prevention strategies need to account for this spatial and structural difference. Since violent crime is a more slowly-varying process, corresponding crime prevention programmes need to address long-standing issues of re-victimisation and re-offending through youth and family support programmes and partnerships with affected communities.

In contrast, our results indicate that to prevent property crime, police need to be aware of its transitory, changing nature. It is driven by localised opportunities, which means that interventions need to target those intersections of opportunity and offender. In particular, the relevance of the taxi feature demonstrates that in large cities, both offenders and victims cross large distances, propagating crime. This implies that police need to consider not only neighbouring areas but also connections to areas that are further away. Anonymous data on human behaviour can be crucial in identifying these links.

At the time of study, the limited availability of Twitter data constrained the time period of study. Future research can exploit different sources of social media activity and investigate whether similar results hold outside of the United States.

7. Conclusion

This paper presents a multi-model solution to predicting the number of crime incidents in a census tract by combining demographic data with aggregated social media, venue and taxi flow data. In addition, it addresses the two-fold concerns of policy makers: preventing crime in the short run through resource allocation and preventing crime in the medium run through prevention programmes.

Using a rolling-window prediction approach, we provide robust evidence that new features accounting for human activity improves forecasts for crimes shaped by local opportunities. By not only relying on previous crime observations and quinquennial census data but rather on abundantly available behavioural data, the models can generalise to new areas or areas with poor reporting rates.

Following an applied perspective, the proposed approach can be employed to predict future problematic crime areas and improve police responsiveness and resource allocation. By analysing underlying mechanisms of different crime types, promising areas for intervention have been identified.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Grid search parameters

Table A.10 details which parameters were optimised during a grid search. We use early stopping when the MSE does not decrease by at least 0.01% for 5 consecutive scores. Where different, we supply the values used for property and violent crime fitting separately.

Table A.10
Range of grid search values for hyperparameter optimisation.

Model	Parameter	Range of values	
		Property	Violent
GBM	Learn rate	0.01–0.2 with 0.01 increments	
	Learn rate annealing	0.990–0.998 with 0.001 increments	
	Maximum allowed tree depth	13–21	7–15
	Row sample rate	0.20–1 with 0.05 increments	
	Column sample rate	0.20–1 with 0.05 increments	
	Column sample rate per tree	0.20–1 with 0.05 increments	
	Minimum number of rows in a terminal node	4, 8, 16, 32, 64, 128, 256, 512	
	Number of bins used for split	16, 32, 64, 128, 256, 512, 1024	
	Error improvement threshold for split	0, 10^{-8}, 10^{-6}, 10^{-4}	
	Histogram type at each node	Quantiles Global, Round Robin	
	Number of trees	10,000	

(*continued on next page*)

Table A.10 (*continued*)

Model	Parameter	Range of values	
		Property	Violent
NN	Learning rate	Adaptive (ADADELTA)	
	Neurons in hidden layer(s)	64, 128, 256, 512	
	Number of hidden layers	1, 2	
	Epochs	1, 10, 20	
	Learning rate decay	0.95, 1 (no decay)	
RF	Maximum allowed tree depth	11–19	7–15
	Row sample rate	0.20–1 with 0.05 increments	
	Column sample rate	0.20–1 with 0.05 increments	
	Minimum number of rows in a terminal node	4, 8, 16, 32, 64, 128, 256, 512	
	Number of bins used for split	16, 32, 64, 128, 256, 512, 1024	
	Error improvement threshold for split	0, 10^{-8}, 10^{-6}, 10^{-4}	
	Histogram type at each node	Quantiles Global, Round Robin	
	Number of trees	10,000	

Appendix B. Coefficients in rolling window estimation

Since we re-estimate the linear models in each window, we obtain a distribution of coefficients over 13 windows. Since setting 8 includes all variables, we present the coefficients for all models for setting 8.

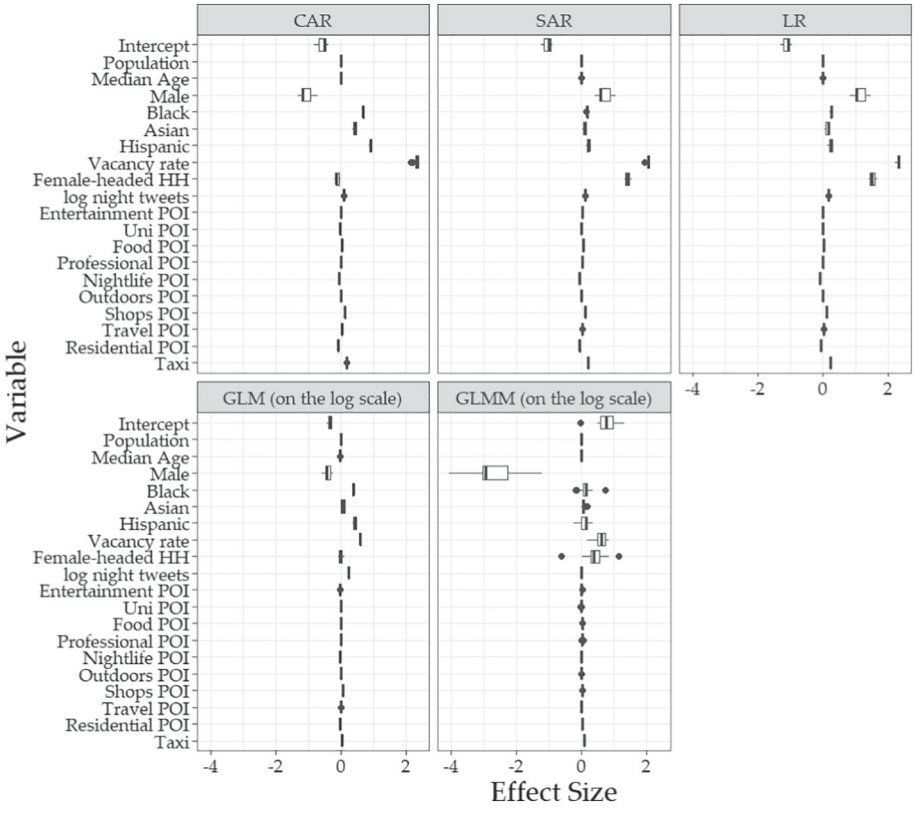


Fig. B.6. Coefficient distribution for property crime for Setting 8 over 13 windows.

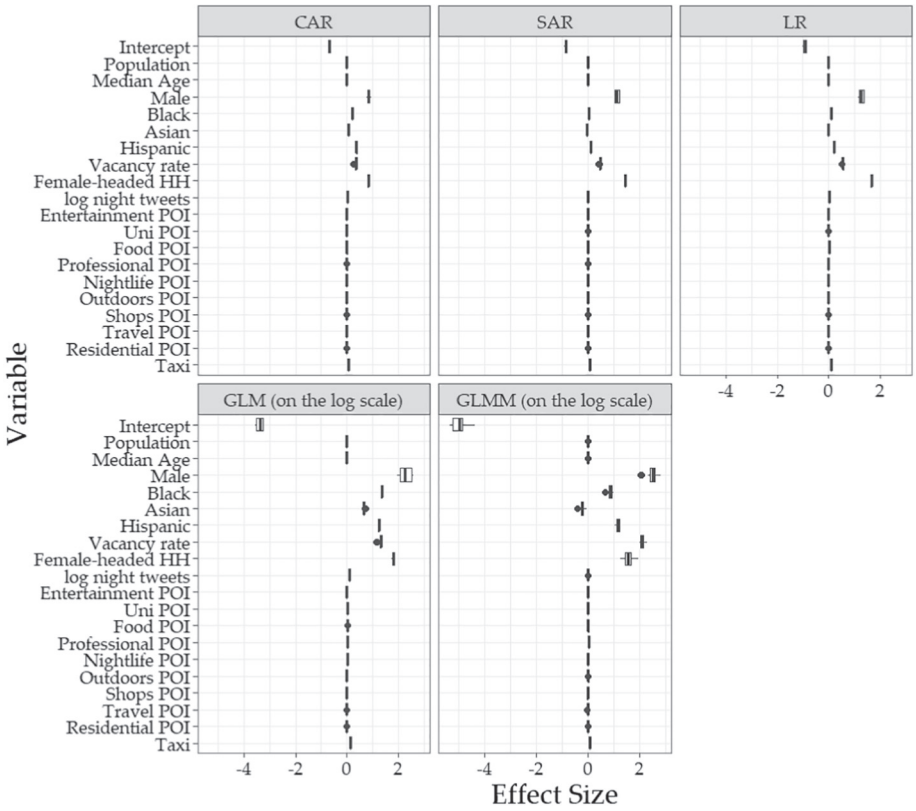


Fig. B.7. Coefficient distribution for violent crime for Setting 8 over 13 windows.

References

[1] S. Aghababaei, M. Makrehchi, Mining Twitter data for crime trend prediction, Intelligent Data Analysis 22 (2018) 117–141.

[2] A. Agresti, An Introduction to Categorical Data Analysis, 2nd ed., John Wiley & Sons, Hoboken, 2007.

[3] L. Anselin, J. Le Gallo, H. Jayet, Spatial panel econometrics, in: L. Mátyás, P. Sevestre (Eds.), The Econometrics of Panel Data, Springer, Berlin, Heidelberg, 2008, pp. 625–660.

[4] B.H. Baltagi, B. Fingleton, A. Pirotte, Estimating and forecasting with a dynamic spatial panel data model, Discussion Paper 95, Spatial Economics Research Centre, 2011.

[5] J. Bendler, A. Ratku, D. Neumann, Crime mapping through geo-spatial social media activity, Proceedings of the 35th International Conference on Information Systems, 2014, pp. 1–16.

[6] S. Bhattacharyya, S. Jha, K. Tharakunnel, J.C. Westland, Data mining for credit card fraud: a comparative study, Decision Support Systems 50 (3) (2011) 602–613.

[7] A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi, A. Pentland, Once upon a crime: towards crime prediction from demographics and mobile data, Proceedings of the 16th International Conference on Multimodal Interaction, 2014, pp. 427–434.

[8] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32.

[9] M. Camacho-Collados, F. Liberatore, A decision support system for predictive police patrolling, Decision Support Systems 75 (2015) 25–37.

[10] N. Carneiro, G. Figueira, M. Costa, A data mining based system for credit-card fraud detection in e-tail, Decision Support Systems 95 (2017) 91–101.

[11] L.E. Cohen, M. Felson, Social change and crime rate trends: a routine activity approach, American Sociological Review 44 (1979) 588–608.

[12] N. Cressie, Statistics for Spatial Data, John Wiley & Sons, Hoboken, 1993.

[13] D. Delen, A comparative analysis of machine learning techniques for student retention management, Decision Support Systems 49 (2010) 498–506.

[14] J.H. Friedman, Stochastic gradient boosting, Computational Statistics & Data Analysis 38 (2002) 367–378.

[15] M.S. Gerber, Predicting crime using Twitter and kernel density estimation, Decision Support Systems 61 (2014) 115–125.

[16] H.-W. Kang, H.-B. Kang, Prediction of crime occurrence from multi-modal data using deep learning, PloS one 12 (2017) 1–19.

[17] J. Kim, P. Kang, Late payment prediction models for fair allocation of customer contact lists to call center agents, Decision Support Systems 85 (2016) 84–101.

[18] T. Kneib, Restricted Maximum Likelihood Estimation of Variance Parameters in Generalized Linear Mixed Models, (2003) https://www.uni-goettingen.de/de/304966.html retrieved 15/02/2017.

[19] C.E. Kubrin, Structural covariates of homicide rates: does type of homicide matter? Journal of Research in Crime and Delinquency 40 (2003) 139–170.

[20] C.E. Kubrin, R. Weitzer, New directions in social disorganization theory, Journal of Research in Crime and Delinquency 40 (2003) 374–402.

[21] C. Kuzey, A. Uyar, D. Delen, The impact of multinationality on firm value: a comparative analysis of machine learning techniques, Decision Support Systems 59 (2014) 127–142.

[22] B.G. Leroux, X. Lei, N. Breslow, Estimation of disease rates in small areas: a new mixed model for spatial dependence, in: M. Halloran, D. Berry (Eds.), Statistical Models in Epidemiology, the Environment, and Clinical Trials, Springer, New York, 2000, pp. 179–191.

[23] N. Malleson, M.A. Andresen, The impact of using social media data in crime rate calculations: shifting hot spots and changing spatial patterns, Cartography and Geographic Information Science 42 (2015) 112–121.

[24] J.D. Morenoff, R.J. Sampson, S.W. Raudenbush, Neighborhood inequality, collective efficacy, and the spatial dynamics of urban violence, Criminology 39 (2001) 517–558.

[25] New York City Police Department, NYPD Complaint Map (Historic), (2016) https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Map-Historic-/57mv-nv28 retrieved 31/01/2017.

[26] NYC Taxi & Limousine Commission, TLC Trip Record Data, (2016) http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml retrieved 02/01/2017.

[27] J. Pfeffer, F. Morstatter, Geotagged Twitter Posts From the United States: A Tweet Collection to Investigate Representativeness, (2016) http://doi.org/10.7802/1166 retrieved with permission 10/02/2017.

[28] G. Rosser, T. Davies, K.J. Bowers, S.D. Johnson, T. Cheng, Predictive crime mapping: arbitrary grids or street networks? Journal of Quantitative Criminology 33 (2017) 569–594.

[29] R.J. Sampson, S.W. Raudenbush, F. Earls, Neighborhoods and violent crime: a multilevel study of collective efficacy, Science 277 (1997) 918–924.

[30] M. Traunmueller, G. Quattrone, L. Capra, Mining mobile phone data to investigate urban crime theories at scale, International Conference on Social Informatics, 2014, pp. 396–411.

[31] U.S. Census Bureau, Profile of General Population and Housing Characteristics: 2010 Census, (2017) https://www.census.gov/data/datasets/2010/dec/summary-file-1.html retrieved 31/01/2017.

[32] H. Wang, D. Kifer, C. Graif, Z. Li, Crime rate inference with big data, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 635–644.

[33] P.-O.H. Wikström, V. Ceccato, B. Hardie, K. Treiber, Activity fields and the dynamics of crime, Journal of Quantitative Criminology 26 (2010) 55–87.

[34] M.L. Williams, P. Burnap, L. Sloan, Crime sensing with big data: the affordances and limitations of using open-source communications to estimate crime patterns, The

British Journal of Criminology 57 (2017) 320–340.
[35] Y. Xue, D.E. Brown, Spatial analysis with preference specification of latent decision makers for criminal event prediction, Decision Support Systems 41 (2006) 560–573.

Lara Vomfell received her bachelor's degree in Political Science and Economics from the University of Muenster and her master's degree in Economics from the University of Berlin. Since 2017, she is a PhD candidate in Behavioural Science at the Warwick Business School at the University of Warwick. Her research interest is in investigating patterns in police and crime data. In particular, she is working on evidence of racial bias in stop and search as well as crime deterrence and displacement effects of stop and search. As part of the Leverhulme Bridges programme, she is interested in developing and deploying more statistically sophisticated models of crime.

Wolfgang Karl Härdle has been director of the Ladislaus von Bortkiewicz Chair of Statistics at the Department of Economics and Business Administration at the Humboldt University of Berlin since 1992. He is Coordinator of the “Collaborative Research Center 649: Economic Risk”. Since October 2013 he has also headed the newly established International Research Training Group, a joint project with Xiamen University in China. His research interests are smoothing methods, discrete choice models, statistical modelling of financial markets and computer-aided statistics. His more recent work deals with the modelling of implied volatilities and the statistical analysis of financial risk. Since February 2014, he is a member of the Integrative Research Institute on Transformations of Human-Environment Systems (IRI THESys).

Stefan Lessmann received a diploma in business administration and a PhD from the University of Hamburg in 2002 and 2007, respectively. He joined the Humboldt University of Berlin in 2014, where he heads the Chair of Information Systems at the School of Business and Economics. His work focuses on the analysis and support managerial decision making. Much of his research is concerned with the development, application, and validation of empirical prediction models. The degree to which such models actually support managers and how to improve their alignment with managers' requirements represent typical research questions. This includes research on the following topics: artificial neural networks, credit risk modelling, ensemble models and forecast combination. He actively participates in knowledge transfer and consulting projects with industry partners; from small start-up companies to global players.

Journal of Empirical Finance 49 (2018) 107–122

Contents lists available at ScienceDirect

Journal of Empirical Finance

journal homepage: www.elsevier.com/locate/jempfin

CRIX an Index for cryptocurrencies

Simon Trimborn [a,b,*], Wolfgang Karl Härdle [a,c]

[a] Humboldt-Universität zu Berlin, C.A.S.E. - Center for Applied Statistics and Economics, Spandauer Str. 1, 10178 Berlin, Germany
[b] Department of Statistics and Applied Probability, National University of Singapore, 21 Lower Kent Ridge Rd, Singapore 119077, Singapore
[c] SKBI School of Business, Singapore Management University, 50 Stamford Road, Singapore 178899, Singapore

ARTICLE INFO

JEL classification:
C51
C52
G10

Keywords:
Index construction
Model selection
Bitcoin
Cryptocurrency
CRIX
Altcoin

ABSTRACT

The cryptocurrency market is unique on many levels: Very volatile, frequently changing market structure, emerging and vanishing of cryptocurrencies on a daily level. Following its development became a difficult task with the success of cryptocurrencies (CCs) other than Bitcoin. For fiat currency markets, the IMF offers the index SDR and, prior to the EUR, the ECU existed, which was an index representing the development of European currencies. Index providers decide on a fixed number of index constituents which will represent the market segment. It is a challenge to fix a number and develop rules for the constituents in view of the market changes. In the frequently changing CC market, this challenge is even more severe. A method relying on the AIC is proposed to quickly react to market changes and therefore enable us to create an index, referred to as CRIX, for the cryptocurrency market. CRIX is chosen by model selection such that it represents the market well to enable each interested party studying economic questions in this market and to invest into the market. The diversified nature of the CC market makes the inclusion of altcoins in the index product critical to improve tracking performance. We have shown that assigning optimal weights to altcoins helps to reduce the tracking errors of a CC portfolio, despite the fact that their market cap is much smaller relative to Bitcoin. The codes used here are available via www.quantlet.de.

1. Introduction

More and more companies have started offering digital payment systems. Smartphones have evolved into a digital wallet, telephone companies offer banking related services: clear signal that we are about to enter the era of digital finance. In fact we are already acting inside a digital economy. The market for e-x (x = "finance", "money", "book", you name it ...) has not only picked up enormous momentum but has become standard for driving innovative activities in the global economy. A few clicks at y and payment at z brings our purchase to location w. Own currencies for the digital market were therefore just a matter of time. Due to organizational difficulties the idea of the Nobel Laureate Hayek, see Hayek (1990), of letting companies offer concurrent currencies seemed for a long time scarcely feasible, but the invention of the *Blockchain* has made it possible to bring his vision to life. Cryptocurrencies (CCs) have surfaced and opened up an angle towards this new level of economic interaction. Since the appearance of Bitcoins, several new CCs have spread through the Web and offered new ways of proliferation. Even states accept them as a legal payment method or part of economic interaction. E.g., the USA classifies CCs as commodities, Kawa (2015), and lately Japan announced that they accept them as a legal currency, EconoTimes (2016). Obviously, the crypto market is fanning out and shows clear signs of acceptance and deepening liquidity, so that a closer look at its general moves and dynamics is called for.

* Corresponding author at: Department of Statistics and Applied Probability, National University of Singapore, 21 Lower Kent Ridge Rd, Singapore 119077, Singapore.
E-mail addresses: simon.trimborn@nus.edu.sg (S. Trimborn), haerdle@hu-berlin.de (W.K. Härdle).

https://doi.org/10.1016/j.jempfin.2018.08.004
Received 1 May 2017; Received in revised form 8 August 2018; Accepted 31 August 2018
Available online 6 October 2018

The transaction graph of Bitcoin (BTC), the Blockchain, has received much attention, see e.g. Ron and Shamir (2013) and Reid and Harrigan (2013). Even the economics of BTC has been studied, e.g. Bolt and Oordt (2016) and Kristoufek (2015). To our best knowledge, the development of the entire CC market has not been studied so far, only subsamples have been taken into account. Wang and Vergne (2017) studied the variations of 5 CCs. Elendner et al. (2017) analyzed the top 10 CCs by market capitalization and found that their returns are weakly correlated with each other. Furthermore, a Principal Component (PC) Analysis, carried out in the same reference, showed 7 out of 10 PC were necessary to describe more than 90% of the variance. These findings indicate the price evolution of CCs is very different from each other. This brings us to the conclusion that BTC, even though it dominates the market in terms of its market capitalization, cannot lead the direction of the market. The movements of other CCs are important too, when one analyzes the market. Having a closer look at the different CCs, it becomes obvious they have different kind of missions and technical aspects. Bitcoin pioneered as the token of the first decentralized, distributed ledger, giving start to multiple interpretations of its nature and purpose: new type of currency, commodity (like gold), alternative asset or innovative technology. The currently second most important CC by market capitalization—Ethereum – was created with a particular goal in mind – to power the blockchain based Ethereum platform for company building (DAO) and smart contract implementation. This idea triggered an unprecedented interest as it allowed companies to enter the field without creating their own blockchain ecosystem. Newcomers could benefit from the existing supporters of the respective platform, which allowed faster entry, adoption and operation. Other CCs, like Ripple (XRP), are intended to fuel the transaction network bridging traditional markets (banks) and the crypto ecosystem. Ripple also became one of the first successful cases of pre-emitted CC, abandoning the idea of decentralization. Since the appearance of BTC many technological advancements took place. Some CCs are designed for faster (or even immediate) transactions, like Litecoin (LTC), some are more efficient energy-wise, like DASH. Many embraced different hashing algorithms, altering the mining process, like Monero. Long ASIC domination is being disrupted, Proof-of-work is replaced by Proof-of-Stake, new ways to motivate those providing computational power are introduced. Regardless the type of CC, one witnesses a new kind of transaction network with a different approach for fees and handling of trust issues. The intended and actual usage can be interpreted as the business model of the different CCs and the participation in either CC can give advantages over others, White (2014).

In the first month of 2017, CCs other than BTC (altcoins) showed a strong gain in their market capitalization, reducing the dominance of BTC in the market. The finding of very different movements of CCs and the stronger position of alternative CCs in the market infers the necessity of a market index for the CC market for tracking the market movements. Comparing CCs against a market index answers economic questions like which business model is more successful than another one, gained recently compared to other CCs, drives the success of the market, is more established. Comparing a CC market index against other market indices answers economic and financial questions like which market proxy is more volatile, has more tail risk, attracts more investments. We construct CRIX, a market index (benchmark) which will enable each interested party to study the outlined economic questions, the performance of the CC market as a whole or of single CCs. Studying the stochastic dynamics of CRIX will allow a la limite to create ETFs or contingent claims.

Many index providers construct their indices with a fixed number of constituents, see e.g. FTSE (2016), S&P (2014) and Deutsche Boerse AG (2013). If the respective index is intended to be a proxy for the performance of a market, this requires huge trust from economists and investors into the choice of the index constituents by the index provider. On the other hand, the CRSP index family, derived for the US market, CRSP (2015), has no boundary on the number of index constituents. The number of constituents is reviewed daily and adjusted until the index members cover a predefined share of the market capitalization. Such a dynamic methodology is important in the market of CCs since the number of CCs changes daily. Additionally the market value of CCs often changes frequently, which increases the market volatility and therefore the need for considering such a CC for the representation of the market. Our intention is extending the idea behind the CRSP indices. Our first goal is constructing a methodology for CRIX which relies on model selection criteria to receive a proxy for the market and to replace the trust problematic with a statistical methodology. The resulting methodology is dynamic in the number of index constituents, like the CRSP indices. By this method only CCs which add informative value to the index are considered, which makes it representative. If more CCs than BTC are necessary to fulfill this requirement, they will be added. However we are concerned with the dominance of BTC in an index solely relying on market capitalization. Thus we introduce a second weighting scheme based on weighting by trading volume. Due to the usage of trading volume, the respective index is constructed in terms of trading focus. If the market participants focus more on altcoins than on BTC, these receive a higher weight. On the other hand, if the market focus is truly on BTC, it will receive a high weight in either index. Our second goal, constructing an investable index will be fulfilled by the methodology itself due to having a sparse index, only consisting of actively traded CCs in a market with low transaction costs. Note that due to the low transaction costs in the CC market, a dynamic methodology creates low additional costs. Additionally to the methodology ensuring an investable index, the proposed trading volume weighting scheme further supports this goal.

Investing into an ETF composed of the constituents of CRIX implies some differences compared to traditional index investing. In the traditional setting only the constituents are reviewed and replaced on the review date – if necessary – according to the index rules. In dynamic index investing the constituents are also reviewed for their number. This requires the manager of the fund to buy and sell more assets on the review date. In a market with high transaction costs, this approach is more costly. But the market of CCs has very low transaction costs, thus this problem will not occur in this market.

To compute CRIX, the differences in the log returns of the market against a selection of possible indices is evaluated. The results show, that the AIC works well to evaluate the differences. It penalizes the index for the number of constituents. For the calculation of the respective likelihoods, a non-parametric approach using the Epanechnikov (1969) kernel is applied. The proof for the impact of the value of an asset in the market on the AIC method is given, thus a top-down approach is applied to select the assets for the benchmarks to choose from, where the sorting depends on either market cap or trading volume. The number of constituents is

recalculated quarterly to ensure an up-to-date fit to the current market situation. With CRIX one may study the contingent claims and the stochastic nature of this index, Chen et al. (2017), or study the CC market characteristics against traditional markets, Härdle and Trimborn (2015).

This paper is structured as follows. Section 2 introduces the topic and reviews the basics of index construction. In Section 3 the method for dynamic index construction for CRIX is described and Section 4 introduces the remaining rules for CRIX. Section 5 describes further variants to create a CRIX family. Their performance is tested in Section 6. In Sections 7 and 8 the new method is applied to the German and Mexican stock markets to check the performance of the methodology against existing indices. The codes used to obtain the results in this paper are available via www.quantlet.de.

2. Index construction

The basic idea of any price index is to weight the prices of its constituent goods by the quantities of the goods purchased or consumed. The Laspeyres index takes the value of a basket of k assets and compares it against a base period:

$$P_{0t}^{L}(k) = \frac{\sum_{i=1}^{k} P_{it} Q_{i0}}{\sum_{i=1}^{k} P_{i0} Q_{i0}} \tag{1}$$

with P_{it} the price of asset i at time t and Q_{i0} the quantity of asset i at time 0 (the base period). For market indices, such as CRSP, S&P500 or DAX, the quantity Q_{i0} is the number of shares of the asset i in the base period. Multiplied with its corresponding price, the market capitalization results, hence the constituents of the index are weighted by their market capitalizations. These indices are often referred to as benchmarks for their respective market. We define the term benchmark:

Definition 1. A benchmark is a measure which consists of a selection of CCs that are representing the market.

But markets change. A company which was representative for market developments yesterday might no longer be important today. On top of that, companies can go bankrupt, a corporation can raise the number of its outstanding shares, or trading in it can become infrequent. All these situations must produce a change in the index structure, so that the market is still adequately represented. Hence companies have to drop out of the index and have to be replaced by others. The index rules determine in which cases such an event happens. The formula of Laspeyres (1) cannot handle such events entirely because a change of constituents will result in a change in the index value that is not due to price changes. Therefore, established price indices like DAX or S&P500, see Deutsche Boerse AG (2013) and S&P (2014) respectively, and the newly founded index CRIX(k), a CRyptocurrency IndeX, thecrix.de, use the adjusted formula of Laspeyres,

$$\mathrm{CRIX}_t(k,\beta) = \frac{\sum_{i=1}^{k} \beta_{i,t_l^-} P_{it} Q_{i,t_l^-}}{Divisor(k)_{t_l^-}} \tag{2}$$

with P, Q and i defined as before, β_{i,t_l^-} the adjustment factor of asset i found at time point t_l^-, l indicates that this is the lth adjustment factor, and t_l^- the last time point when Q_{i,t_l^-}, $Divisor(k)_{i,t_l^-}$ and β_{i,t_l^-} were updated. In the classical setting, β_{i,t_l^-} is defined to be $\beta_{i,t_l^-} = 1$ for all i and l. Anyhow, some indices use β_{i,t_l^-} to achieve maximum weighting rules, e.g. Deutsche Boerse AG (2013) and MEXBOL (2013). The *Divisor* ensures that the index value of CRIX has a predefined value on the starting date. It is defined as

$$Divisor(k,\beta)_0 = \frac{\sum_{i=1}^{k} \beta_{i0} P_{i0} Q_{i0}}{\text{starting value}}. \tag{3}$$

The starting value could be any possible number, commonly 100, 1000 or 10000. It ensures that a positive or negative development from the base period will be revealed. Whenever changes to the structure of CRIX occur, the *Divisor* is adjusted in such a way that only price changes are reflected by the index. Defining k_1 and k_2 as number of constituents, it results

$$\frac{\sum_{i=1}^{k_1} \beta_{i,t_{l-1}^-} P_{i,t-1} Q_{i,t_{l-1}^-}}{Divisor(k_1,\beta)_{t_{l-1}^-}} = \mathrm{CRIX}_{l-1}(k_1,\beta) = \mathrm{CRIX}_t(k_2,\beta) = \frac{\sum_{j=1}^{k_2} \beta_{j,t_l^-} P_{j,t} Q_{j,t_l^-}}{Divisor(k_2,\beta)_{t_l^-}}. \tag{4}$$

In indices like FTSE, S&P500 or DAX the number of index members is fixed, $k_1 = k_2$, see FTSE (2016), S&P (2014) and Deutsche Boerse AG (2013). As long as the goal behind these indices is the reflection of the price development of the selected assets, this is a straightforward approach. But, e.g., DAX is also meant to be an indicator for the development of the market as a whole, see Janßen and Rudolph (1992). This raises automatically the question of whether the included assets and the weighting scheme are representing the market. Since the constituents are chosen using a top-down approach, meaning that the biggest companies by market capitalization are included, the intuitive answer is yes. But it leaves a sour taste that additional assets may describe the market more appropriately. Furthermore different weighting schemes provide another view on the market. One may object by referring to total market indices like the Wilshire 5000, S&P Total Market Index or CRSP U.S. Total Market Index, see Wilshire Associates (2015), S&P (2015) and CRSP (2015), that are providing a full description. But financial practice has shown that smaller indices like DAX30 and S&P500 receive more attention in evaluating the movements of their corresponding markets, probably because they are easier to invest in due to the smaller number of constituents. It is therefore appealing to know which are the representative assets in a market and which smaller number of index constituents eases the handling of a tracking portfolio. Additionally, one may be concerned that an index would include illiquid and non-investable assets which makes the management of a tracking portfolio even more difficult. Fig. 1 shows that this is indeed a problem in the CC market. Some CCs have a fairly high market capitalization while their respective trading volume

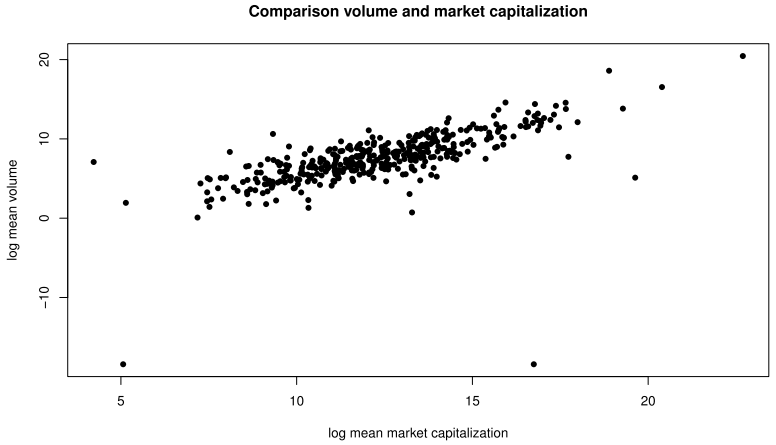


Fig. 1. Comparison of the log mean trading volume and log mean market capitalization, both measured in USD, for all CCs in the dataset over the time period 20140401–20170325. VolMarketCapComparison.

Table 1
Weighting schemes for derivation of CRIX.

	Market cap weighting	Liquidity weighting
β_{i,t_l^-}	1	$\frac{Vol_{i,t_l^-}}{P_{i,t_l^-}Q_{i,t_l^-}}$

is very low. This is problematic, because an asset which is not frequently traded cannot add enough information to a market index to display market changes and is difficult to trade for an investor. Hence, one goal behind constructing CRIX is making it investable by concentrating on liquid CCs:

Definition 2. Between investment portfolios with equal performance, the one with the least assets is preferable.

We react to the goals and problems in two ways: First, these thoughts raise the question which value of k is "optimal" for building an investable benchmark for the market. Additionally, especially young and innovative markets may change their structure over time. Therefore, a quantification of an accurate CC benchmark with sparse number of constituents is asked for. Since the CC market shows a frequently changing market structure with a huge number of illiquid CCs, a time varying index selection structure is applied. The later described selection method omits illiquid CCs by construction, because only CCs who show changes in their return series can be selected to be added to CRIX by the method. Due to the low transaction costs in this market, a dynamic methodology is applicable since it does not raise the costs of restructuring a tracking portfolio too much. Secondly, we apply two kind of weighting schemes, Table 1. We apply the classical setting to build a proper market index which is only flexible in terms of the dynamic constituents and tackles the illiquidity issue due to the applied selection method. The liquidity weighting allows one to weight CCs higher, which are more traded relative to their market capitalization and therefore implicitly acquire more financial attention. This weighting scheme bails (2) down to weighting the price development by their trading volume,

$$\text{LCRIX}_t(k,\beta) = \frac{\sum_{i=1}^{k} \frac{Vol_{i,t_l^-}}{P_{i,t_l^-}Q_{i,t_l^-}} P_{it} Q_{i,t_l^-}}{Divisor(k)_{t_l^-}} = \frac{\sum_{i=1}^{k} \frac{Vol_{i,t_l^-}}{P_{i,t_l^-}} P_{it}}{Divisor(k)_{t_l^-}}. \tag{5}$$

The latter is referred to as Liquidity CRIX (LCRIX). This approach has the potential to diminish the influence of e.g. Bitcoin stronger than the market cap weighting, if the relation of trading volume to market cap is higher for other CCs. In Section 6 we show that LCRIX has a better mean directional accuracy than CRIX and puts more weight on altcoins, Table 8, therefore tackling the issue of BTC dominance when the actual trading amount suggests a different result.

3. Dynamic index construction

This section is dedicated to describing the composition rule which is used to find the number of index members—the spine of CRIX and LCRIX. Since CRIX will be a benchmark for the CC market, the dimension and evaluation of the market has to be defined:

Definition 3. The total market (TM) consists of all CCs in the CC universe. Its value is the combined market value of the CCs.

To compare the TM with a benchmark candidate, it will be normalized by a Divisor,

$$\text{TM}(K)_t = \frac{\sum_{i=1}^{K} P_{it} Q_{i,t_l^-}}{Divisor(K)_{t_l^-}} \tag{6}$$

with K the number of all CCs in the CC universe. Note that no adjustment factor is used for $\text{TM}(K)_t$. For the volume weighting, the TM is defined as LTM respectively,

$$\text{LTM}(K)_t = \frac{\sum_{i=1}^{K} \frac{Vol_{i,t_l^-}}{P_{i,t_l^-} Q_{i,t_l^-}} P_{it} Q_{i,t_l^-}}{Divisor(K)_{t_l^-}}. \tag{7}$$

In the further explanations, the focus lies on the TM. However when LCRIX is derived, it is optimized against LTM. The results can be easily extended to the case of LTM. Further define the log returns:

$$\varepsilon(K)_t^{TM} = \log\{\text{TM}(K)_t\} - \log\{\text{TM}(K)_{t-1}\} \tag{8}$$

$$\varepsilon(k,\beta)_t^{CRIX} = \log\{\text{CRIX}(k,\beta)_t\} - \log\{\text{CRIX}(k,\beta)_{t-1}\}, \tag{9}$$

where $\text{CRIX}(k,\beta)_t$ is the CRIX with k constituents at time point t.

The goal is to optimize k and β so that a sparse but accurate approximation in terms of

$$\min_{k,\beta} \|\varepsilon(k,\beta)\|^2 = \min_{k,\beta} \|\varepsilon(K)^{TM} - \varepsilon(k,\beta)^{CRIX}\|^2, \tag{10}$$

is achieved, where $\varepsilon(k,\beta)$ is the difference in the log returns of $\text{TM}(K)$ and $\text{CRIX}(k,\beta)$. A squared loss function is chosen in (10), since it heavily penalizes deviations.

Since the value of $\text{TM}(K)_t$ is unknown and not measurable due to a lack of information, the total market index will be defined and used as a proxy for the $\text{TM}(K)$. The definition is inspired by total market indices like CRSP (2015), S&P (2015) and Wilshire Associates (2015). They use all stocks for which prices are available.

Definition 4. The total market index (TMI) contains all CCs in the CC universe for which prices are available. The CCs are weighted by their market capitalization.

This changes (6) to

$$\text{TMI}_t(k_{max}) = \frac{\sum_{i=1}^{k_{max}} P_{it} Q_{i,t_l^-}}{Divisor(k_{max})_{t_l^-}}$$

with k_{max} the maximum number of CCs with available prices and (10) to

$$\min_{k,\beta} \|\widehat{\varepsilon}(k,\beta)\|^2 = \min_{k,\beta} \|\varepsilon(k_{max})^{TMI} - \varepsilon(k,\beta)^{CRIX}\|^2 \tag{11}$$

$$\begin{aligned} \text{s.t.: } & 1 \le k \le k^u \\ & k = k_1 + s \\ & k^u \in [1, k_{max}] \\ & s \in [1, k_{max} - k_1] \\ & \beta^{1\times k} = (1, \dots, 1, \beta_{k_1+1}, \dots, \beta_{k_1+s})^\top \\ & \beta_{k_1+1}, \dots, \beta_{k_1+s} \in (-\infty, \infty), \end{aligned} \tag{12}$$

where $\varepsilon(k_{max})^{TMI}$ are the log returns for TMI. In the derivation of LCRIX, the optimization is performed against LTMI and $\beta^{1\times k} = (\beta_1, \dots, \beta_k, \beta_{k_1+1}, \dots, \beta_{k_1+s})^\top$ where $\beta_i = \frac{Vol_{i,t_l^-}}{P_{i,t_l^-} Q_{i,t_l^-}}$ for $i = 1, \dots, k_1$ and $\beta_{k_1+1}, \dots, \beta_{k_1+s} \in (-\infty, \infty)$.

Several constraints were introduced with (11). The parameters $\beta_{k+1}, \dots, \beta_{k+s}$ are included to evaluate if adding s more assets to the index explains the difference between $\varepsilon(k_{max})^{TMI}$ and $\varepsilon(k,\beta)^{CRIX}$ better. The first k assets (k_1) will not be adjusted by a parameter, so no parameter estimation is necessary. This makes the first term a constant. The choice of k_1 is important since it defines the number of base CCs to be included in the index. The parameters of the next s assets have to be estimated, so (2) becomes

$$\text{CRIX}_t(k,\beta) = \frac{\sum_{i=1}^{k_1} P_{it} Q_{i,t_l^-} + \sum_{j=k_1+1}^{k_1+s} \beta_{j,t_l^-} P_{jt} Q_{j,t_l^-}}{Divisor(k_1)_{t_l^-}}.$$

A number of criteria are applicable. Model selection (SC) criteria can be categorized by their property to be either asymptotic optimal or consistent in choosing the true model. In this context will be investigated: Generalized Cross Validation (GC), Generalized Full Cross Validation (GFC), Mallows' C_p, Shibata (SH), Final Prediction Error (FPE) and Akaike Information Criterion (AIC), all asymptotic optimal criteria under the assumption of Gaussian distributed residuals. Since CRIX is supposed to be a benchmark model, all possible models under certain restrictions for the number of parameters are included in the test set,

$$\Theta_{SC} = \{\text{CRIX}(k_1,\beta), \text{CRIX}(k_2,\beta), \dots\}, \tag{13}$$

where $k_1, k_2, \dots$ are predefined values and $SC \in \{\text{GC}, \text{GFC}, C_p, \text{SH}, \text{FPE}, \text{AIC}\}$. Recall that the intention behind CRIX is to discover under a squared loss function the best model to describe the data (benchmark), which supports the choice of an asymptotic optimal criteria. The GC criterion, see Craven and Wahba (1978), is defined as

$$\text{GC}\{\widehat{\varepsilon}(k,\beta), s\} = \frac{T^{-1} \sum_{t=1}^{T} \widehat{\varepsilon}(k,\beta)_t^2}{(1 - T^{-1}s)^2} \tag{14}$$

by assuming that $s < T$. One shall note that s and not $k + s$ defines the number of variables to penalize for, since k parameters are set to be 1 and need not be estimated. According to Arlot and Celisse (2010), the asymptotic optimality of GC was shown in several frameworks. The GFC, see Droge (1996):

$$\mathrm{GFC}\{\widehat{\varepsilon}(k,\beta),s\} = T^{-1}\sum_{t=1}^{T}\widehat{\varepsilon}(k,\beta)_t^2(1+T^{-1}s)^2 \tag{15}$$

is an alteration.

A further score, SH,

$$\mathrm{SH}\{\widehat{\varepsilon}(k,\beta),s\} = \frac{T+2s}{T^2}\sum_{t=1}^{T}\widehat{\varepsilon}(k,\beta)_t^2, \tag{16}$$

was shown to be asymptotically optimal, Shibata (1981), and asymptotically equivalent to Mallows' C_p and AIC.

Mallows (1973)' C_p:

$$\mathrm{C}_p\{\widehat{\varepsilon}(k,\beta),s\} = \frac{\sum_{t=1}^{T}\widehat{\varepsilon}(k,\beta)_t^2}{\widehat{\sigma}(k,\beta)^2} - T + 2\cdot s \tag{17}$$

with $\widehat{\sigma}(k,\beta)^2$ the variance of $\widehat{\varepsilon}(k,\beta)$. $C_p\{\widehat{\varepsilon}(k,\beta),s\}$ tends to choose models which overfit and is not consistent in selecting the true model, see Mallick and Yi (2013), Woodroofe (1982) and Nishii (1984).

The FPE uses the formula

$$\mathrm{FPE}\{\widehat{\varepsilon}(k,\beta),s\} = \frac{T+s}{(T-s)T}\sum_{t=1}^{T}\widehat{\varepsilon}(k,\beta)_t^2, \tag{18}$$

see Akaike (1970)

So far, the discussed criteria depend on little data information. Just the squared residuals and, in the case of Mallows' C_p, the variance are taken into account. The AIC uses more information by depending on the maximum likelihood, derived by

$$L\{\widehat{\varepsilon}(k,\beta)\} = \max_{\beta}\prod_{t} f\{\widehat{\varepsilon}(k,\beta)_t\}, \tag{19}$$

where f, in (21), represents the density of the $\widehat{\varepsilon}(k,\beta)_t$ over all t. The AIC is defined to be

$$\mathrm{AIC}\{\widehat{\varepsilon}(k,\beta),s\} = -2\log L\{\widehat{\varepsilon}(k,\beta)\} + s\cdot 2, \tag{20}$$

Akaike (1998). If the true model is of finite dimension, then the AIC is not consistent, compare Hurvich and Tsai (1989). Shibata (1983) showed the asymptotic efficiency of Mallows' C_p and AIC under the assumption of an infinite number of regression variables or an increasing number of regression variables with the sample size. Due to the usage of the density in deriving the AIC, it uses more information about the dataset. Considering that (10) implies the criteria are derived under an expected squared loss function,

$$\mathrm{E}(\|\varepsilon(k,\beta)\|^2) = \int_{-\infty}^{\infty}\|\varepsilon(k,\beta)\|_2^2 f\{\varepsilon(k,\beta)\}d\varepsilon(k,\beta), \tag{21}$$

the density, f, can be estimated different from the Gaussian distribution. Here, f is estimated nonparametrically with an Epanechnikov kernel, since according to Härdle et al. (2004) the Epanechnikov (1969) kernel shows a good balance between variance optimization and numerical performance. In nonparametric estimation with an Epanechnikov kernel, Epa, the estimator of f is derived by

$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n}\mathrm{Epa}\left(\frac{x-x_i}{h}\right), \quad \mathrm{Epa}(u) = \frac{3}{4\sqrt{5}}\left(1-\frac{u^2}{5}\right)\mathbf{I}(|u|\le\sqrt{5})$$

where h is the bandwidth.

The bandwidth selection is performed with the plug-in selector by Sheather and Jones (1991) and further described in Wand and Jones (1994). The plug-in selector is derived under the loss function Mean Integrated Squared Error, MISE. Hall (1987) found that the Kullback–Leibler (KL) loss function for selecting the smoothing parameter of the kernel density is highly influenced by the tails of the distribution. Devroye and Györfi (1985) mention that Mean Integrated Error (MIE) is stronger affected than MISE by the tails of the distribution and Kanazawa (1993) claims that MIE shall be used if interest is in modeling the tails. Kanazawa (1993) investigates that the use of a Kullback–Leibler loss function would put more weight on the tails compared to MISE. Since this is not in our interest, the choice of the density smoothing parameter, h, is performed under MISE.

Due to the richer information basis of the AIC, we decide to use it as the selection criteria for CRIX. The choice is supported by an empirical analysis in Section 6.

To decide with the AIC which number k should be used, a procedure was created which compares the squared difference between log returns of the TMI, see Definition 4, and several candidate indices,

$$\|\widehat{\varepsilon}(k_j,\beta)\|^2 = \|\varepsilon(k_{max})^{TMI} - \varepsilon(k_j,\beta)^{CRIX}\|^2, \tag{22}$$

where $\varepsilon(k_j,\beta)^{CRIX}$ is the log return of CRIX version with k_j constituents and $\widehat{\varepsilon}(k_j,\beta)$ is the respective difference. The candidate indices, $\mathrm{CRIX}(k_j,\beta)$, have different numbers of constituents which fulfill $k_1 < k_2 < k_3 < \cdots$, where $k_j = k_1 + s(j-1)$. Therefore, the

number of constituents between the indices are equally spaced. The procedure implies that the selection method evaluates if s more assets add information to CRIX. If so, these assets are added to the intercept and the next s assets are tested for. Assets with a higher market capitalization are expected to have a higher influence on the AIC, so the following theorem is formulated:

Theorem 1. *The rate of improvement of the AIC depends on the relative value of an asset in the market.*

The proof for Theorem 1 is given in the Appendix A.1, under the assumption of normally distributed error terms. Therefore, we will follow the common practise to include the assets with the highest market capitalization in the index,

$$\arg\max_{i} \sum_{j=1}^{k} P_{j,i,t_l^-} Q_{j,i,t_l^-}, \quad i \in \{1, \dots, K\}. \tag{23}$$

Thus, a top-down approach to decide about the number of index constituents is applied.

For the sorting of the index constituents by highest market capitalization, just the closing data of the last day of a month are used. We chose to do so, since the next periods CRIX will just depend on Q_{i,t_l^-}, (2), and not on data which lie further in the past. This is in line with the methodology of e.g. the DAX. For LCRIX, the CCs with the highest trading volume are chosen respectively,

$$\arg\max_{i} \sum_{j=1}^{k} Vol_{j,i,t_l^-}, \quad i \in \{1, \dots, K\}. \tag{24}$$

Since the differences between the TMI(k_{max}) and CRIX(k_j, β) are caused over time by the missing time series in CRIX(k_j, β), the independence assumption of the $\hat{\varepsilon}(k_j, \beta)$ for all j cannot be fulfilled by construction. But Györfi et al. (1989) give arguments that under certain conditions in case of nonparametric density estimation, the rate of convergence is essentially the same as for an independent sample. Summarizing the described procedure, results to:

1. At time point $T+1$, construct TMI(k_{max})
2. Set $j = 2$
3. Construct CRIX($k_1, 1$) and CRIX(k_j, β), $k_1 < k_2 < k_3 < \cdots$
4. Compute $\hat{\varepsilon}(k_j, \beta)$ and $\hat{\varepsilon}(k_1, 1)$
5. Kernel density estimation (KDE) for density $f(\hat{\varepsilon}(k_1, 1))$
   (a) Compute the log likelihood (20) for $\hat{\varepsilon}(k_j, \beta)$ with KDE for $\hat{\varepsilon}(k_1, 1)$.
   (b) Sum the log likelihoods
6. Derive AIC$\{\hat{\varepsilon}(k_j, \beta), k_j - k_1\}$ and AIC$\{\hat{\varepsilon}(k_1, 1), 0\}$
7. If $j = (k_{max} - k_1)/k_1$: stop, else jump to 3. and $j = j + 1$

The next section describes the further index rules for CRIX.

4. CRIX family rules

The constituents of the indices are regularly checked so that the corresponding index always represents its asset universe well. It is common to do this on a quarterly basis. In case of CRIX this reallocation is much faster. In the past, coins have shown a very volatile behavior, not just in the manner of price volatility. In some weeks, many occur out of nothing in the market and many others vanish from the market even when they were before very important, e.g., Auroracoin. This calls for a faster reallocation of the market benchmark than on a quarterly basis. A monthly reallocation is chosen to make sure that CRIX catches the momentum of the CC market well. Therefore, on the last day of every month, the CCs which had the highest market capitalization on the last day in the last month will be checked and the first k will be included in CRIX for the coming month. Accordingly for LCRIX the ones with the highest trading volume are chosen.

Since a review of an index is commonly performed on a quarterly basis the number of index members of CRIX will be checked on a quarterly basis too. The described procedure from Section 3 will be applied to the observations from the last three months on the last day of the third month after the markets closed. The number of index constituents, k, will be used for the next three months. Thus, CRIX corresponds to a monthly rebalanced portfolio which number of constituents is reviewed quarterly.

It may happen that some data are missing for some of the analyzed time series. If an isolated missing value occurs alone in the dataset, meaning that the values before and after it are not missing, then Missing At Random (MAR) is assumed. This assumption means that just observed information cause the missingness, Horton and Kleinman (2007). The Last-Observation-Carried-Forward (LOCF) method is then applied to fill the gap for the application of the AIC. We did not choose a different approach since a regression or imputation method may alter the data in the wrong direction. By LOCF, no change is implied and the CC is not excluded. If two or more data are missing in a row, then the MAR assumption may be violated, therefore no method is applied. The corresponding time series is then excluded from the computation in the derivation period. If data are missing during the computation of the index values, the LOCF method is applied too. This is done to make the index insensitive to this CC at this time point. CRIX should mimic market changes, therefore an imputation or regression method for the missing data would distort the view on the market.

Before continuing, the described rules are summarized:

- Quarterly altering of the number of index constituents

- Monthly altering of the index constituents
- Model selection for index derivation with AIC
- Nonparametric estimation of the density
- Application of a top-down approach to select the assets for the subset analysis
- Application of LOCF if trading of an asset stops before next reallocation.

5. The CRIX family

Using the described methods and rules from above, three indices will be proposed. This indices provide a different look at the market.

1. CRIX/LCRIX: The first and leading index is CRIX and for volume weighting LCRIX. While the choice for the best number of constituents is made, their numbers are chosen in steps of 5. It is common in financial industry to construct market indices with a number of constituents which is evenly divisible by 5, see e.g. FTSE (2016), S&P (2014), Deutsche Boerse AG (2013). Therefore this selection is applied for CRIX(k), $k = 5, 10, 15, \dots$ with $k_1 = 5$. Since the global minimum for the selection criterion may involve many index constituents, but a sparse index is the goal, the search for the optimal model terminates at level j whenever

$$\text{AIC}\{\hat{\varepsilon}(k_j, \beta), k_j - 5\} < \text{AIC}\{\hat{\varepsilon}(k_{j-1}, \beta), k_{j-1} - 5\} \tag{25}$$

and k_{j-1} index constituents are chosen. Therefore merely a local optimum will be achieved in most of the cases for $\Theta = \Theta_{AIC}$, in (13). But the choice is still asymptotically optimal by defining $\Theta = \{\Theta_{AIC} | k_i \leq k_j \forall i\}$. In Section 6 it will be shown that the performance of the index is already very good.
2. ECRIX/LECRIX: The second constructed index is called Exact CRIX (ECRIX) and Liquidity ECRIX respectively. It follows the above rules too. But the number of its constituents is chosen in steps of 1. Therefore the set of models contains CRIX(k), $k = 1, 2, 3, \dots$ with $k_1 = 1$ and stops when

$$\text{AIC}\{\hat{\varepsilon}(k_j, \beta), k_j - 1\} < \text{AIC}\{\hat{\varepsilon}(k_{j-1}, \beta), k_{j-1} - 1\}. \tag{26}$$

3. EFCRIX/LEFCRIX: Since the decision procedures for CRIX and ECRIX terminate when the AIC rises for the first time, Exact Full CRIX and Liquidity EFCRIX will be constructed to visualize whether the decision procedure works fine for the covered indices. The intention is to have an index which may approach the TMI but only in case even small assets help improve the view on the total market, a benchmark for the benchmarks. It will be derived with the AIC procedure, compare Section 3. For $k = 1, 2, 3, \dots$ with $k_1 = 1$ the decision rule is based on

$$\min_{k_j, \beta} \text{AIC}\{\hat{\varepsilon}(k_j, \beta), k_j - 1\} \tag{27}$$

for $\Theta = \Theta_{AIC}$, in (13). This index computes the AIC for every possible number of constituents and the number is chosen where the AIC becomes minimal.

6. Performance analysis

The indices CRIX, ECRIX, EFCRIX with market cap weighting and LCRIX, LECRIX, LEFCRIX with volume weighting have been proposed to give insight into the CC market. Our RDC CC database covers data for over 1000 CCs, kindly provided by CoinGecko. The data used for the analysis cover daily closing data for prices, market volume and market capitalization in USD for each CC in the time period from 2014-04-01 to 2017-03-25. Crypto exchanges are open on the weekends, therefore data for weekend closing prices exist. Since CC exchanges do not finish trading after a certain time point every day, a time point which serves as a closing time has to be defined. CoinGecko used 12 am UTC time zone. One should note that missing data are observed in the dataset, therefore the last rules from Section 4 will come into play.

Fig. 2 shows the performance of CRIX, and Fig. 6 the differences between CRIX and both ECRIX and EFCRIX. For the purpose of comparison, the indices were recalibrated on the recalculation dates since the index constituents change then. We do not provide each index plot individually since they perform almost equally. However, the AIC method gave very different numbers of constituents for the corresponding indices. The numbers of constituents are given in Table 4. For comparison, the number of constituents under the other discussed model selection criteria are provided too. The variance of C_p was derived with a GARCH(1,1) model, Bollerslev (1986). The corresponding information for ECRIX and EFCRIX are given in the same Table 4. Interestingly the methodology of EFCRIX causes its number of constituents to reach a relatively stable value for each period. ECRIX has mostly much fewer constituents than CRIX and EFCRIX due to the fact that this index just runs until a local optimum. Comparing the number of constituents for CRIX derived with AIC against the other criteria, one sees that GC, GFC and SH tend to choose more or the same number of constituents than AIC. Also all three criteria suggest the same result. C_p stops at the initial value for CRIX, ECRIX and EFCRIX. For CRIX, ECRIX and EFCRIX, AIC mostly chooses less constituents compared to all other criteria, except C_p which terminates very early. For LCRIX, LECRIX and LEFCRIX mostly less constituents were chosen than for CRIX, ECRIX and EFCRIX, compare Table 5. Note that the AIC gave the sparsest result again.

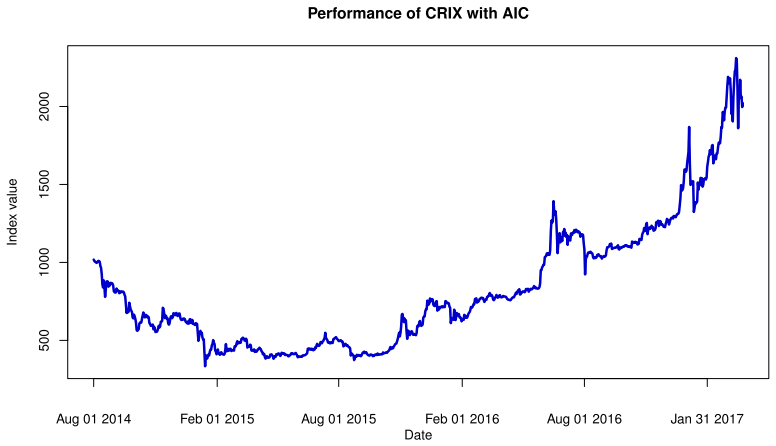


Fig. 2. Performance of CRIX. CRIXindex CRIXcode.

Table 2
Comparison of CRIX, ECRIX, EFCRIX, derived under different penalizations, against TMI under mean of monthly Mean Squared Error, compared with btc.

	AIC	GC	GFC	SH	Cp	FPE
CRIX	0.4769	0.4883	0.3755	0.3598	1.9844	0.0042
ECRIX	11.0988	10.3673	10.3673	10.4667	79.3979	0.0048
EFCRIX	3.1394	0.0116	0.0049	0.0049	79.3979	0.0048
LCRIX	0.6417	0.1497	0.1217	0.1211	0.6638	0.0049
LECRIX	22.8782	16.7187	16.7187	16.7187	125.0620	0.0047
LEFCRIX	7.9158	0.0645	0.0126	0.0126	125.0620	0.0047
btc	79.3979	79.3979	79.3979	79.3979	79.3979	79.3979

Table 3
Comparison of CRIX, ECRIX, EFCRIX, derived under different penalizations, against TMI under mean of monthly Mean Directional Accuracy, compared with btc.

	AIC	GC	GFC	SH	Cp	FPE
CRIX	0.9896	0.9908	0.9918	0.9928	0.9835	1.0000
ECRIX	0.9576	0.9586	0.9586	0.9586	0.9133	1.0000
EFCRIX	0.9794	0.9990	1.0000	1.0000	0.9133	1.0000
LCRIX	0.9928	0.9949	0.9959	0.9959	0.9917	1.0000
LECRIX	0.9692	0.9700	0.9700	0.9700	0.9501	1.0000
LEFCRIX	0.9855	0.9979	1.0000	1.0000	0.9501	1.0000
btc	0.9133	0.9133	0.9133	0.9133	0.9133	0.9133

The indices optimized until a local optimum are expected to perform less optimal than the globally optimized ones against the TMI/LTMI. Tables 2 and 3 give the mean over monthly Mean Squared Error (MSE) and Mean Directional Accuracy (MDA), defined as

$$\mathrm{MSE}\{\mathrm{CRIX}(k)\} = \frac{1}{t_l^+ - t_l^-} \sum_{t=t_l^-}^{t_l^+} \{\mathrm{CRIX}(k)_t - \mathrm{TMI}(k_{\max})_t\}^2 \tag{28}$$

$$\begin{aligned}\mathrm{MDA}\{\mathrm{CRIX}(k)\} = \frac{1}{t_l^+ - t_l^-} \sum_{t=t_l^-}^{t_l^+} \mathbf{I}[&\mathrm{sign}\{\mathrm{TMI}(k_{\max})_t - \mathrm{TMI}(k_{\max})_{t-1}\} \\ &= \mathrm{sign}\{\mathrm{CRIX}(k)_t - \mathrm{CRIX}(k)_{t-1}\}]\end{aligned} \tag{29}$$

where t_l^- and t_l^+ are the beginning and end of the month respectively, $\mathbf{I}(\cdot)$ is the indicator function and $\mathrm{sign}(\cdot)$ gives the sign of the respective equation. Apparently CRIX performs best, which can be explained due to its larger number of index constituents. The CRIX, ECRIX and EFCRIX are close in terms of the MDA but the MSE is much better for CRIX. Comparing all the model selection criteria, FPE has the best performance in terms of MSE and MDA, due to choosing high numbers of constituents. The trading volume weighted indices are close in terms of MSE and MDA to their market weighted corresponding indices. At the same time the number of constituents are mostly sparser for the volume weighted ones.

CRIX was constructed with steps of five which is common in practice and performed best under AIC. For this case the number of constituents was the most stable, while achieving the best performance for MSE and MDA. Additionally, the analysis showed that it is indeed unnecessary from a practical viewpoint to choose the global optimal AIC under steps of 1. Even a local optimum and a much

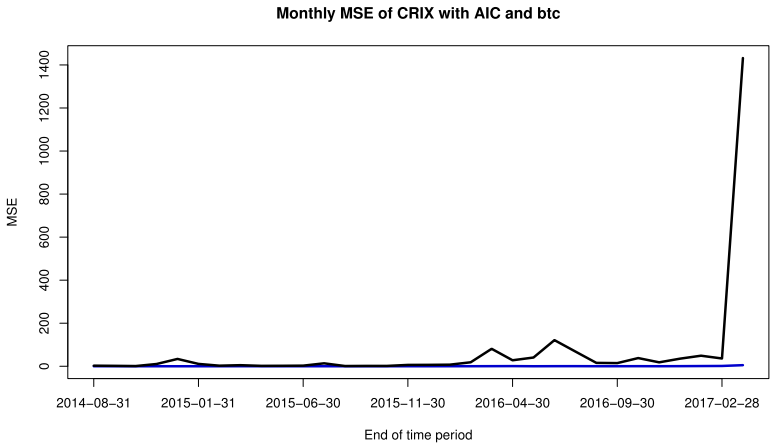


Fig. 3. Performance of CRIX compared to BTC.

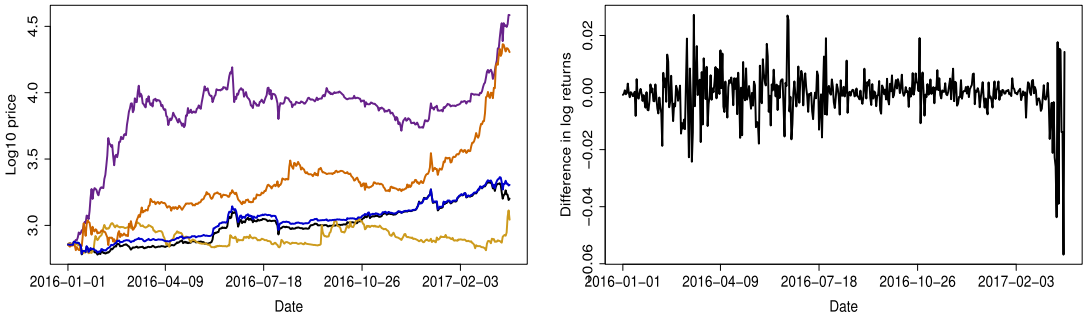
Table 4
Comparison of AIC, GC, GFC, SH, Cp and the FPE method for the selection of the number of index constituents for the CRIX, ECRIX and EFCRIX in the 11 periods.

	CRIX						ECRIX						EFCRIX						
	AIC	GC	GFC	SH	Cp	FPE	AIC	GC	GFC	SH	Cp	FPE	AIC	GC	GFC	SH	Cp	FPE	max
1	5	10	10	10	10	35	2	2	2	3	1	36	2	7	30	30	1	36	36
2	10	15	15	15	5	100	3	3	3	3	1	93	3	94	93	93	1	93	113
3	5	10	35	35	5	100	5	5	5	5	1	93	5	94	93	93	1	93	158
4	10	10	10	40	5	95	3	3	3	3	1	90	3	91	90	90	1	90	182
5	10	20	20	20	5	100	2	4	4	4	1	93	12	94	93	93	1	93	169
6	10	10	20	20	5	100	2	2	2	2	1	93	2	94	93	93	1	93	171
7	5	20	20	20	5	100	1	1	1	1	1	93	16	94	93	93	1	93	176
8	15	20	20	20	5	95	3	4	4	4	1	91	3	92	91	91	1	91	140
9	15	5	5	5	5	100	3	3	3	3	1	93	3	94	93	93	1	93	188
10	15	15	25	25	5	100	3	5	5	5	1	93	3	94	93	93	1	93	207
11	10	35	45	45	5	100	2	2	2	2	1	93	4	94	93	93	1	93	221

more stable number of constituents is able to mimic the market movements very well in terms of the MDA and MSE. Furthermore, even for ECRIX there was more than one constituent selected most of the time. This shows that Bitcoin, which currently clearly dominates the market in terms of market capitalization and trading volume, does not account for all the variance in the market. Other CCs are important for the market movements too.

Depending on the theoretical and empirical analysis, we decided to continue with the AIC. From the theoretical viewpoint, the AIC uses the most information about the data, since it relies on the density. From the empirical analysis, the AIC chooses much less constituents than GC, GFC, SH and FPE, while its performance in terms of MSE and MDA is close to the three outlined criteria. The better performance was achieved due to overparametrization of the index by GC, GFC, SH and FPE. Therefore, CRIX will be derived with the AIC criterion.

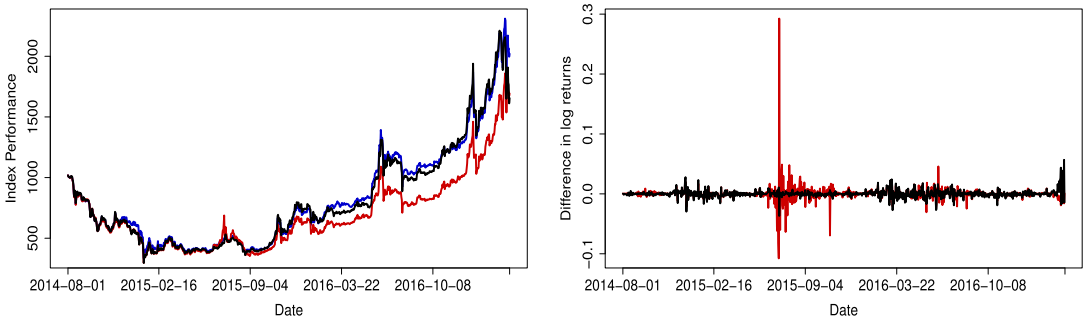
Comparing CRIX with the development of BTC, it tracks the market development better over time. Fig. 3 shows the monthly MSE of CRIX with AIC and BTC. In 2016 CRIX tracked the market development much better than BTC, and in the beginning of 2017 even better due to the huge impact of the price gain of altcoins like Ethereum, Ripple and Dash. Their performance is visualized in Fig. 4, clearly showing the better performance of CRIX in this time period, driven by price gains in altcoins. Due to the log scale and the high gains of altcoins, the difference between CRIX and BTC appears little, while in fact being considerable. Fig. 4(b) shows the difference in the log returns of CRIX and BTC. One sees differences in their return series, which are particularly strong beginning of 2016 and in March 2017. Comparing the performance of CRIX and LCRIX against BTC, one observes an increasing spread between the indices, Fig. 5. It indicates a lower weight of BTC in LCRIX, thus tackling the issue of dominance of BTC in CRIX by liquidity weighting. Having a look at the actual differences in the log return series compared to CRIX, Fig. 5(b), stronger spikes are observed, thus showing the difference in the performance from CRIX and LCRIX driven by the stronger weights on altcoins in LCRIX. Table 8 shows the actual weights given to BTC and altcoins in the respective indices. In the liquidity indices altcoins frequently receive a higher weight compared to the respective indices based on market capitalization weighting. Once the altcoins received even 52% of the weights in LCRIX. The results show the market focus in terms of trading is stronger for altcoins than their market capitalization suggests, thus an index accounting for this is called for, LCRIX. Simultaneously the weighting scheme tackles the dominance of BTC in a market capitalization index.



(a) Performance of rescaled log price series of CRIX, BTC, ETH, XRP and DASH

(b) Difference in the log returns of CRIX and BTC

Fig. 4. Comparison of performance of CRIX, BTC, ETH, XRP and DASH.



(a) Performance of index series of CRIX, BTC and LCRIX

(b) Difference in the log returns of CRIX and LCRIX and CRIX and BTC

Fig. 5. Comparison of performance of CRIX, BTC and LCRIX.

Table 5
Comparison of AIC, GC, GFC, SH, Cp and the FPE method for the selection of the number of index constituents for the LCRIX, LECRIX and LEFCRIX in the 11 periods.

	LCRIX						LECRIX						LEFCRIX						
	AIC	GC	GFC	SH	Cp	FPE	AIC	GC	GFC	SH	Cp	FPE	AIC	GC	GFC	SH	Cp	FPE	max
1	5	10	15	15	5	35	2	3	3	3	1	36	2	6	16	16	1	36	36
2	5	10	10	10	5	100	2	4	4	4	1	93	2	94	93	93	1	93	113
3	5	5	20	20	5	100	3	4	4	4	1	93	6	94	93	93	1	93	158
4	15	20	20	30	5	95	3	2	2	2	1	90	3	91	90	90	1	90	182
5	5	5	5	5	5	100	1	1	1	1	1	93	93	94	93	93	1	93	169
6	5	5	5	5	5	100	2	2	2	2	1	93	9	94	93	93	1	93	171
7	10	25	30	35	5	100	1	2	2	2	1	93	1	94	93	93	1	93	176
8	10	20	35	35	5	95	1	1	1	1	1	91	3	92	91	91	1	91	140
9	5	10	10	10	5	100	2	2	2	2	1	93	2	94	93	93	1	93	188
10	10	10	10	10	5	100	1	1	1	1	1	93	1	94	93	93	1	93	207
11	5	15	15	15	5	100	2	3	3	3	1	93	2	94	93	93	1	93	221

7. Application to the German stock market

The CRIX methodology was derived with the idea of finding a method which allows mimicking young and fast changing markets appropriately. But well known major markets usually change their structure too. So the proposed methodology is tested on the German stock market, which has four major indices: DAX, MDAX, SDAX and TecDAX. The DAX is used to determine the overall market direction, Janßen and Rudolph (1992). Since it is chosen from the so called prime segment, it has some prior restrictions. It is interesting to see whether our methodology yields the DAX as an adequate benchmark for the total market. Since the indices are derived with market cap weighting scheme, only this methodology is tested. Following Definition 4, all available stocks are defined as the TMI and our new method is applied to find an appropriate index. Again, the 7-step method from Section 3 was applied to find the number of constituents, but it starts at 30 members to check if more constituents are necessary. The method for the identification of k and the reallocation of the included assets is performed quarterly, like DAX. To be in line with the DAX reallocation dates, the

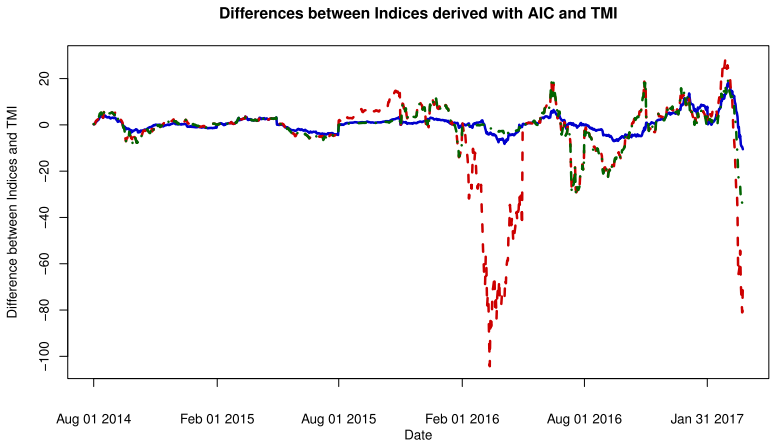


Fig. 6. Realized difference between TMI and CRIX (solid), ECRIX (dashed), EFCRIX (dotdashed). CRIXfamdiff CRIXcode.

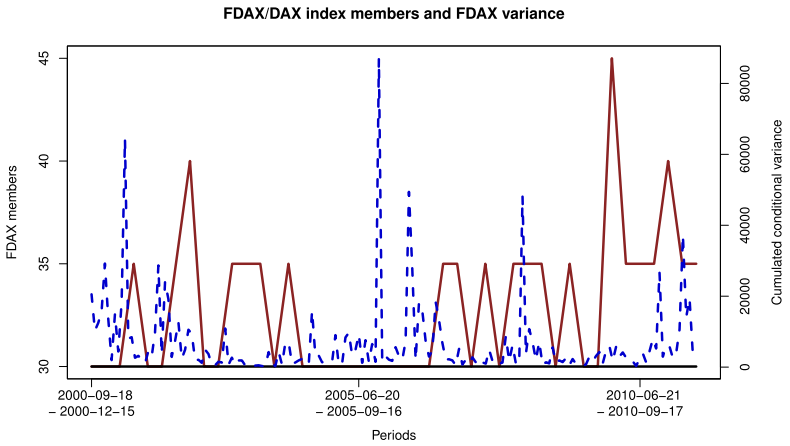


Fig. 7. Number of constituents of FDAX (solid), DAX (horizontal solid) and cumulated monthly variance of FDAX (dashed). CRIXdaxmembersvar CRIXcode.

index calculation will start after the third Friday of September and the reallocation dates are the third Fridays of December, March, June and September, see Deutsche Boerse AG (2013).

The data were fetched from Datastream in the period 2000-06-16 until 2015-12-18. All stocks which are German companies and are traded on XETRA are chosen. Any time series for which Datastream reported an error either for the price or market capitalization data was excluded from the analysis. The index, computed with the new methodology, is called Flexible DAX (FDAX). One should note that the analysis starts three months after the starting point of the dataset due to the initialization period of FDAX.

Fig. 7 shows the number of members of FDAX and DAX in the respective periods. Most of the time, the number of index constituents for FDAX is higher than the 30 members of DAX. Just around 2004–2005 is the k more frequently 30. Especially while the turmoil of the financial markets, starting from 2008/2009, is the number of index constituents much higher. One might hint that a higher reported variability in one period should cause an increase in k in the next period, since it was shown that the selection method depends on the variance, see Appendix A. Fig. 7 shows that this idea can partially be supported. The derivation of the conditional variance was performed with a GARCH(1,1) model, Bollerslev (1986), and the daily results were summed up. Obviously, in the extreme cases increases the k in the next period, see 2001, 2002, 2006 and 2011.

The computation of the MSE and MDA, see Table 6, shows that FDAX is a more accurate benchmark for the total market as DAX. Since Janßen and Rudolph (1992) state that DAX may be used to analyze the movements of the total market, an MDA of 92% is indeed good. But FDAX mimics the market even better, with an MDA of 96%. Also the MSE for FDAX is much lower than the one of DAX. Therefore the methodology fulfilled its goal to find a sparse, investable and accurate benchmark, depending on the MDA.

8. Application to Mexican stock market

The Mexican stock market is represented by the IPC35, MEXBOL (2013). One of its rules is a readjustment of the weights to lower the effect of dominant stocks. In the CC market BTC is such a dominant asset. The CRIX methodology could help to circumvent arbitrary rules and develop an index to represent the market accurately.

Table 6
Comparison of DAX with CRIX methodology (FDAX) and rescaled DAX against TMI.

	MSE	MDA
FDAX vs. TMI	6.36	0.96
DAX vs. TMI	51.02	0.92

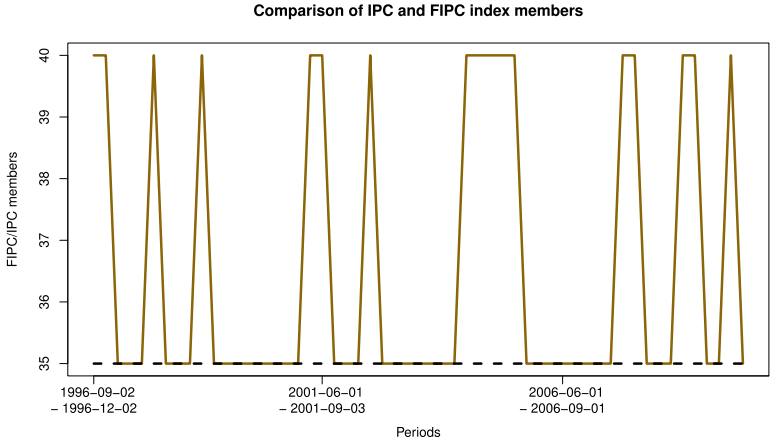


Fig. 8. Number of constituents of FIPC (solid) and IPC (dashed) in the respective periods. CRIXipcmembers CRIXcode.

Table 7
Comparison of IPC with CRIX methodology (FIPC) and rescaled IPC against TMI.

	MSE	MDA
FIPC vs. TMI	24.97	0.97
IPC vs. TMI	4743.50	0.91

The data were fetched from Datastream for the period 1996-06-01 until 2015-05-29 and cover all Mexican companies listed in Datastream. The specifications of the methodology are the same as for the German stock market except for the recalculation date. In line with the methodology of the IPC35, the index is recalculated with the closing data of the last business days of August, November, February and May, therefore the recalculated index starts on the first business days of September, December, March and June. The TMI will be all fetched companies. The choice of k starts with 35 since this is the amount of constituents of IPC.

Again, the CRIX methodology works well. The MSE is very low compared to the one for the IPC35 and the MDA gives a much better performance too, see Table 7. We can conclude that the methodology helped to circumvent the usage of arbitrary rules for the weights in the rules of the indices and enhances at the same time the performance of the market index. Fig. 8 shows the number of index members of the FIPC compared to the IPC. Obviously, the methodology also suggests using more than 35 index members half of the time which is the number of members of the IPC.

9. Conclusion

The movements of CCs are very different from each other, Elendner et al. (2017). So studying the entire market of CCs requires an instrument which adequately captures and displays the market movements, an index. But index construction for CCs requires a new methodology to find the right number of index members. Innovative markets, like the one for CCs, change their structure frequently. The proposed methods were applied to oracle a new family of indices, which are displayed and updated on a daily basis. The performance of the new indices were studied and it was shown that the dynamic AIC based methodology results in indices with stable properties. The results show that a market like the CC market – momentarily dominated by Bitcoin – still needs a representative index since Bitcoin does not account for all the variance in the market. The diversified nature of the CC market makes the inclusion of altcoins in the index product critical to improve tracking performance. We have shown that assigning optimal weights to altcoins helps to reduce the tracking errors of a CC portfolio, despite the fact that their market cap is much smaller relative to Bitcoin.

Besides the classical market capitalization weighting, a volume weighting scheme was proposed. The corresponding indices are sparser in terms of constituents while having a comparable performance, which gives support to this weighting scheme under the goals of the study. The AIC based method was also applied to the German stock market. The results yield a more accurate benchmark in terms of MDA. In applying the CRIX methodology to the Mexican stock market, which is dominated by Telmex, one finds high accuracy of it in terms of MSE and MDA.

We conclude, that the CRIX technology enhances the construction of an index if the goal is to find a sparse, investable and accurate benchmark.

Table 8
Average monthly weights of BTC and altcoins in the respective periods in the 6 indices.

Periods	CRIX			LCRIX			ECRIX			LECRIX			EFCRIX			LEFCRIX		
	k	BTC	altcoins	k	BTC	altcoins	k	BTC	altcoins	k	BTC	altcoins	k	BTC	altcoins	k	BTC	altcoins
2014/08	5	0.96	0.04	5	0.96	0.04	2	0.98	0.02	2	0.97	0.03	2	0.98	0.02	2	0.97	0.03
2014/09	5	0.94	0.06	5	0.82	0.18	2	0.97	0.03	2	0.86	0.14	2	0.97	0.03	2	0.86	0.14
2014/10	5	0.93	0.07	5	0.86	0.14	2	0.97	0.03	2	0.91	0.09	2	0.97	0.03	2	0.91	0.09
2014/11	10	0.92	0.08	5	0.95	0.05	3	0.94	0.06	2	0.99	0.01	3	0.94	0.06	2	0.99	0.01
2014/12	10	0.85	0.15	5	0.95	0.05	3	0.86	0.14	2	0.97	0.03	3	0.86	0.14	2	0.97	0.03
2015/01	10	0.82	0.18	5	0.93	0.07	3	0.85	0.15	2	0.98	0.02	3	0.85	0.15	2	0.98	0.02
2015/02	5	0.86	0.14	5	0.90	0.10	5	0.86	0.14	3	0.91	0.09	5	0.86	0.14	6	0.90	0.10
2015/03	5	0.90	0.10	5	0.96	0.04	5	0.90	0.10	3	0.96	0.04	5	0.90	0.10	6	0.96	0.04
2015/04	5	0.90	0.10	5	0.94	0.06	5	0.90	0.10	3	0.95	0.05	5	0.90	0.10	6	0.94	0.06
2015/05	10	0.90	0.10	15	0.96	0.04	3	0.92	0.08	3	0.96	0.04	3	0.92	0.08	3	0.96	0.04
2015/06	10	0.87	0.13	15	0.86	0.14	3	0.90	0.10	3	0.88	0.12	3	0.90	0.10	3	0.88	0.12
2015/07	10	0.88	0.12	15	0.48	0.52	3	0.90	0.10	3	0.48	0.52	3	0.90	0.10	3	0.48	0.52
2015/08	10	0.88	0.12	5	0.59	0.41	2	0.93	0.07	1	1.00	0.00	12	0.88	0.12	93	0.58	0.42
2015/09	10	0.89	0.11	5	0.53	0.47	2	0.93	0.07	1	1.00	0.00	12	0.88	0.12	93	0.52	0.48
2015/10	10	0.90	0.10	5	0.59	0.41	2	0.96	0.04	1	1.00	0.00	12	0.90	0.10	93	0.58	0.42
2015/11	10	0.92	0.08	5	0.82	0.18	2	0.97	0.03	2	0.83	0.17	2	0.97	0.03	9	0.82	0.18
2015/12	10	0.93	0.07	5	0.84	0.16	2	0.98	0.02	2	0.84	0.16	2	0.98	0.02	9	0.84	0.16
2016/01	10	0.92	0.08	5	0.87	0.13	2	0.97	0.03	2	0.87	0.13	2	0.97	0.03	9	0.87	0.13
2016/02	5	0.89	0.11	10	0.90	0.10	1	1.00	0.00	1	1.00	0.00	16	0.87	0.13	1	1.00	0.00
2016/03	5	0.83	0.17	10	0.86	0.14	1	1.00	0.00	1	1.00	0.00	16	0.81	0.19	1	1.00	0.00
2016/04	5	0.85	0.15	10	0.93	0.07	1	1.00	0.00	1	1.00	0.00	16	0.84	0.16	1	1.00	0.00
2016/05	15	0.83	0.17	10	0.75	0.25	3	0.87	0.13	1	1.00	0.00	3	0.87	0.13	3	0.75	0.25
2016/06	15	0.85	0.15	10	0.65	0.35	3	0.88	0.12	1	1.00	0.00	3	0.88	0.12	3	0.65	0.35
2016/07	15	0.84	0.16	10	0.71	0.29	3	0.90	0.10	1	1.00	0.00	3	0.90	0.10	3	0.71	0.29
2016/08	15	0.83	0.17	5	0.70	0.30	3	0.90	0.10	2	0.72	0.28	3	0.90	0.10	2	0.72	0.28
2016/09	15	0.82	0.18	5	0.73	0.27	3	0.88	0.12	2	0.77	0.23	3	0.88	0.12	2	0.77	0.23
2016/10	15	0.84	0.16	5	0.84	0.16	3	0.89	0.11	2	0.85	0.15	3	0.89	0.11	2	0.85	0.15
2016/11	15	0.87	0.13	10	0.94	0.06	3	0.91	0.09	1	1.00	0.00	3	0.91	0.09	1	1.00	0.00
2016/12	15	0.89	0.11	10	0.94	0.06	3	0.93	0.07	1	1.00	0.00	3	0.93	0.07	1	1.00	0.00
2017/01	15	0.88	0.12	10	0.92	0.08	3	0.93	0.07	1	1.00	0.00	3	0.93	0.07	1	1.00	0.00
2017/02	10	0.89	0.11	5	0.93	0.07	2	0.94	0.06	2	0.94	0.06	4	0.92	0.08	2	0.94	0.06
2017/03	10	0.81	0.19	5	0.74	0.26	2	0.87	0.13	2	0.82	0.18	4	0.84	0.16	2	0.82	0.18

Acknowledgments

We would like to thank the editor and an anonymous referee for their valuable comments to this article. Our thanks extends to David Lee Kuo Chuen and Ernie G. S. Teo for their comments in several discussions. Financial support from the Deutsche Forschungsgemeinschaft, Germany via CRC 649 "Economic Risk" and IRTG 1792 "High Dimensional Non Stationary Time Series", Humboldt-Universität zu Berlin, Germany is gratefully acknowledged.

Appendix A

A.1. Proof of Theorem 1

Proof. Assume normally distributed error terms, (10) and (22): $\epsilon(k,\beta) \sim \mathrm{N}\{0, \sigma(k,\beta)^2\}$, $\hat{\epsilon}(k,\beta) \sim \mathrm{N}\{0, \hat{\sigma}(k,\beta)^2\}$. Then

$$\log L\{\epsilon(k,\beta)\} = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log\sigma(k,\beta)^2 - \frac{1}{2\sigma(k,\beta)^2}\sum_{t=1}^{T}\epsilon(k,\beta)_t^2. \tag{30}$$

Denote $RSS\{\hat{\epsilon}(k,\beta)\} = \sum_{t=1}^{T}\hat{\epsilon}(k,\beta)_t^2$ and $\hat{\sigma}(k,\beta)^2 = T^{-1}RSS\{\hat{\epsilon}(k,\beta)\}$. Then

$$\log L\{\hat{\epsilon}(k,\beta)\} = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log T^{-1}RSS\{\hat{\epsilon}(k,\beta)\} - \frac{1}{2T^{-1}RSS\{\hat{\epsilon}(k,\beta)\}}RSS\{\hat{\epsilon}(k,\beta)\} \tag{31}$$

$$= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log T^{-1}RSS\{\hat{\epsilon}(k,\beta)\} - \frac{T}{2} \tag{32}$$

$$= -\frac{T}{2}\log T^{-1}RSS\{\hat{\epsilon}(k,\beta)\} + C \tag{33}$$

with $C = -\frac{T}{2}\log(2\pi) - \frac{T}{2}$. Since C does not depend on any model parameters, just on the data length T, this part of the equation could be omitted.

$$\mathrm{AIC}\{\hat{\epsilon}(k,\beta), s\} = T\log T^{-1}RSS\{\hat{\epsilon}(k,\beta)\} + 2\cdot s \tag{34}$$

$$= T\log\hat{\sigma}(k,\beta)^2 + 2\cdot s \tag{35}$$

The enhancement in the fit to the Total Market Index (TMI) by adding more constituents, s, determines the degree of improvement of the likelihood.

With the linearity property of the expectation operator, assume without loss of generality

$$\begin{aligned}
\mathsf{E}\{\varepsilon(k_{max})^{TM}\} = \mathsf{E}\{\varepsilon(k,\beta)^{CRIX}\} &= 0\\
t &\in \{1,\dots,T\}\\
t_l^- &= 0\\
s &= 1
\end{aligned}$$

$$\begin{aligned}
\widehat{\sigma}(k,\beta) &= \mathrm{Var}\{\widehat{\varepsilon}(k,\beta)\}\\
&= \mathrm{Var}\{\varepsilon(k_{max})^{TM} - \varepsilon(k,\beta)^{CRIX}\}\\
&= \sum_{t=1}^{T}\Bigg[\log\left\{\sum_{i=1}^{k_{max}} P_{it}Q_{i,0}\Big(\sum_{i=1}^{k} P_{i,t-1}Q_{i,0} + \beta_1 P_{k+1,t-1}Q_{k+1,0}\Big)\right\}\\
&\qquad - \log\left\{\sum_{i=1}^{k_{max}} P_{i,t-1}Q_{i,0}\Big(\sum_{i=1}^{k} P_{i,t}Q_{i,0} + \beta_1 P_{k+1,t}Q_{k+1,0}\Big)\right\}\Bigg]^2\\
&= \sum_{t=1}^{T}\Bigg[\log\left\{\sum_{i=1}^{k_{max}} P_{it}Q_{i,0}\sum_{i=1}^{k} P_{i,t-1}Q_{i,0} + \sum_{i=1}^{k_{max}} P_{it}Q_{i,0}\beta_1 P_{k+1,t-1}Q_{k+1,0}\right\}\\
&\qquad - \log\left\{\sum_{i=1}^{k_{max}} P_{i,t-1}Q_{i,0}\sum_{i=1}^{k} P_{i,t}Q_{i,0} + \sum_{i=1}^{k_{max}} P_{i,t-1}Q_{i,0}\beta_1 P_{k+1,t}Q_{k+1,0}\right\}\Bigg]^2
\end{aligned}$$

Using the relation $\log(a+b) = \log(a) + \log(1+\frac{b}{a})$, it results:

$$\begin{aligned}
&= \sum_{t=1}^{T}\Bigg[\log\left\{\sum_{i=1}^{k_{max}} P_{it}Q_{i,0}\sum_{i=1}^{k} P_{i,t-1}Q_{i,0}\right\} + \log\left\{1 + \frac{\sum_{i=1}^{k_{max}} P_{it}Q_{i,0}\beta_1 P_{k+1,t-1}Q_{k+1,0}}{\sum_{i=1}^{k_{max}} P_{it}Q_{i,0}\sum_{i=1}^{k} P_{i,t-1}Q_{i,0}}\right\}\\
&\qquad - \log\left\{\sum_{i=1}^{k_{max}} P_{i,t-1}Q_{i,0}\sum_{i=1}^{k} P_{i,t}Q_{i,0}\right\} + \log\left\{1 + \frac{\sum_{i=1}^{k_{max}} P_{i,t-1}Q_{i,0}\beta_1 P_{k+1,t}Q_{k+1,0}}{\sum_{i=1}^{k_{max}} P_{i,t-1}Q_{i,0}\sum_{i=1}^{k} P_{i,t}Q_{i,0}}\right\}\Bigg]^2\\
&= \sum_{t=1}^{T}\Bigg(\log\left\{\sum_{i=1}^{k_{max}} P_{it}Q_{i,0}\sum_{i=1}^{k} P_{i,t-1}Q_{i,0}\right\} - \log\left\{\sum_{i=1}^{k_{max}} P_{i,t-1}Q_{i,0}\sum_{i=1}^{k} P_{i,t}Q_{i,0}\right\}\\
&\qquad + \left[\log\left\{1 + \frac{\beta_1 P_{k+1,t-1}Q_{k+1,0}}{\sum_{i=1}^{k} P_{i,t-1}Q_{i,0}}\right\} - \log\left\{1 + \frac{\beta_1 P_{k+1,t}Q_{k+1,0}}{\sum_{i=1}^{k} P_{i,t}Q_{i,0}}\right\}\right]\Bigg)^2
\end{aligned} \tag{36}$$

Solving the derivation and writing the terms which do not depend on β_1 as A_t and the last part of (36) as B_t:

$$\begin{aligned}
\widehat{\sigma}(k,\beta) &= \sum_{t=1}^{T} A_t + 2\log\left\{\sum_{i=1}^{k_{max}} P_{it}Q_{i,0}\sum_{i=1}^{k} P_{i,t-1}Q_{i,0}\right\} B_t - 2\log\left\{\sum_{i=1}^{k_{max}} P_{i,t-1}Q_{i,0}\sum_{i=1}^{k} P_{i,t}Q_{i,0}\right\} B_t + B_t^2\\
&= \sum_{t=1}^{T} A_t + 2B_t\left[\log\left\{\sum_{i=1}^{k_{max}} P_{it}Q_{i,0}\sum_{i=1}^{k} P_{i,t-1}Q_{i,0}\right\} - \log\left\{\sum_{i=1}^{k_{max}} P_{i,t-1}Q_{i,0}\sum_{i=1}^{k} P_{i,t}Q_{i,0}\right\}\right] + B_t^2\\
&= \sum_{t=1}^{T} A_t + 2B_t\left[\varepsilon(k_{max})^{TM} - \varepsilon(k,1)^{CRIX}\right] + B_t^2
\end{aligned}$$

Since normally distributed error terms are assumed, note that $\beta_1 = \frac{Cov\{\widehat{\varepsilon}(k,1),\varepsilon_{k+1}\}}{Var\{\varepsilon_{k+1}\}}$, where ε_{k+1} is the log return of $P_{i,t}Q_{i,0}$. The change in the variance will depend on the additional variance which the new constituent can explain, see β_1. Furthermore, it depends on the value of $P_{k+1,t}Q_{k+1,0}$ relative to $\sum_{i=1}^{k} P_{i,t}Q_{i,0}$, (36), which is the summed market value of the constituents in the index. This infers that constituents with a higher market capitalization are more likely to be part of the index. ■

This gives support to using the often applied top-down approach, which we use for the construction of CRIX too.

References

Akaike, H., 1970. Statistical predictor identification. Ann. Inst. Statist. Math. 22 (1), 203–217.

Akaike, H., 1998. Information theory and an extension of the maximum likelihood principle. In: Parzen, E., Tanabe, K., Kitagawa, G. (Eds.), Selected Papers of Hirotugu Akaike. In: Springer Series in Statistics, Springer New York, pp. 199–213.

Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. Stat. Surv. 4, 40–79.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. J. Econometrics 31 (3), 307–327.
Bolt, W., Oordt, M.v., 2016. On the Value of Virtual Currencies. SSRN Scholarly Paper ID 2842557. Social Science Research Network, Rochester, NY.
Chen, S., Chen, C.Y.-H., Härdle, W.K., Lee, T.M., Ong, B., 2017. Econometric analysis of a cryptocurrency index for portfolio investment. In: Lee Kuo Chuen, D., Deng, R. (Eds.), Handbook of Digital Finance and Financial Inclusion: Cryptocurrency, FinTech, InsurTech, and Regulation, vol. 1. Elsevier.
Craven, P., Wahba, G., 1978. Smoothing noisy data with spline functions. Numer. Math. 31 (4), 377–403.
CRSP, 2015. CRSP U.S. Equity Indexes Methodology Guide, crsp.com/.
Deutsche Boerse AG, 2013. Guide to the equity Indizes of Deutsche Boerse, www.dax-indices.com.
Devroye, L., Györfi, L., 1985. Nonparametric Density Estimation The L1 View. Wiley.
Droge, B., 1996. Some comments on cross-validation. In: Härdle, W.K., Schimek, M.G. (Eds.), Statistical Theory and Computational Aspects of Smoothing. In: Contributions to Statistics, Physica-Verlag HD, pp. 178–199.
EconoTimes, 2016. Japans cabinet Approves New Bitcoin Regulations, econotimes.com.
Elendner, H., Trimborn, S., Ong, B., Lee, T.M., 2017. The cross-section of crypto-currencies as financial assets. In: Lee Kuo Chuen, D., Deng, R. (Eds.), Handbook of Digital Finance and Financial Inclusion: Cryptocurrency, FinTech, InsurTech, and Regulation, vol. 1. Elsevier.
Epanechnikov, V., 1969. Non-parametric Estimation of a Multivariate Probability Density. Theory Probab. Appl. 14 (1), 153–158.
FTSE, 2016. FTSE UK Index Series, www.ftse.com.
Györfi, L., Härdle, W.K., Sarda, P., Vieu, P., 1989. Nonparametric curve Estimation from Time Series. In: Györfi, L., Härdle, W.K., Sarda, P., Vieu, P. (Eds.), Lecture Notes in Statistics, vol. 60. Springer New York.
Hall, P., 1987. On Kullback-Leibler loss and density estimation. Ann. Statist. 15 (4), 1491–1519.
Härdle, W.K., Müller, M., Sperlich, S., Werwatz, A., 2004. Nonparametric and Semiparametric Models. Springer Science & Business Media.
Härdle, W.K., Trimborn, S., 2015. CRIX or evaluating Blockchain based currencies, Oberwolfach Report No. 42/2015 "The Mathematics and Statistics of Quantitative Risk.
Hayek, F.A., 1990. Denationalization of Money: An Analysis of the Theory and Practice of Concurrent Currencies, third ed. Institute of Economic Affairs, London.
Horton, N.J., Kleinman, K.P., 2007. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. Amer. Statist. 61 (1), 79–90.
Hurvich, C.M., Tsai, C.-L., 1989. Regression and time series model selection in small samples. Biometrika 76 (2), 297–307.
Janßen, B., Rudolph, B., 1992. Der deutsche Aktienindex DAX. Fritz Knapp Verlag.
Kanazawa, Y., 1993. Hellinger distance and Kullback–Leibler loss for the kernel density estimator. Statist. Probab. Lett. 18 (4), 315–321.
Kawa, L., 2015. Bitcoin Is Officially a Commodity, According to U.S. Regulator, Bloomberg.com.
Kristoufek, L., 2015. What are the main drivers of the Bitcoin price? Evidence from wavelet coherence analysis. PLoS One 10, 1–15.
Mallick, H., Yi, N., 2013. Bayesian methods for High Dimensional Linear Models. J. Biometr. Biostat. 1.
Mallows, C.L., 1973. Some comments on Cp. Technometrics 15 (4), 661–675.
MEXBOL, 2013. Prices and quotations Index (MEXBOL) - Methodology Note, bmv.com.
Nishii, R., 1984. Asymptotic properties of criteria for selection of variables in multiple regression. Ann. Statist. 12 (2), 758–765.
Reid, F., Harrigan, M., 2013. An analysis of anonymity in the bitcoin system. In: Altshuler, Y., Elovici, Y., Cremers, A.B., Aharony, N., Pentland, A. (Eds.), Security and Privacy in Social Networks. Springer New York, pp. 197–223.
Ron, D., Shamir, A., 2013. Quantitative analysis of the full bitcoin transaction graph. In: Sadeghi, A.-R. (Ed.), Financial Cryptography and Data Security. In: Lecture Notes in Computer Science, vol. 7859, Springer Berlin Heidelberg, pp. 6–24.
Sheather, S.J., Jones, M.C., 1991. A reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation. Journal of the Royal Statistical Society. Series B. Methodological 53, 683–690.
Shibata, R., 1981. An optimal Selection of Regression Variables. Biometrika 68 (1), 45–54.
Shibata, R., 1983. Asymptotic mean efficiency of a selection of regression variables. Ann. Inst. Statist. Math. 35 (1), 415–423.
S&P, 2014. Index mathematics - Methodology, us.spindices.com.
S&P, 2015. Dow jones Total Stock Market Indices Methodology, us.spindices.com.
Wand, M.P., Jones, M.C., 1994. Multivariate plug-in bandwidth selection. Comput. Stat. 9 (2), 97–116.
Wang, S., Vergne, J.P., 2017. Buzz factor or innovation potential: What explains cryptocurrencies returns?. PLoS One 12 (1), e0169556.
White, L.H., 2014. The market for Cryptocurrencies. SSRN Scholarly Pape ID 2538290. Social Science Research Network, Rochester, NY.
Wilshire Associates, 2015. Wilshire 5000 Total Market Index Methodology, wilshire.com.
Woodroofe, M., 1982. On model selection and the ARC sine laws. Ann. Statist. 10 (4), 1182–1194.

Scientometrics (2018) 117:1023–1040
https://doi.org/10.1007/s11192-018-2910-8

How to measure the performance of a Collaborative Research Center

Alona Zharova[1] · Janine Tellinger-Rice[2] · Wolfgang Karl Härdle[1,3]

Received: 3 March 2018 / Published online: 22 September 2018

Abstract
New Public Management helps universities and research institutions to perform in a highly competitive research environment. Evaluating publicly financed research improves transparency, helps in reflection and self-assessment, and provides information for strategic decision making. In this paper we provide empirical evidence using data from a Collaborative Research Center (CRC) on financial inputs and research output from 2005 to 2016. After selecting performance indicators suitable for a CRC, we describe main properties of the data using visualization techniques. To study the relationship between the dimensions of research performance, we use a time fixed effects panel data model and fixed effects Poisson model. With the help of year dummy variables, we show how the pattern of research productivity changes over time after controlling for staff and travel costs. The joint depiction of the time fixed effects and the research project's life cycle allows a better understanding of the development of the number of discussion papers over time.

Keywords Research performance · Fixed effects panel data model · Network · Collaborative Research Center

Mathematics Subject Classification 62-07 · 62-09 · 62P20

JEL C23 · C13 · M19

Financial support from the German Research Foundation (DFG) via Collaborative Research Center 649 "Economic Risk" and International Research Training Group 1792 "High Dimensional Nonstationary Time Series", Humboldt-Universität zu Berlin, is gratefully acknowledged. We are thankful for the assistance provided by Nicole Hermann und Dominik Prugger.

✉ Alona Zharova
alona.zharova@hu-berlin.de

[1] School of Business and Economics, Humboldt-Universität zu Berlin, Berlin, Germany

[2] domino e.V., Birkenwerder, Germany

[3] Singapore Management University, Singapore, Republic of Singapore

Introduction

New Public Management (NPM) emerged in the 1980s (Hood 1991) with the goal of improving efficiency and overall performance of public sector institutions by using business management approaches and models. NPM places a strong focus on permanent monitoring and evaluation of performance. Measuring research performance allows an analysis of the structural issues in science. It can thus facilitate the development of a scientific system and strengthen excellence in research.

This paper discusses Collaborative Research Centers (CRC)—long-term university-based research institutions funded by the German Research Foundation (DFG 2018). Evaluating publicly financed research results improves transparency, helps in reflection and self-assessment, and provides information for strategic decision making. Periodic monitoring of resource use and interim results allows CRC management to keep the finger on the pulse and to react to unfavourable phenomena promptly or to develop options for improvement; thereby, supporting success of the CRC.

There are numerous studies that concentrate on the evaluation of university research or research institutions in general (Pastor et al. 2015; Van den Berghe et al. 1998). Lee (2010) and Bolli and Somogyi (2011) discuss performance measurements for departments and research units. Jansen et al. (2007) and Carayol and Matt (2004) further investigate performance indicators for research groups. However, a CRC differs from common research units or institutions, because of its interdisciplinary background. The performance indicators used for the evaluation of a CRC should be designed specifically for its needs and purposes in order to reflect the behaviour of involved research fields and other underlying characteristics.

In this paper we focus on a selection of performance indicators for intermediate and final evaluations suitable for broad applicability within CRCs and identifying a relationship between productivity and resource use of CRCs that may have implications for funding policy. The goals of this paper include: (1) selecting performance indicators suitable for a CRC; (2) visualizing goals vs. results, societal impact and the interdisciplinarity structure of research results of a CRC; (3) analysis of a dependence structure between financial inputs and research output of a CRC and development of research productivity over time.

To achieve these objectives, we use twelve years (2005–2016) of Collaborative Research Center 649 “Economic Risk” (CRC 649) data on 35 sub-projects. For each sub-project we observe yearly staff costs, travel costs and number of discussion papers (DPs). The life span of each sub-project varies, which results in an unbalanced panel.

Schröder et al. (2014) indicate that the proposal for funding determines objectives for the research activity. To examine the correspondence between objectives and research results of the CRC, we carry out a semantic analysis of proposals and abstracts from published DPs. As a result, we find that both use 50% of the same words.

Apart from research activity, a CRC has an impact on society through public events, transfer of knowledge or promotion of young researchers. For instance, young researchers usually perform specific theoretical or practical research that is also used for their Ph.D. thesis. Collecting data on their further career helps to better understand this impact. With the help of a mosaic plot, we visualize three important dimensions of young researchers careers after receiving their Ph.D. within the CRC: gender, location and area of work. For example, we show that almost 70% of young researchers who received their Ph.D. during CRC membership found later a job in academia.

Through a network analysis, we illustrate the interdisciplinarity structure of the research results and find out that most DPs were published in the fields of mathematical and quantitative methods, followed by financial economics, macroeconomics and monetary economics.

To study the relationship between research outcomes and funding for the CRC, we regress the number of DPs on staff and travel costs using sub-project-level data. With the help of year dummy variables added to the model, we show how the pattern of the sub-projects' productivity changed from 2005 to 2016 after controlling for staff and travel costs. Since the level of spending from the previous year and the preceding number of DPs may influence the current number of DPs, we additionally control for the lagged variables. The productivity of each sub-project may differ due to some heterogeneity or individual effects, such as the skills of a principal investigator (PI), average abilities or skills of researchers employed at the sub-project, or the specific behavior of a research field. For instance, working on a publication with one vs. more co-authors, writing in English vs. other languages, or publishing in books vs. articles may affect the research outcomes (Zharova et al. 2017). Therefore, we allow for the possibility of individual sub-project's effects. Considering the data structure, we apply a time fixed effects panel data (FE) model. Since the number of DPs is a count variable, we also apply a fixed effects Poisson (FEP) model.

We show that an increase of staff costs by 100% leads to an expected increase in the number of DPs by roughly 43% (FE) or 1.62 DPs (FEP). Travel costs have a diminishing effect on the number of DPs according to estimation results of the considered models. The previous level of both staff and travel costs negatively influence the number of DPs. We depict the estimates of coefficients of the dummy variables for years and find that the development trend corresponds with the stages of a project's life cycle. For instance, the most significant declines in the number of DPs take place during the stage of theoretical and empirical research, whereas the finalization stage corresponds with the growth in the number of published DPs.

The programmed R codes are available on the web-based repository hosting service and collaboration platform GitHub.

The remainder of the paper is structured as follows. Literature review on performance indicators is presented in Sect. 2. Section 3 describes the data and provides some preliminary descriptive analyses. Section 4 introduces the methodology and shows empirical results. Finally, Sect. 5 summarizes the results.

Literature review

The combination of a peer-reviewed process and quantitative indicators is common practice in research performance assessment. The German Council of Science and Humanities (WR, germ.—Wissenschaftsrat) suggests evaluating the research institutions within three dimensions (research, promoting young researchers and knowledge transfer), which contain nine research performance criteria (WR 2004). We select five criteria relevant to a CRC and provide a literature review on suitable indicators that may reflect the performance of the CRC.

1. *Research quality* shows originality and novelty of research outputs, trustworthiness of methodology, impact and relevance for further research (Table 1).

2. *Effectiveness* reflects the contribution of all sub-projects to the development of expertise in the research field within the CRC and beyond (Table 2).

Table 1 Research quality

Indicator	Definition	Literature
Relative reception success		
C_{Pub}	Relation of total number of citations (NC_{Pub}) to the total number of publications (N_{Pub})	Wissenschaftsrat (2012), Diem and Wolter (2013), Donner and Aman (2015)
C_{Pub}/FC_m	Number of citations per publication in relation to the citation's average of the field	Wissenschaftsrat (2012), Abramo and D'Angelo (2011), Moed et al. (2011), Van den Berghe et al. (1998)
C_{Pub}/JC_m	Number of citations per publication in relation to the citation's average of the journal	Moed (2010), Wissenschaftsrat (2012)

3. The *efficiency* criterion describes a quantity of research outputs in relation to a specific input, i.e. total costs, staff expenditures, number of staff, etc. (Table 3).

4. *Research enabling* relates to scientific activities that facilitate and support the research of young researchers (Table 4).

5. *Knowledge transfer* defines the transfer of research results and products or distribution of knowledge (Table 5).

Data

Collaborative Research Centers (CRC) are interdisciplinary research institutions financed through the German Research Foundation (germ.—Deutsche Forschungsgemeinschaft, DFG). The goal of a CRC is to pursue interdisciplinary innovative research by bringing together scholars from different research fields within multiple research projects, also called sub-projects. The classical CRC consolidates cooperation between several universities or non-university research institutions with at least 60% of all sub-projects based in the coordinating university (DFG 2018).

CRCs are granted for four years and depending on the results of the interim evaluations can be prolonged twice for a maximum period of twelve years. During the assessment each sub-project undergoes a critical appraisal. Depending on a change in research program or staff turnover (professors), a CRC can also submit proposals for new sub-projects. As a result, the number of research projects may vary between phases.

In this paper we provide empirical evidence using data from a Collaborative Research Center 649 "Economic Risk" (hereinafter referred to as the CRC). The CRC was launched in 2005 for a four-year term and extended twice, for a total life span of twelve years. As an interdisciplinary research center, it combined economics, mathematics and statistics and pursued research within three primary areas: (1) microeconomics, in particular individual and contractual answers to risk; (2) quantitative projects, in particular financial markets and risk assessment; (3) macroeconomic risks. For more information, we refer to the website of the CRC (CRC 649 2016).

The total number of the CRC sub-projects within three four-year phases is 35, but the number of sub-projects per phase varies from 16 to 21. Since the sub-projects of the CRC have different life periods, the data set does not have the observations for all years that indicates an unbalanced panel, see Fig. 1. The main reason for the panel being unbalanced

Table 2 Effectiveness

Indicator	Definition	Literature
Research activity		
N_{Costs}	Total amount of the third party expenses (TPE)	Wissenschaftsrat (2012), Schmoch and Schubert (2009)
N_{Staff}	Total number of staff financed from third party funds (TPF)	Carayol and Matt (2004), Wissenschaftsrat (2012)
RA_{unit}	Research activity of unit (sub-project, SP)—multiplication of the total number of publications and the total number of citations of a unit with regard to the institutions-wide number of citations for the analyzed period ($\text{RA}_{\text{SP}} = N_{\text{Pub}_{\text{SP}}} * \text{C}_{\text{Pub}_{\text{SP}}}/C_{\text{Pub}_{\text{CRC}}}$)	Pastor et al. (2015)
Research productivity		
N_{Pub}	Total number of publications	Wissenschaftsrat (2012), Abramo and D'Angelo (2011), Diem and Wolter (2013), Moed et al. (2011), Hornbostel (1991)
NC_{Pub}	Total number of citations	Wissenschaftsrat (2012)
FN_{Pub}	Fractional productivity—total number of contributions to publications, where each contribution is a publication divided by the number of co-authors	Abramo et al. (2009), Abramo and D'Angelo (2011)
ScS_{Pub}	Scientific strength—weighted sum of publications authored by each person, where the weight for each publication is the number of citations per publication in relation to the citation's average of the field ($C_{\text{Pub}}/\text{FC}_{\text{m}}$)	Abramo and D'Angelo (2011), Abramo et al. (2009)
h	*h*-index	Hirsch (2005), Bornmann (2013)
Visibility of the CRC		
AbsC_{Pub}	Absolute citation count in the light of maximum citation count of a single publication ($C_{\text{Pub}_{\text{max}}}$) and the number of non-cited publications (N_{ncPub})	Wissenschaftsrat (2012)
Reputation		
	List of scientific prizes and awards	Zheng and Liu (2015), Wissenschaftsrat (2012)
Professional activity		Wissenschaftsrat (2012)
	Editorships	
	Review activities	
	Editorial board memberships	
	Academic functions	
	Academic memberships	
	Organized conferences and workshops	

is the attrition of sub-projects, as a result of research project's termination or the leave of principal investigators to other universities, and the establishment of new research projects during the prolongation phases. For instance, twelve sub-projects had a life cycle of four

Table 3 Efficiency

Indicator	Definition	Literature
N_{Pub}/N_{Staff}	Relation of the number of publications (N_{Pub}) to the number of research staff (N_{Staff})	Pastor and Serrano (2016), Wissenschaftsrat (2012), Abramo and D'Angelo (2011)
NC_{Pub}/N_{Staff}	Relation of the number of citations of publications (N_{Pub}) to the number of research staff (N_{Staff})	Wissenschaftsrat (2012), Lee (2010)
N_{Costs}/N_{Staff}	Relation of the TPE to the total number of research staff (N_{Staff})	Wissenschaftsrat (2012), Pastor and Serrano (2016), Barra and Zotti (2016)

Table 4 Research Enabling / Promotion of young researchers

Indicator	Definition	Literature
Promotion of young researchers		
N_{YR}	Total number of positions for young researchers	Wissenschaftsrat (2012)
$N_{Ph.D.}$	Total number of defended Ph.D.	Wissenschaftsrat (2012), Diem and Wolter (2013), Grözinger and Leusing (2006), Schmoch and Schubert (2009)
$D_{Ph.D.}$	Average duration of Ph.D. study	Wissenschaftsrat (2004)
$N_{Pub_{Ph.D.}}$	Total number of publications by young researchers	Wissenschaftsrat (2004)
	List of awards and prizes of young researchers	Wissenschaftsrat (2012)
	List of calls and appointments for young researchers	Wissenschaftsrat (2012)

Table 5 Knowledge transfer

Indicator	Definition	Literature
N_{Pat}	Number of patents	Wissenschaftsrat (2011), Carayol and Matt (2004)
	List of Transfer projects	
	List of activities in public relations	Wissenschaftsrat (2012)
	List of research products and teaching materials	Wissenschaftsrat (2012)

years, eleven sub-projects lasted for eight years and five sub-projects existed twelve years (see Fig. 1).

Principal investigators (PIs) lead sub-projects. From 35 sub-projects 83% have one PI and 17% have two PIs. Since three PIs participate in two sub-projects, the CRC counts 38

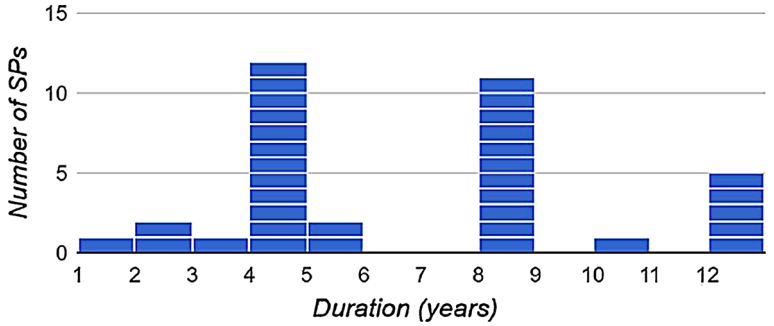


Fig. 1 Distribution of sub-projects (SP) over life span in years

PIs in total over twelve years. PIs of all three academic ranks participate in the CRC: full professors (76%), junior professors (19%) and postdoctoral researchers (5%).

The CRC uses 62% of resources on average to finance the research staff working within sub-projects, in particular doctoral (Docs) and postdoctoral (PostDocs) researchers. In addition, all members of the CRC may use its central funds for travel costs, organizing conferences and workshops, inviting guest lecturers and researchers, gender equality etc.

The amount of research staff working within sub-projects differs, depending on the scope and complexity of the research program. Each sub-project counts from 0.5 to 2.5 full-time equivalents (FTEs) of researcher positions per year. The FTEs are often split and used to hire more research staff, i.e. 2 researchers with 50% financing, or to top up researchers that are already employed and who are financed by other sources. Figure 2 shows the distribution of sub-projects according to the number of FTEs per year. For instance, 21 sub-projects have one FTE per year on average, eight sub-projects hire staff on 0.5 FTEs, four sub-projects use 1.5 FTEs and two sub-projects have each 2 and 2.5 FTEs.

In this paper we use data from annual financial reports, internal publications' and discussion papers' (DPs) databases and CRC's newsletter. Additional insight is gathered from the texts of one proposal for a launch and two proposals for a prolongation of the CRC 649 (2005–2008, 2009–2012, 2013–2016) which were submitted to the DFG. On the one hand, one can see such proposals as goals that the CRC sets for each period. On the other hand, the published DPs encompass the achieved results of the research activity. We undertake a semantic analysis on both informational sources, i.e. 61 summaries of sub-projects from three proposals and abstracts of 771 DPs. The two word clouds of the top 75 keywords are illustrated in Fig. 3. We find that both use 50% of the same words. The

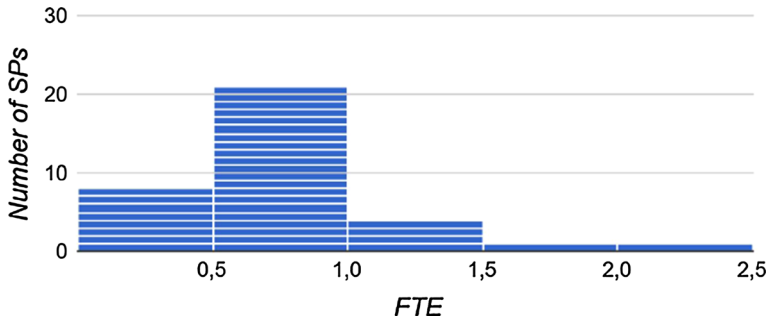


Fig. 2 Distribution of sub-projects according to the number of research staff (in FTE per year)

different size of the same words, for instance the word "risk", indicates that the number of times the word is mentioned in the proposals and abstracts differs.

One of the primary goals of a CRC is the high-quality instruction, supervision and support of young researchers. The common result of this process is a Ph.D. defence. Collecting data on the further career of the young researchers helps to better understand the impact on society. For instance, one may wonder how many females that worked and defended their Ph.D. thesis at the CRC are afterward working in academia in Germany? To visualize such data we use a mosaic plot in Fig. 4.

The vertical axis splits the individuals according to their gender. The data are further divided into two groups on the upper horizontal axis according to the location of the job. The lower horizontal axis shows how many people received a contract in academia or other fields. The width and height of each segment represent the number of observations within each group. Consider the 65 members of the CRC that received their Ph.D. from 2005 to 2016. There are 11 female researchers that received jobs in academia in Germany and 6 in other countries. For males that stayed in academia, the number is 21 for Germany and 7 for other countries. This means that almost 70% of young researchers who received their Ph.D. during CRC membership found later a job in an academic institution.

The proportion of 36.9% of female researchers is quite low in comparison to 50.4% for female doctoral students within CRCs in social sciences and humanities, but higher than 25.7% within CRC in mathematical and natural sciences (DFG 2017). However, since the CRC pursued interdisciplinary research in both social and mathematical sciences, the CRC proportion corresponds to the value in-between. As a part of the communication processes with alumni and mentoring of CRC young researchers, the CRC invited its former members who got promoted in academia as guest lecturers for CRC seminars or as guest researchers to work on papers jointly with PIs and/or younger CRC generations.

In order to understand if the intended interdisciplinarity occurred, we analyze DPs that serve as an outcome of the CRC research activity. Almost each DP has codes indicating subject fields according to the Journal of Economic Literature (JEL) classification in the economic sciences (see JEL 2018).

We show the network of collaborating disciplines in Fig. 5. The small gold circles introduce the DPs, whereas the nodes leading to the bigger blue circles indicate the JEL

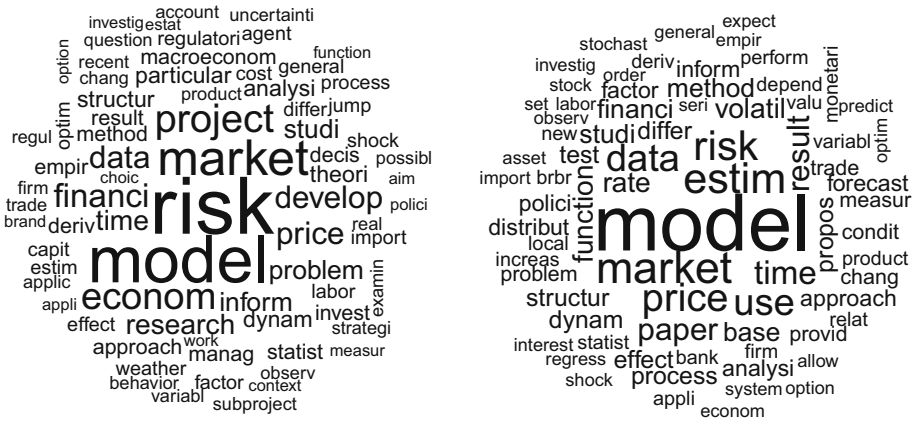


Fig. 3 Semantic analysis of goals (left; 61 summaries from sub-projects of three proposals for the CRC) versus results (right; 771 abstracts from DP)

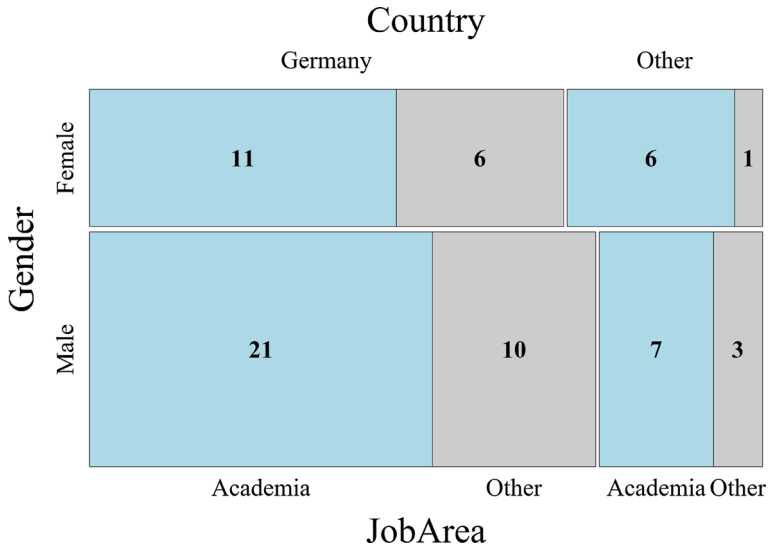


Fig. 4 Mosaic plot of job type, location and gender of 65 CRC members who received their Ph.D. between 2005 and 2016 (as of Dec 2016)

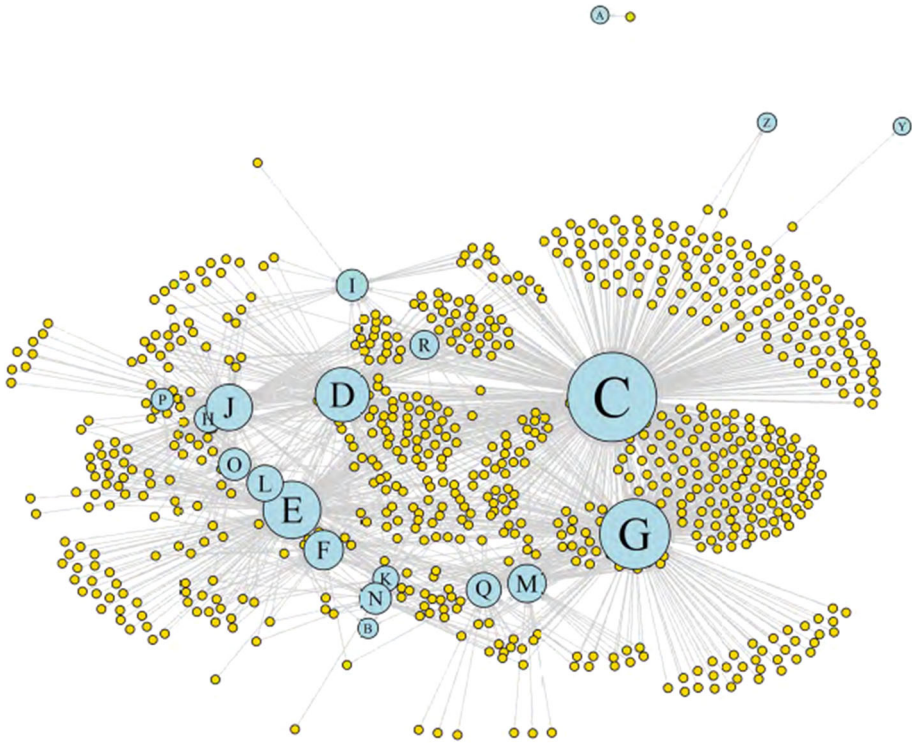


Fig. 5 Network of 760 discussion papers (yellow) and 20 JEL codes (blue) published from 2005 to 2016. (Color figure online)

code of the corresponding research area. The size of each blue circle reflects the relative number of references to DPs. The explanation of JEL codes is given in Table 6. For instance, most of the DPs were published in the C area, i.e. mathematical and quantitative methods. They are followed by G (financial economics), E (macroeconomics and monetary economics) and D (microeconomics). These four fields with higher research output correspond to the three primary areas of the CRC. Note that the DPs that involve research in more than one field are connected to two or more JEL codes simultaneously. This confirms the interdisciplinary character of the CRC research output.

One more factor influencing the variability of the number of DPs across research fields is the area of expertise of PIs and research staff. Figure 6 shows the cumulative number of PIs within their areas of expertise and Fig. 7 depicts the cumulative number of CRC research staff (in FTE) working within same research areas for twelve years. Since the attrition of some sub-projects and establishment of new ones influences the availability of PIs and research staff and accordingly their expertise within the CRC life cycle, we use cumulative numbers. We also use weights for the number of the sub-projects and expertise areas for each PI to equalize the total time available for research. For example, the PI who is an expert in four research areas receives 0.25 for each JEL code and the PI who leads two sub-projects has 0.5 for the distribution within JEL areas of each project.

Figures 6 and 7 show, for instance, that the area D reveals 24 years of PIs expertise and 15 years of research staff (in FTE) work. Both figures provide evidence that the most expertise is concentrated within the area C, followed by E, D, G and Q. This also explains the concentration of research output within corresponding JEL areas in Fig. 5. The

Table 6 JEL Classification System

Code	Research field
A	General Economics and Teaching
B	History of Economic Thought, Methodology, and Heterodox Approaches
C	Mathematical and Quantitative Methods
D	Microeconomics
E	Macroeconomics and Monetary Economics
F	International Economics
G	Financial Economics
H	Public Economics
I	Health, Education, and Welfare
J	Labor and Demographic Economics
K	Law and Economics
L	Industrial Organization
M	Business Administration and Business Economics/Marketing/Accounting/Personnel Economics
N	Economic History
O	Economic Development, Innovation, Technological Change, and Growth
P	Economic Systems
Q	Agricultural and Natural Resource Economics/Environmental and Ecological Economics
R	Urban, Rural, Regional, Real Estate, and Transportation Economics
Y	Miscellaneous Categories
Z	Other Special Topics

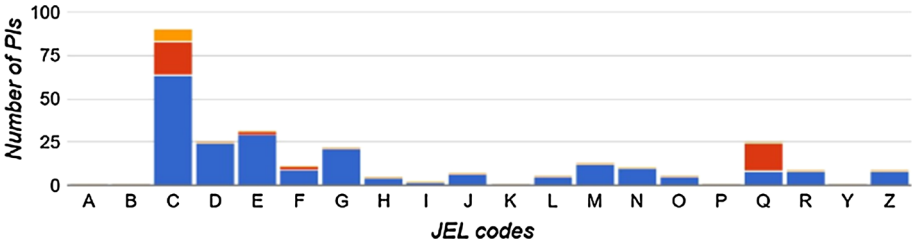


Fig. 6 Cumulative number of PIs (in PI years; full professors—blue, junior professors—red, postdoctoral researchers—orange) from 2005 to 2016 (weighted by the number of research fields and sub-projects) with expertise in corresponding JEL research fields. (Color figure online)

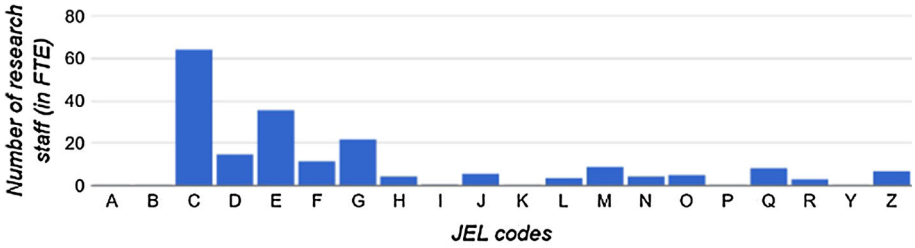


Fig. 7 Cumulative number of research staff in FTE (in staff years; weighted by the number of research fields) from 2005 to 2016 working within corresponding JEL research areas

correlation between the number of DPs and number of PIs specializing in the same JEL areas is 93.8% (95% for full professors only), whereas the correlation between the number of DPs and the amount of research staff (in FTE) working within same fields is 95.1%.

Analysis of research productivity

The observed time series across the same sub-projects indicate the longitudinal or panel structure of the data. To investigate the relationship between the input and the output variables, we use the methods designed for panels.

Methodology

The basic framework for the panel data analysis shows the model (Wooldridge 2002):

$$y_i = \beta X_i + u_i, \quad i = 1, \ldots, K, \tag{1}$$

where $y_i = (y_{i1}, \ldots, y_{iT})^\top$ is a $(1 \times T)$ vector of observations for $t = 1, 2, \ldots, T$, $X_i = (x_{i1}^\top, \ldots, x_{iT}^\top)^\top$ is a $(K \times T)$ matrix of observations, β is a $(K \times 1)$ vector of coefficients and u_i is a $(1 \times T)$ vector of unobservables.

The unobserved sub-project's effect may contain such factors as publishing behavior in a research field, average researchers' abilities or skills of principal investigators of sub-projects that should be roughly constant over time.

We allow for arbitrary correlation between the unobserved sub-project's heterogeneity or fixed effects c_i and the observed explanatory variables x_{it} and, therefore, use the fixed effects model for each i (Wooldridge 2016):

$$y_{it} = \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + c_i + u_{it}, \quad t = 1, 2, \ldots, T, \quad i = 1, 2, \ldots, K, \tag{2}$$

where y_{it} includes dependent variables and x_{it} independent variables for individual i at time t, $\beta_1, \ldots, \beta_k$ are the unknown coefficients, c_i is individual effect or individual heterogeneity and u_{it} are idiosyncratic errors that change across individuals i and time t.

The fixed effects estimator (or the within estimator) is obtained as the pooled OLS estimator on the time-demeaned variables. The strict exogeneity assumption on explanatory variables, $E(u_{it}|\mathbf{X}_i, c_i) = 0$, provides that the fixed effects estimator is unbiased (Wooldridge 2016). As the number of sub-projects (clusters) is large, statistical inference after OLS should be based on cluster-robust standard errors to account for heteroscedasticity and within-panel serial correlation (Cameron and Miller 2015).

Next, we are interested in the pattern of sub-projects' productivity, i.e. number of produced discussion papers, in different time periods. For this purpose we use time fixed effects that change over time but are constant across sub-projects. We include the dummy variables for $T - 1$ years to avoid the multicollinearity. Usually the first year is selected as a base year. The time fixed effects model (FE) is (Stock and Watson 2003):

$$y_{it} = \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + \delta_1 + \delta_2 D_2 + \cdots + \delta_T D_T + c_i + u_{it}, \tag{3}$$

where $D_2, \ldots, D_T$ are time effects and $\delta_1, \ldots, \delta_T$ are the parameters to estimate.

When the dependent variable involves count data, it has a Poisson distribution instead of a normal distribution. Hausman et al. (1984) introduce a fixed effects Poisson model (FEP) as:

$$E(y_{it}|x_i, a_i) = a_i \mu(x_{it}, \beta_0), \quad t = 1, 2, \ldots, T, \tag{4}$$

where β_0 is a $(1 \times K)$ vector of unknown parameters to be estimated and μ is the conditional mean. Wooldridge (1999) further derives a consistent estimator for FEP using a quasi-conditional maximum likelihood estimator (QCMLE).

Empirical results

Before presenting the estimates, we explain some specifications of the model. Since the yearly staff and travel costs are in nominal Euros, a slight increase may happen due to inflation. One possibility to deal with this is an adjustment using a Consumer Price Index (CPI). Another way to track the effect of real spendings is the use of a logarithmic form. The interpretation of the estimation results is then done using the level-log model. Here we use the second approach.

Table 7 presents the results of FE (1) and (2), and FEP (3) and (4) models for the number of DP as a dependent variable. The parameters of interest are staff costs $\beta_{\text{logStaffCosts}}$, travel costs $\beta_{\text{logTravelCosts}}$ and year-specific influence δ_{year}. We also include lagged variables into the models (2) and (4), since the current number of research outputs may be affected by the previous number of publication and invested funds in economic sciences and mathematics (Zharova et al. 2017). The models (2) and (4) encompass the number of DPs $\beta_{nDP_{t-1}}$, staff costs $\beta_{\text{logStaffCosts}}$ and travel costs $\beta_{\text{logTravelCosts}}$ in the time $t - 1$. The intercept *const* is the average of individual effects c_i across all sub-projects that is reported by Stata. We use cluster-robust standard errors to account for heteroscedasticity. The significance level of all estimates decreases as a result of standard error adjustment (Wooldridge 2016).

Table 7 Estimation results for time fixed effects (within) regression (models (1) and (2)) and fixed effects Poisson regression (models (3) and (4)) with number of DP (*nDP*) as the dependent variable and with robust standard errors adjusted for clusters in sub-projects

Dependent variable: *nDP*	FE model		FEP model	
	(1)	(2)	(3)	(4)
$\beta_{\text{logStaffCosts}}$	1.38**	1.62*	0.47***	0.43**
	(0.61)	(0.88)	(0.12)	(0.19)
$\beta_{\text{logTravelCosts}}$	− 0.94*	− 0.34	− 0.22**	− 0.04
	(0.55)	(0.47)	(0.10)	(0.09)
δ_{2006}	1.61	1.92	0.25	0
	(1.36)	(1.61)	(0.26)	(omit.)
δ_{2007}	− 1.20	− 2.55	− 0.30	− 0.98***
	(1.38)	(2.46)	(0.31)	(0.25)
δ_{2008}	− 0.95	− 2.03	− 0.23	− 0.97***
	(1.30)	(2.10)	(0.32)	(0.36)
δ_{2009}	− 2.05*	− 3.16	− 0.54*	− 1.20***
	(1.13)	(1.98)	(0.33)	(0.23)
δ_{2010}	− 1.93*	− 2.13	− 0.51*	− 1.03***
	(1.14)	(2.68)	(0.30)	(0.31)
δ_{2011}	1.10	0	0.33*	0
	(0.70)	(omit.)	(0.20)	(omit.)
δ_{2012}	− 2.79*	− 3.60*	− 0.71**	− 1.90***
	(1.46)	(1.78)	(0.34)	(0.20)
δ_{2013}	− 2.98**	− 3.18	− 0.80**	− 1.32***
	(1.30)	(2.52)	(0.32)	(0.41)
δ_{2014}	− 1.36	− 1.73	− 0.44	− 0.99***
	(0.95)	(1.61)	(0.27)	(0.37)
δ_{2015}	− 2.55**	− 1.90	− 0.74**	− 1.02***
	(1.17)	(1.77)	(0.33)	(0.31)
δ_{2016}	− 0.30	0	− 0.31	− 0.69*
	(1.79)	(omit.)	(0.36)	(0.41)
const	− 2.37	0.05		
	(5.29)	(10.09)		
$\beta_{nDP_{t-1}}$		0.02		− 0.01*
		(0.16)		(0.03)
$\beta_{\text{logStaffCosts}_{t-1}}$		− 0.66		− 0.25
		(0.59)		(0.23)
$\beta_{\text{logTravelCosts}_{t-1}}$		− 0.21		− 0.02
		(0.58)		(0.13)
R^2	0.20	0.21		
AIC	706	437	463	253
BIC	742	469	501	258

***, ** and * indicate a statistical significance at 1%, 5% and 10% level, respectively. Standard deviation is provided in brackets

In (2) and (4) two years were omitted because of collinearity. In (3) five observations were dropped out of the analysis because there was only one observation per group. Performing analysis on unbalanced data slightly increases the estimated effects of considered variables, but the general idea remains unchanged (Wooldridge 2016).

In the model (1) we see the positive, significant effect of staff costs on the number of DPs. 1.38/100 is the unit change in *n DP* when staff expenses increase by 1%. In other words, a 100% increase in staff costs leads to an increase in the number of DPs by 1.38. Similarly, the model (2) shows that a 100% increase in staff costs increases the number of DPs by 1.62, holding other variables constant. The fit of the FE models in (1) and (2) in Table 7 with *nDP* as the dependent variable is almost the same, indicating that including lagged variables does not significantly improve the model.

The FEP estimates have a different interpretation. For instance, the coefficient on $\beta_{\text{logStaffCosts}}$ shows that a rise of staff costs by 100% leads to an increase of the number of DPs by 47% and 43% for models (3) and (4) correspondingly. The coefficients on staff costs estimates for four models in Table 7 are significant at 1% to 10% level. The influence of previous values of staff costs on the number of DPs is negative and insignificant.

Travel costs have a diminishing effect on the number of DPs according to estimation results of considered models. The coefficient on $\beta_{\text{logTravelCosts}}$ implies that, if we increase the travel costs by 100%, we expect the number of DP to decrease by 0.94 DP due to FE model (1). The Poisson coefficient in (3) means that an increase in *logTravelCosts* by 10% decreases *nDP* by 2% (0.22×0.10).

The coefficients on the year dummy variables reveal how the average productivity of sub-projects changes over time. As 2005 is selected as the base year, it is not reported with a coefficient. The coefficient on δ_{2006} in model (1) shows that, on average, 1.6 DPs are attributed to the year effect of 2006 holding other factors fixed. In Poisson case (3) one suggests that the expected number of DPs in 2006 is 25% higher than on average. The coefficients on δ_{2006} and δ_{2011} indicate a positive increase in the number of DPs even without changing expenses. The omission of year dummies would lead to the attribution of this positive effects to the effects of costs change.

One can see that the year effects have a negative impact on the number of DPs in the majority of years for all models. The project's life cycle could explain this. Research projects generally have five main stages: proposal development, funding review, project start-up, performing research and finalization of the project. We map the estimates of coefficients of the models and fit the stages of life cycles in Figs. 8 and 9. Proposal development and funding review take place before 2005 and are not depicted in these Figures.

A highly demanding application for a CRC requires extensive preliminary research. The results of this preliminary research are published as DPs in the first year 2005, thus, creating a specific bias towards later research outputs produced during the CRC's life time. The three following increases in the number of DPs take place mainly in the finalization stage caused by the publishing of research results in the final stage of projects. The research output of the last phase in 2016 shows part of the positive trend. In fact, 28 DPs were published in 2017, after the CRC was officially finished and financing ended. Three major declines could be explained by the theoretical and empirical stage of the research in the middle of each project life cycle. In summary, the joint depiction of the time fixed effects and the research project's life cycle allows a better understanding of the development of the number of DPs over time.

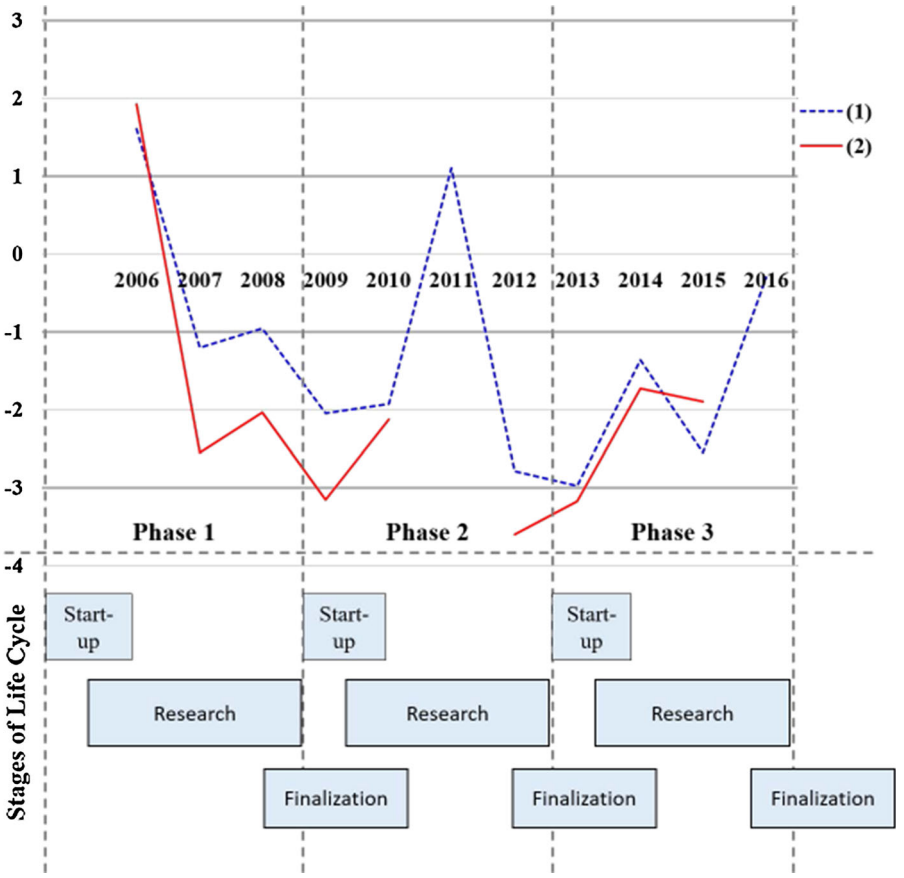


Fig. 8 Estimates of coefficients on the year dummy variables for the time fixed effects (within) regression (models (1) and (2)). The lower part of the figure shows the corresponding stage of the research project life cycle

Conclusions

Our findings show that the performance indicators suitable for the intermediate or final evaluation of a CRC facilitate a better understanding of the dependence structure between research productivity and financial inputs, and provide relevant information for successful decision and policy making.

As a result of semantic analysis of the text from proposals for the CRC submitted to the DFG and the abstracts from published DPs, we find out that two word clouds standing for goals and results use 50% of the same words. Aiming to visualize a further career path of young researchers that received their Ph.D. within the CRC, we use mosaic plot with dimensions gender, location and area of work. We show that almost 37% are females and 70% of young researchers found a job in academia.

We describe the interdisciplinary structure of research results with the help of the network analysis. We show that such fields as mathematical and quantitative methods, financial economics, macroeconomics and monetary economics and microeconomics are

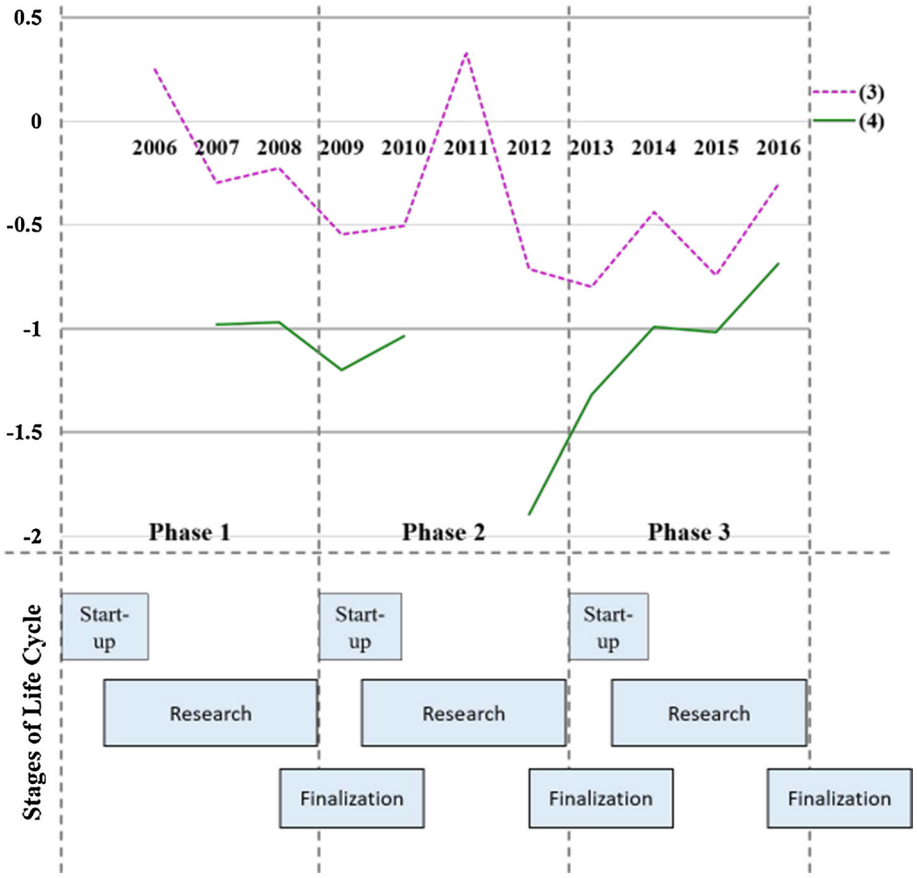


Fig. 9 Estimates of coefficients on the year dummy variables for the fixed effects Poisson regression (models (3) and (4)). The lower part of the figure shows the corresponding stage of the research project life cycle

the most reflected in the published DPs. These fields correspond to the primary research areas of the CRC. Moreover, the most of research output takes place in the areas that have more PIs with corresponding expertise. Additionally, the sub-projects with more research staff are expected to produce more DPs. The network visualization provides also evidence that one of the main goals of the interdisciplinary research center—interdisciplinarity—is achieved.

Using time fixed effects panel data model and fixed effects Poisson model, we show that increasing staff costs by 100% raises the number of DPs of a sub-project by 1.62 or 43% according to the estimates of FE and FEP models correspondingly. Travel costs have diminishing effect on the number of DPs according to our estimation results. We analyse the change in productivity of the CRC over time for reasons not captured by the other independent variables using the dummy variables for years. We depict the estimates of coefficients for years and show the possible association between the trend and the stages of a project's life cycle. For instance, the major declines in the number of DPs take place during the stage of theoretical and empirical research, whereas the finalization stage may correspond to the growth in the number of published DPs.

References

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2009). Research collaboration and productivity: Is there correlation? *Higher Education*, *57*(2), 155–171.

Abramo, G., & D'Angelo, C. A. (2011). National-scale research performance assessment at the individual level. *Scientometrics*, *86*, 347–364.

Barra, C., & Zotti, R. (2016). Measuring efficiency in higher education: An empirical study using a bootstrapped data envelopment analysis. *International Advances in Economic Research*, *22*, 11–33.

Bolli, T., & Somogyi, F. (2011). Do competitively acquired funds induce universities to increase productivity? *Research Policy*, *40*(1), 136–147.

Bornmann, L. (2013). How to analyze percentile citation impact data meaningfully in bibliometrics: The statistical analysis of distributions, percentile rank classes, and top-cited papers. *Journal of the American Society for Information Science and Technology*, *64*(3), 587–595.

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster–robust inference. *Journal of Human Resources*, *50*(2), 317–373.

Carayol, N., & Matt, M. (2004). Does research organization influence academic production? Laboratory level evidence from a large European university. *Research Policy*, *33*(8), 1081–1102.

CRC 649. (2016). http://sfb649.wiwi.hu-berlin.de/about/index.php. Accessed 11 January 2018.

CRC Project on GitHub. http://www.github.com/QuantLet/CRC. Accessed 11 January 2018.

Deutsche Forschungsgemeinschaft (DFG, eng. - German Research Foundation). (2017). Chancengleichheits-Monitoring 2017. Antragstellung und -erfolg von Wissenschaftlerinnen bei der DFG. http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/chancengleichheit/chancengleichheits_monitoring_2017.pdf. Accessed 01 August 2018.

Deutsche Forschungsgemeinschaft (DFG, eng. - German Research Foundation). (2018). http://www.dfg.de/en//research_funding/programmes/coordinated_programmes/collaborative_research_centres/index.html. Accessed 11 January 2018.

Diem, A., & Wolter, S. C. (2013). The use of bibliometrics to measure research performance in educational sciences. *Research in Higher Education*, *54*, 86–114.

Donner, P., & Aman, V. (2015). Quantilbasierte Indikatoren für impact und Publikationsstartegie. Ergebnisse für Deutschland in allen Fachdisziplinen in den Jahren 2000 bis 2011, (p. 8). IFQ: Studien zum deutschen Innovationssystem.

Grözinger, G. & Leusing, B. (2006). Wissenschaftsindikatoren an Hochschulen. Europa-Universität Flensburg, International Institute of Management, Discussion Papers. 012, https://EconPapers.repec.org/RePEc:fln:wpaper:012.

Hausman, J., Hall, B. H., & Griliches, Z. (1984). Econometric models for count data with an application to the patents-R & D relationship. *Econometrica*, *52*(4), 909–938.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–16572.

Hood, C. (1991). A public management for all seasons? *Public Administration*, *69*(1), 3–19.

Hornbostel, S. (1991). Drittmitteleinverbung. Ein Indikator für universitäre Forschungsleistungen. *Beiträge zu Hochschulforschung*, *1*, 57–84.

Jansen, D., Wald, A., Franke, K. et al. (2007). Drittmittel als Performanzindikator der Wissenschaftlichen Forschung. *Koelner Z. Soziol. u. Soz. Psychol. 59*(1), 125–149.

JEL (Journal of Economic Literature) Classification System. (2018). https://www.aeaweb.org/econlit/jelCodes.php?view=jel. Accessed 11 January 2018.

Lee, G. J. (2010). Assessing publication performance of research units: Extensions through operational research and economic techniques. *Scientometrics*, *84*(3), 717–734.

Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, *4*(3), 265–277.

Moed, H. F., de Moya-Anegón, F., López-Illescas, C., & Visser, M. (2011). Is concentration of university research associated with better research performance? *Journal of Informetrics*, *5*, 649–658.

Pastor, J. M., & Serrano, L. (2016). The determinants of the research output of universities: Specialization, quality and inefficiencies. *Scientometrics*, *1029*(2), 1255–1281.

Pastor, J. M., Serrano, L., & Zaera, I. (2015). The research output of European higher education institutions. *Scientometrics*, *102*(3), 1867–1893.

Schmoch, U., & Schubert, T. (2009). Sustainability of incentives for excellent research—The German case. *Scientometrics*, *81*(1), 195–218.

Schröder, S., Welter, F., Leisten, I., Richert, A., & Jeschke, S. (2014). Research performance and evaluation? Empirical results from collaborative research centers and clusters of excellence in Germany. *Research Evaluation*, *23*(3), 221–232.

Stock, J. H., & Watson, M. W. (2003). *Introduction to econometrics* (1st ed.). London: Pearson.

Van den Berghe, H., Houben, J. A., de Bruin, R. E., Moed, H. F., Kint, A., Luwel, M., et al. (1998). Bibliometric indicators of university research performance in Flanders. *Journal of the American Society for Information Science*, *49*(1), 59–67.

Wissenschaftsrat. (2004). *Empfehlungen zu Rankings im Wissenschaftssystem Teil 1: Forschung* (pp. 6285–04). Drs: Hamburg.

Wissenschaftsrat. (2011). Empfehlungen zur Bewertung und Steuerung von Forschungsleistung, Drs 1656–11.

Wissenschaftsrat. (2012). *Bericht der Steuerungsgruppe zur Pilotstudie zur Weiterentwicklung des Forschungsratings* (pp. 2815–12). Drs: Köln.

Wooldridge, J. M. (1999). Distribution-free estimation of some nonlinear panel data models. *Journal of Econometrics*, *90*, 77–97.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data* (1st ed.). Cambridge: MIT Press Books.

Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach* (6th ed.). Boston: Cengage Learning.

Zharova, A., Härdle, W.K. & Lessmann, S. (2017). Is scientific performance a function of funds? SFB 649 Discussion Paper, 2017(28).

Zheng, J., & Liu, N. (2015). Mapping of important international academic awards. *Scientometrics*, *104*(3), 763–791.

Research Article

Ying Chen*, Wolfgang K. Härdle, Qiang He and Piotr Majer

Risk related brain regions detection and individual risk classification with 3D image FPCA

https://doi.org/10.1515/strm-2017-0011
Received April 14, 2017; revised August 3, 2018; accepted October 13, 2018

Abstract: Understanding how people make decisions from risky choices has attracted increasing attention of researchers in economics, psychology and neuroscience. While economists try to evaluate individual's risk preference through mathematical modeling, neuroscientists answer the question by exploring the neural activities of the brain. We propose a model-free method, 3-dimensional image functional principal component analysis (3DIF), to provide a connection between active risk related brain region detection and individual's risk preference. The 3DIF methodology is directly applicable to 3-dimensional image data without artificial vectorization or mapping and simultaneously guarantees the contiguity of risk related brain regions rather than discrete voxels. Simulation study evidences an accurate and reasonable region detection using the 3DIF method. In real data analysis, five important risk related brain regions are detected, including parietal cortex (PC), ventrolateral prefrontal cortex (VLPFC), lateral orbifrontal cortex (lOFC), anterior insula (aINS) and dorsolateral prefrontal cortex (DLPFC), while the alternative methods only identify limited risk related regions. Moreover, the 3DIF method is useful for extraction of subjective specific signature scores that carry explanatory power for individual's risk attitude. In particular, the 3DIF method perfectly classifies both strongly and weakly risk averse subjects for in-sample analysis. In out-of-sample experiment, it achieves 73 %–88 % overall accuracy, among which 90 %–100 % strongly risk averse subjects and 49 %–71 % weakly risk averse subjects are correctly classified with leave-k-out cross validations.

Keywords: fMRI, FPCA, GLM, risk attitude, SVD

MSC 2010: 62H12, 62P10

1 Introduction

Understanding people's risk preferences and how people make decisions under risk have both attracted much attention in industry and academia alike. Accurate risk classification is of benefit both to creditors including banks, retailers, mail order companies, utilities and various other organizations, and to the applicants avoiding over commitment, see [16]. While the traditional classification approaches rely on expert knowledge,

***Corresponding author: Ying Chen,** Department of Mathematics, National University of Singapore, Singapore; and Department of Statistics and Applied Probability, National University of Singapore, Singapore; and Risk Management Institute, National University of Singapore, Singapore, e-mail: stacheny@nus.edu.sg. http://orcid.org/0000-0002-2577-7348
Wolfgang K. Härdle, Ladislaus von Bortkiewicz Chair of Statistics, C.A.S.E. Center for Applied Statistics & Economics, Humboldt-Universität zu Berlin, Berlin, Germany; and Sim Kee Boon Institute (SKBI) for Financial Economics at Singapore Management University, Singapore, e-mail: haerdle@wiwi.hu-berlin.de
Qiang He, Department of Statistics and Applied Probability, National University of Singapore, Singapore, e-mail: hq19861027@gmail.com
Piotr Majer, Ladislaus von Bortkiewicz Chair of Statistics, C.A.S.E. Center for Applied Statistics & Economics, Humboldt-Universität zu Berlin, Berlin, Germany, e-mail: piotr.majer71@gmail.com

experience and even a subjective feeling to categorize an individual to be risk averse or risk seeking, there has been an increasing demand in statistical methods for quantitative complements to the formal art alike analysis systems. Discriminant analysis, linear regression, logistic regression and decision trees have been developed and implemented in literature.

To explain the decision making behaviors, classical expected utility theory has been widely adopted in economics, see [23, 28, 31, 50]. The utility theory assumes that a rational decision maker chooses a strategy that maximizes the average or expected value of a concave utility function among possible outcomes, see e.g. [36] for the properties of utility functions. The utility functions depend on parameters that represent individual's risk preferences and are estimated based on the individual's characteristics. Alternatively, risk-return models [30] determine the average or expected returns and the associated risks of different choices, and compute a risk-compensated value in the capital asset pricing models, see [43, 51, 52]. The traditional models, though demonstrating some decision making philosophy in a common sense, are unable to explain the heterogeneity in decision-making under similar risk attitudes from person to person in the experiments of behavioral economics and neuroscience, see [3, 5, 10, 21, 44].

Decision-making is indeed a complex neural process involving both cognitive and emotional factors. According to [23] and [44], individuals not only estimate the expected value of utility or return, but more importantly, they seem to adapt these estimates by subjective factors, such as risk preference. It thus becomes scientifically necessary and important to answer which parts of the human brain regulate specific decision-making tasks and which neural processes drive investment decisions, see [25, 33, 37, 41]. It is also interesting to ask whether the identification of the risk related brain regions helps to explain the heterogeneity of individual risk preference and its impact on making decision from the neural aspect.

The recent development on neural image data collection allows quantitative analysis to be possible. In modern risk perception and investment decision (RPID) experiments, subjects are requested to make decisions with uncertain outcomes and simultaneously their brain reactions are recorded as neural images by the functional magnetic resonance imaging (fMRI) scanner. The neural images or fMRI data reflects the changes in the brain's blood flow at volume and oxygen level during neural activities. The blood-oxygen-level-dependent (BOLD) signals are captured on 3-dimensional (3D) spatial maps of brain voxels during the experiments.

Given the fMRI data collected in the risk related experiments, specific brain regions have been found to be associated with risk related decision making. Tobler, O'Doherty, Dolan and Schultz [45] demonstrated that lateral orbifrontal cortex (lOFC) and medial orbifrontal cortex (mOFC) are related to the evaluation and the contrast of risky or sure choices. Mohr, Biele, Krugel, Li and Heekeren [33] discovered that risk averse individuals have greater brain activities in lateral orbifrontal cortex (lOFC) and posterior cingulate cortex (PCC). Mohr, Biele and Heekeren [32] evidenced the importance of anterior insula (aINS) and ventrolateral prefrontal cortex (VLPFC) to value processing, risk and uncertainty. Van Bömmel, Song, Majer, Mohr, Heekeren and Härdle [47] found parietal cortex (PC) is associated with value processing and selective attention. The risk related regions are quantified as the voxels significantly activated by the stimulus, which turn out to be contiguous in modest size relative to the visual or audial cortex. Two techniques – general linear model (GLM) method and principal component analysis (PCA) method – are by far the most popular to identify the risk related regions.

The model-based GLM technique depends on a parametric structure, see e.g. [9, 11, 48]. It only focuses on the neural information with a pre-defined design matrix and ignores any neural activity other than the priori specified modeling. The PCA technique is model free and has potential to detect risk related regions without making any constraint or subjective assumptions, see [2, 4, 27]. Without losing much variability, it extracts spatial factors to represent the risk related brain regions, while the individual risk attitude of the subject is explained by the factor loadings named signature scores via an orthogonal decomposition.

The PCA method however needs a conversion of the fMRI data to a vector of discrete signals, leading to extremely high dimensionality when applied to the high resolution image data. To solve the estimation challenge, singular value decomposition (SVD) has been proposed with a reduced dimension of covariance matrix, see [13]. Nevertheless, the PCA and SVD methods conducted in a discrete framework cannot guarantee the contiguity of risk related regions rather discrete voxels, see [19].

This motivates the adoption of functional principal component analysis (FPCA), see [39, 40]. In FPCA, the vectorized fMRI data is smoothed as a continuous curve, for which eigen-decomposition is performed, see [29, 47, 49]. Zipunnikov, Caffo, Yousem, Davatzikos, Schwartz and Crainiceanu [54] further proposed the functional SVD (FSVD) approach that improved computational efficiency with the utilization of the SVD technique. It is worth noting that the FPCA and FSVD methods both request vectorizing the BOLD signals that are naturally defined on 3D location coordinates to 1D domain. Given the high resolution of fMRI data, without sufficient knowledge of spatial interdependence of the brain, the pre-processing vectorization potentially impairs accuracy and efficiency for the risk related region detection and further for the risk classification.

It is necessary to ask why not directly analyze the fMRI signals in the 3D domain and how much accuracy can be improved by employing such a new technique. In our study, we propose a model-free 3-dimensional image functional principal component analysis (3DIF) method to identify risk related regions and extract subject signature scores. Simulation study and real data analysis demonstrate good quality of the detected risk related regions with stable accuracy and contiguity property. The 3DIF regions are further found to carry explanatory power for subjects' risk attitudes. In the application of risk classification, the 3DIF method reaches 100 % accuracy for in-sample analysis and 73 %–88 % overall accuracy for out-of-sample analysis. In particular, it correctly classifies 90 %–100 % strongly risk averse subjects and 49 %–71 % weakly risk averse subjects by using leave-k-out cross validations.

The remainder of the paper is structured as follows. Section 2 presents the RPID experiment and data. Section 3 details the 3DIF methodology and briefly reviews the alternative methods in literature. Section 4 reports the performance of the proposed 3DIF method under different scenarios. In Section 5, we implement the 3DIF to real data. Section 6 concludes.

2 RPID experiment and data

To investigate the mechanism of brain processes during the process of making decisions under risk, we analyze functional magnetic resonance imaging (fMRI) data on seventeen subjects who were exposed to an RPID experiment designed in [33]. The experiment uses streams of investment returns as stimuli and hypothesizes how individual risk attitude affects decisions in risky choices against sure choices. Figure 1 displays a graphic illustration of the experimental setup. Each experiment trial composes of two phases. The presentation phase displays a random Gaussian distributed return stream with ten observations that are sequentially displayed over 2×10 seconds. After a 2.5 seconds break, subjects are exposed in the decision phase to one of three types of tasks and have to give an answer within the next 7 seconds. The three types of tasks included the *decision* task, where subjects choose either a 5 % fixed return (sure choice) or the investment of the random return stream just shown (risky choice). In the other two tasks subjects report their *subjective expected return* (scaling from 5 % to 15 %) and *perceived risk* (from 0 = no risk to 100 = maximum risk) of the just displayed investment. Each trial is repeated 27 times, with the types of tasks randomly selected. In total, there are 3×27 trails for each subject. During the experiment, subjects were placed in the fMRI scanner and high resolution ($91 \times 109 \times 91$) images were acquired every 2.5 seconds.

The seventeen subjects were native German speakers, healthy and right-handed. All participants had no history of neurological or psychiatric diseases. They were paid for their participation and gave written informed consent. The return streams were independent from trial to trial, randomly drawn from a Gaussian distribution. The expected value of the return streams varied from 6 %, 9 %, to 12 % and standard deviations from 1 %, 5 % to 9 %. The combinations generated in total nine different Gaussian distributions associated with various risk-return relationships, e.g. low return (6 %) and low risk (1 %) as well as high return (12 %) and high risk (9 %).

The same data had been studied by two works in the existing literature. Mohr, Biele, Krugel, Li and Heekeren [33] conducted the general linear model (GLM) with six design factors. The factors are either subject specific values including e.g. return stream, perceived risk, expected value of the return stream, or dummy variables. The study detected value-reward related brain activity in bilateral dorsolateral prefrontal cortex

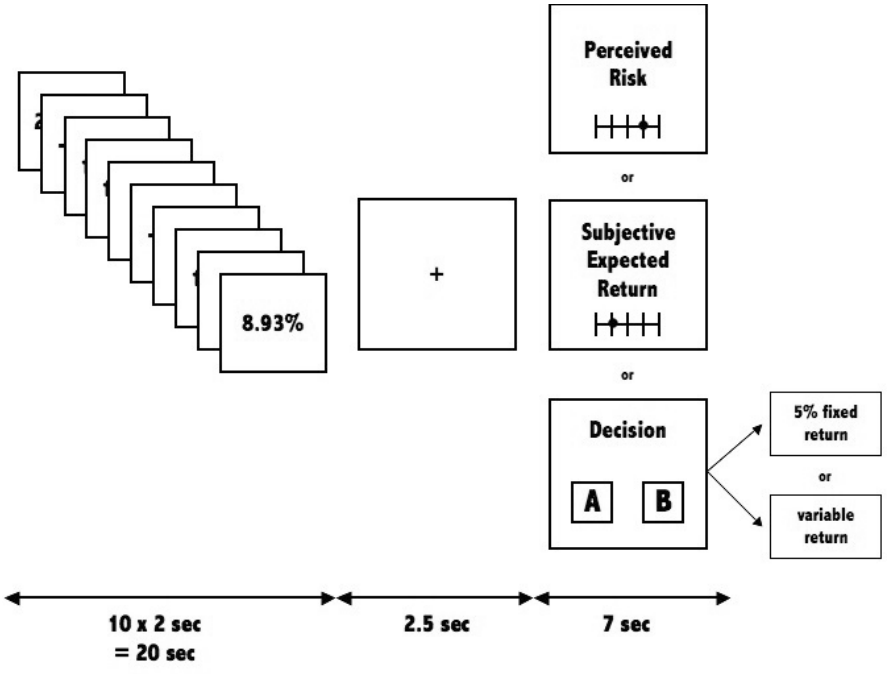


Figure 1: Graphic illustration of one trail of the RPID experiment, see [33].

(DLPFC), posterior cingulate cortex (PCC), ventrolateral prefrontal cortex (VLPFC), and medial prefrontal cortex (MPFC), which is consistent to [1, 22, 24–26, 35, 46]. It also found that perceived risk correlated significantly with the BOLD signal in the anterior insula (aINS), as documented in a variety of studies by [8, 14, 20, 34, 37, 38, 42]. However, GLM detection depends on the significance of statistical tests, which are hard to extract subject specific signals for further analysis.

Van Bömmel, Song, Majer, Mohr, Heekeren and Härdle [47] proposed a panel version of the dynamic semiparametric factor model (PDSFM) to reanalyze the data. The approach however only detected two important risk-related regions and did not contain any activation regions previously reported in [33] except Parietal Cortex (PC). Subject signature scores were extracted and used in risk classification. Using the variance of these stimuli responses as input for the classification algorithm, it obtained very high classification rates at 97 % for strongly risk averse individuals and 75 % for weakly risk averse with the SVM classifier by applying the double leave-one-out cross-validation algorithm.

3 Method

Our interest is to propose a dimension reduction technique on 3D space to improve prediction in the fMRI study of association between risk preferences and brain activity. In this section, we detail the 3D image functional principal component analysis (3DIF) method that is directly applicable to high-dimensional functional data and guarantees the contiguity of detected risk related brain regions. We show how to identify common spatial factors and extract subjective specific scores. The spatial factors are used to construct common risk activation regions that do not dependent on subjects, while the heterogeneity of individual risk attitude is explained by the subjective specific scores.

Let $Y_t^{(j)}(x_1, x_2, x_3)$ denote the observed fMRI signal at time $t = 1, \ldots, N$ for subject $j = 1, \ldots, J$ at 3D spatial location (x_1, x_2, x_3), where $x_1 \in \mathcal{P}_1, x_2 \in \mathcal{P}_2, x_3 \in \mathcal{P}_3$ are defined in a bounded cube $\mathcal{P}_1 \times \mathcal{P}_2 \times \mathcal{P}_3 \subset \mathbb{R}^3$. In our study, $J = 17$ subjects and $N = 1360$ scanned images. The brain is measured in a cube of size $[1, 91] \times [1, 109] \times [1, 91]$, i.e. around 10^6 voxels per scan. A tensor B-spline smoother is used to smooth each time-specific brain image and it leads to continuous 3D functional data, denoted as $f_t^{(j)}(x_1, x_2, x_3)$.

3.1 3D image functional principal component analysis (3DIF)

For any continuous functional data $f_t(\boldsymbol{x})$ with $\boldsymbol{x} = (x_1, x_2, x_3)$, one can represent it in a vector format

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{C}\boldsymbol{\phi}(\boldsymbol{x}), \tag{3.1}$$

where $\boldsymbol{C}$ is an $(N \times K^3)$-dimensional matrix of B-spline coefficients, N is the number of time points in the fMRI data and K refers to the number of knots in each spatial direction, and

$$\boldsymbol{\phi}(\boldsymbol{x}) = [\phi_1(x_1, x_2, x_3), \phi_2(x_1, x_2, x_3), \ldots, \phi_{K^3}(x_1, x_2, x_3)]^\top$$

are the continuous basis functions generated by tensor products of univariate splines. Thus K^3 is the total number of the basis functions.

In the factor extraction experiment, we are able to assume the fMRI images to be independent and identically distributed. Denote the covariance function of the functional data

$$G(\boldsymbol{x}, \boldsymbol{s}) = \mathrm{Cov}\{f(\boldsymbol{x}), f(\boldsymbol{s})\}$$

and its sample estimator

$$\widehat{G}(\boldsymbol{x}, \boldsymbol{s}) = N^{-1} \sum_{t=1}^{N} f_t(\boldsymbol{x}) f_t(\boldsymbol{s}). \tag{3.2}$$

The covariance operator V is defined as

$$Vf = \int_{\mathcal{P}_1} \int_{\mathcal{P}_2} \int_{\mathcal{P}_3} G(\cdot, \boldsymbol{x}) f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}.$$

Similarly to the orthogonal decomposition in the multivariate PCA, we have for the 3D image functional data

$$V\xi = \int_{\mathcal{P}_1} \int_{\mathcal{P}_2} \int_{\mathcal{P}_3} G(\cdot, \boldsymbol{x}) \xi(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \lambda \xi(\boldsymbol{x}),$$

where $\xi(\boldsymbol{x})$ and λ denote the eigenfunction on the 3D domain and the eigenvalue respectively. The eigenvalues are real and non-negative $\lambda_1 > \lambda_2 > \cdots \geq 0$. Without spatial information loss or distortion due to vectorization in e.g. FPCA, the first functional factor $\xi_1(x_1, x_2, x_3)$ corresponding to the largest eigenvalue λ_1 accounts for as much of the variability in the data as possible, and each succeeding functional factor $\xi_\ell(x_1, x_2, x_3)$ in turn has the highest variance possible under the constraint that it is uncorrelated with the preceding ones.

Plugging (3.1) into (3.2), we obtain

$$\widehat{G}(\boldsymbol{s}, \boldsymbol{x}) = N^{-1} \boldsymbol{\phi}^\top(\boldsymbol{s}) \boldsymbol{C}^\top \boldsymbol{C} \boldsymbol{\phi}(\boldsymbol{x}),$$

and the orthogonal decomposition equation as

$$\iiint N^{-1} \boldsymbol{\phi}^\top(\boldsymbol{s}) \boldsymbol{C}^\top \boldsymbol{C} \boldsymbol{\phi}(\boldsymbol{x}) \boldsymbol{\phi}^\top(\boldsymbol{x}) \boldsymbol{b} \, \mathrm{d}(\boldsymbol{x}) = \lambda \boldsymbol{\phi}^\top(\boldsymbol{s}) \boldsymbol{b},$$

where the eigenfunction $\xi = \boldsymbol{\phi}^\top \boldsymbol{b}$ with $\boldsymbol{b}$ being a coefficient vector. Define

$$\boldsymbol{W} = \iiint \boldsymbol{\phi}(\boldsymbol{x}) \boldsymbol{\phi}^\top(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}.$$

By solving

$$N^{-1} \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{C}^\top \boldsymbol{C} \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{u} = \lambda \boldsymbol{u}, \tag{3.3}$$

where $\boldsymbol{u} = \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{b}$ and the coefficient vector $\boldsymbol{b}$ satisfies $\boldsymbol{b}_i^\top \boldsymbol{W} \boldsymbol{b}_i = 1$ and $\boldsymbol{b}_i^\top \boldsymbol{W} \boldsymbol{b}_j = 0$, we obtain the eigenfunction that contains spatial information and hence will be used to construct the common spatial factors of the fMRI data.

3.2 Multilinear model

To obtain common spatial factors across subjects, we borrow the idea of panel data analysis by averaging fMRI signals over subjects at each time t:

$$\bar{Y}_t(x_1, x_2, x_3) = J^{-1} \sum_{j=1}^{J} Y_t^{(j)}(x_1, x_2, x_3), \quad t = 1, \ldots, N.$$

The averaged signals are smoothed over a tensor B-spline regression with $K = 16$ knots in every spatial direction. The eigenfunctions are obtained by the 3DIF orthogonal decomposition in Section 3.1.

The eigenfunctions consist of not only important regions attributed to risk perception and investment decisions but also other neural activities unrelated to the investigated stimuli and possible magnetic noises. To remove the impact of noises, the spatial factors are constructed by trimming the eigenfunctions at extreme quantiles such as [0.05 %, 99.95 %] levels and replacing the "non-active" voxels with zeros. Moreover, we only consider the first L eigenfunctions and denote the trimmed factors as common risk related regions, denoted as $\widehat{\xi}_\ell(x_1, x_2, x_3)$ with $\ell = 1, \ldots, L$, since only the first spatial factors are fundamental and necessary. By doing this, the original high dimensionality is reduced to a small number of common spatial factors.

Heterogeneity of individual risk attitude are extracted in the multilinear regression that projects the raw fMRI signals on the common spatial regions:

$$Y_t^{(j)}(x_1, x_2, x_3) = \sum_{\ell=1}^{L} Z_{\ell,t}^{(j)} \widehat{\xi}_\ell(x_1, x_2, x_3) + \varepsilon_t^{(j)}(x_1, x_2, x_3), \tag{3.4}$$

where $\varepsilon_t^{(j)}(x_1, x_2, x_3)$ denotes the idiosyncratic noise of the j-th subject, which is independently and identically distributed with zero mean and constant variance. The subject-specific factor loadings $Z_{\ell,t}^{(j)}$ are calculated by ordinary least squares regression at time t for subject j:

$$\min_{Z_{\ell,t}^{(j)}} \sum_{x_1, x_2, x_3} \left\{ Y_t^{(j)}(x_1, x_2, x_3) - \sum_{\ell=1}^{L} Z_{\ell,t}^{(j)} \widehat{\xi}_l(x_1, x_2, x_3) \right\}^2.$$

The multi-subject 3DIF estimation procedure can now be summarized as follows:

(1) Take the average $\bar{Y}_t(x_1, x_2, x_3)$ of the raw 3D fMRI data across all subjects and obtain the smoothed 3D image functional data $f_t(x_1, x_2, x_3)$.
(2) Perform 3DIF to construct common spatial functional factors $\widehat{\xi}_\ell(x_1, x_2, x_3)$ via (3.3) and trim out insignificant active regions at e.g. 0.05 %– and 99.95 %+ quantiles.
(3) For every subject, estimate the subject-specific factor loadings $Z_{\ell,t}^{(j)}$ with the multilinear regression (3.4) that will be further used to classify risk attitude of the subject.

4 Simulation

Before implementing the proposed 3DIF method to real data, we perform a simulation study to investigate its performance under known data generating processes. Our primary interest is to see how much the 3DIF method will improve the detection accuracy of the risk related brain regions compared to the alternative 1-dimensional functional approach. Moreover, we study how robust is the region detection with respect to the size of the risk activation brain regions.

Our simulation studies are designed to properly reflect real data at hand. The fMRI signals are generated for a "brain" defined in the dimensions of $[1, 91] \times [9, 100] \times [11, 81]$. In previous literature five regions including PC, VLPFC, lOFC, aINS and DLPFC have been identified to be active under risk related tasks. In the first simulation study, we consider five regions that are contained in the literature documented places and specify each of them to a $3 \times 3 \times 3$ cube for a simple demonstration. In particular, PC is defined at location $[51, 53] \times [25, 27] \times [60, 62]$, VLPFC at $[27, 29] \times [89, 91] \times [38, 40]$, lOFC at $[54, 56] \times [97, 99] \times [30, 32]$,

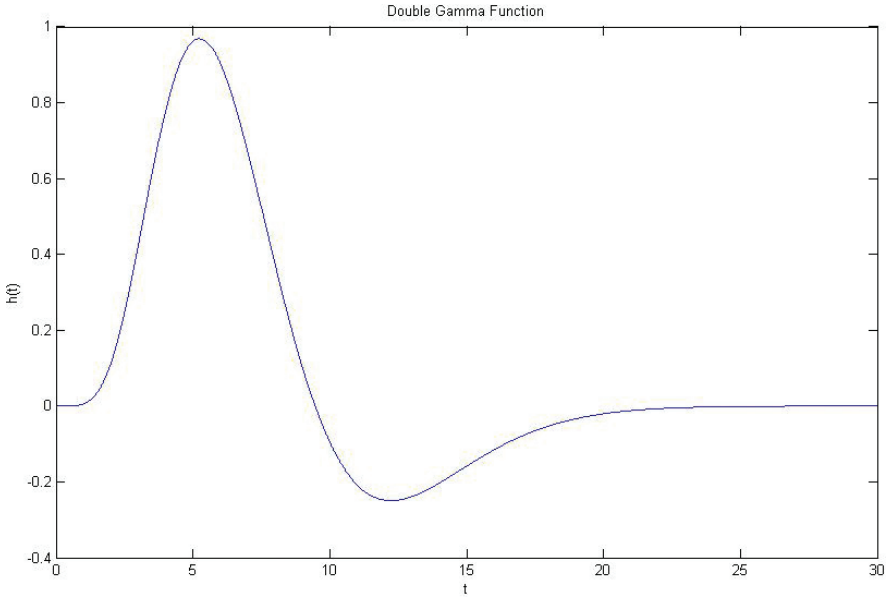


Figure 2: Visualization of the double gamma function.

aINS at $[63, 65] \times [75, 77] \times [37, 39]$, and DLPFC at $[66, 68] \times [77, 79] \times [53, 55]$. The regions are constant in the data generation.

Two kinds of factor loadings are considered: Gaussian distributed random loadings, and a more realistic situation by incorporating the haemodynamic response function (HRF) in the random loadings. The HRF is generated by a double gamma function (see [12, 15, 19, 53]):

$$h(t) = \left(\frac{t}{a_1 b_1}\right)^{a_1} e^{-\frac{t-a_1 b_1}{b_1}} - c\left(\frac{t}{a_2 b_2}\right)^{a_2} e^{-\frac{t-a_2 b_2}{b_2}},$$

where $a_1 = 6$, $a_2 = 12$, $b_1 = b_2 = 0.9$ and $c = 0.35$. Compared to the pure random factor loadings, the HRF scenario mimics the working process of the fMRI scanners, where HRF triggers brain activities. Figure 2 illustrates how the double gamma function reflects the haemodynamic response function (HRF).

Figure 3 gives an illustration of one simulated convolution of double gamma function and the generated factor loadings with HRF.

The 3D image signals are generated to represent brain signals recorded by the fMRI scanner during an RPID experiment:

$$f_t^{(\mathrm{NFL})}(x_1, x_2, x_3) = \sum_{\ell=1}^{5} Z_{\ell t}\xi_\ell(x_1, x_2, x_3) + \varepsilon_t(x_1, x_2, x_3),$$

$$f_t^{(\mathrm{HRF})}(x_1, x_2, x_3) = \sum_{\ell=1}^{5} \{Z_{\ell t} + h(t)\}\xi_\ell(x_1, x_2, x_3) + \varepsilon_t(x_1, x_2, x_3),$$

where NFL refers to the scenario with only normal random factor loadings, while HRF incorporates the impact of HRF in the fMRI signals. The five functional factors $\xi_\ell(x_1, x_2, x_3)$ have been defined in the locations (x_1, x_2, x_3) as mentioned before and are constant over time. The factor loading $Z_{\ell t}$ corresponds to the ℓ-th spatial factor at time point $t = 1, \ldots, 1\,000$. In both the NFL and HRF scenario, the factor loadings are Gaussian distributed with mean zero and standard deviations of 7.6, 5.8, 5.2, 1.8, and 1.7 respectively learned from the real data. The random noise $\varepsilon_t(x_1, x_2, x_3)$ is standard normal distributed and independent from each other. Each generation is repeated 100 times.

We implement two methods to identify the common spatial factors: 3DIF and FSVD proposed by [54]. Both methods handle continuous functional data, however 3DIF directly analyze the fMRI signals in 3D space while FSVD is only applicable for 1D functional data though the latter employs the singular value decomposition (SVD) approach to achieve better estimation feasibility and accuracy. In the simulation study, we chose

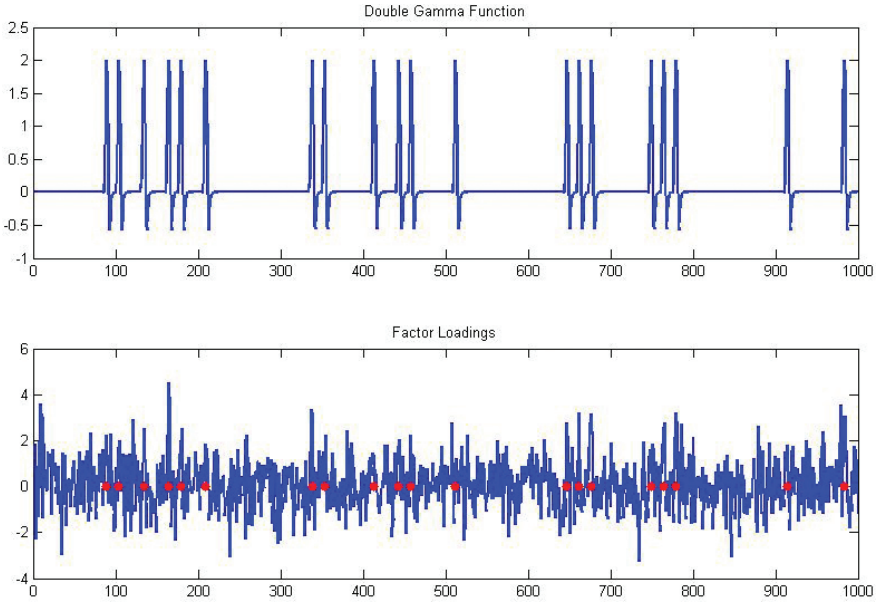


Figure 3: Simulated factor loadings. On top is the double gamma function. The bottom is the simulated factor loadings, which are the sum of the double gamma function and the normal random loadings. The red dots highlight time points when the stimulus are triggered.

$K = 16$ in each direction leading to $K^3 = 4096$ basis functions to utilize the largest computational power for each direction. It is worth noting that the designed risk related regions are only used in the fMRI data generation and will not be utilized in the following decomposition and factor computation. Instead, they are retained to evaluate the detection accuracy. In both methods, the active regions are defined as the trimmed spatial functional factors over the 99.999 % quantile and below the 0.001 % quantile.

As an illustration, Figure 4 displays one active region lOFC associated with evaluating and contrasting different option choices [45]. From top to bottom, one observes the generated (true) region, the identified regions by the 3DIF method and the FSVD approach. The active regions are highlighted as bright areas. Both methods detect the region, however 3DIF performs better in several aspects. In the NLF case, 3DIF explains more variation for the fMRI signals than FSVD, i.e. 56.3 % against 55.2 %, see Table 1. The variance explained increases when the number of factor increases. Moreover, 3DIF provides more clear-cut results, i.e. if the identified spacial factor corresponds to only one actual region, and simultaneously has less mis-detection, i.e. by wrongly identifying non-active regions. See Table 2 for the average percentage of the true regions detected by each estimated functional factor. More than 60 % of the estimated functional factors correspond to exactly one region in 3DIF. The value drops to 43.33 % in FSVD. As for mis-detection, 3DIF mistakenly detects 28 % and FSVD has more at 36.83 %. More importantly, 3DIF provides contiguous regions instead of discrete voxels thanks to its mathematical properties, see the contour plot of lOFC in Figure 5. On the other hand, FSVD identifies discrete voxels, due to the adoption of SVD in the discrete space, which improves estimation efficiency but at cost of contiguity. The relative good performance applies to the HRF scenario, too. While 3DIF explains 69.5 % variation, FSVD reaches to 55.9 %. When using 3DIF, 70 % of the detected risk regions correspond to exactly one active region, 23.33 % are mis-detected and less than 7 % are mixture of risk regions. The alternative FSVD method has only 54 % of one-to-one match, more than 30 % mis-detection and 15 % of mixture. Again, 3DIF accurately and reasonably detects a contiguous region, while the FSVD gives discrete voxels.

Now we repeat the above two experiments with different designs on the active regions to investigate the robustness of 3DIF. In particular, the five active regions are generated with varying sizes to reflect a more realistic situation. Following the study of [33] on the size of identified brain regions, our spatial moderate assumptions state that the spatial factors are active at location $[51, 54] \times [25, 28] \times [60, 63]$ for Parietal Cortex (64 voxels), $[27, 29] \times [88, 91] \times [38, 41]$ for VLPFC (48 voxels), $[52, 59] \times [92, 99] \times [28, 35]$ for lOFC (512 voxels), $[62, 66] \times [74, 78] \times [37, 39]$ for aINS (75 voxels), and $[64, 70] \times [73, 79] \times [51, 57]$ for

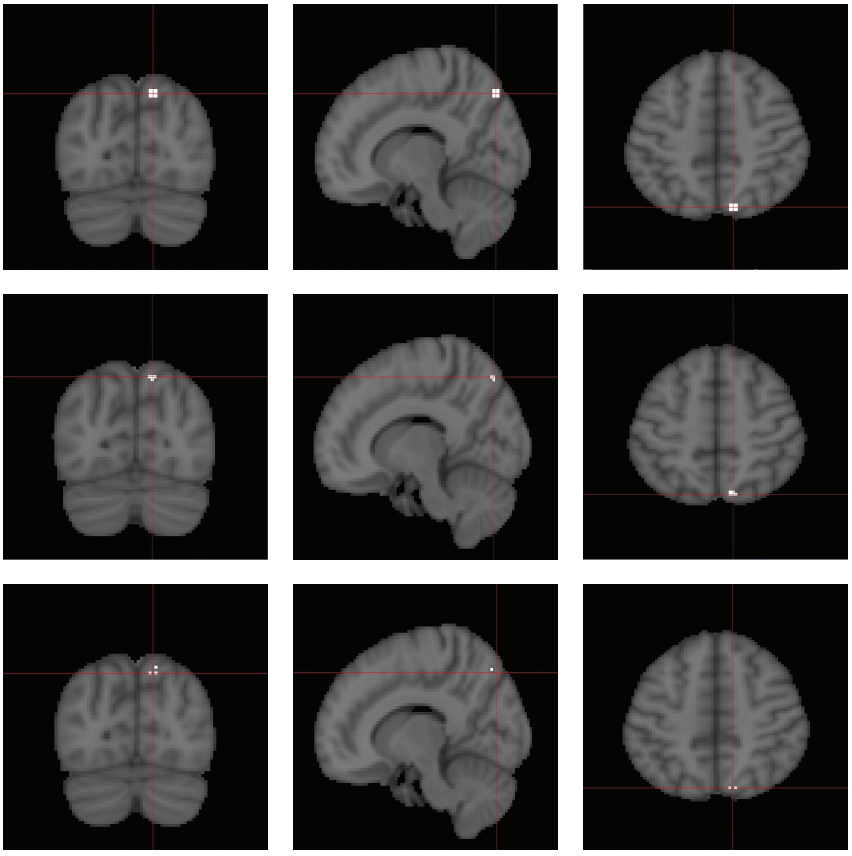


Figure 4: Functional factors on lOFC. From top to bottom are the generated (true) region, the estimated region with 3DIF and the estimated region with the FSVD method.

	Factor						
	1	**2**	**3**	**4**	**5**	**6**	**Total**
NFL: 3DIF	24.2 %	4.5 %	4.2 %	9.9 %	1.7 %	11.7 %	56.3 %
NFL: FSVD	19.2 %	0.7 %	1.6 %	21.5 %	4.8 %	7.4 %	55.2 %
HRF: 3DIF	25.9 %	4.9 %	7.0 %	16.2 %	5.7 %	9.8 %	69.5 %
HRF: FSVD	20.5 %	2.2 %	3.3 %	17.8 %	1.2 %	10.7 %	55.9 %

Table 1: Variance explained by different number of spatial factors for NFL with Gaussian random factor loadings and HRF incorporating HRF in the factor loadings. Two methods have been implemented: 3DIF and FSVD.

	Regions			
	0	**1**	**2**	**≥ 3**
NFL: 3DIF	28.00 %	60.67 %	11.33 %	0.00 %
NFL: FSVD	36.83 %	43.33 %	19.50 %	0.33 %
HRF: 3DIF	23.33 %	70.00 %	6.67 %	0.00 %
HRF: FSVD	31.33 %	54.00 %	14.67 %	0.00 %

Table 2: Average percentage of the estimated functional factors that detect the true regions; "0 region" means no active region and hence a nonzero values indicates mis-detection.

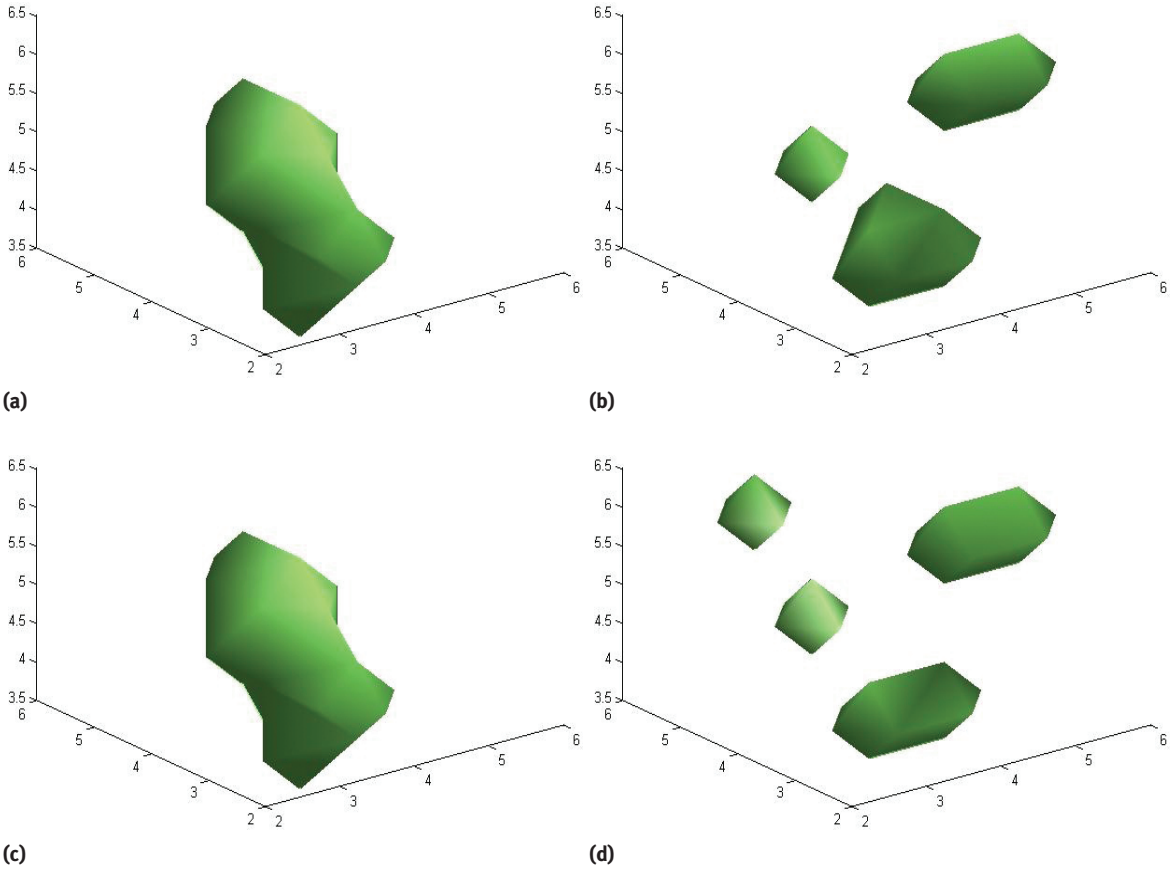


Figure 5: Contour plot of the estimated active region lOFC in NFL (top) and HRF (bottom) cases. On the left is the estimated region with 3DIF and on the right is the estimated region with FSVD.

	Regions			
	0	**1**	**2**	**≥ 3**
NFL: 3DIF	27.00 %	62.67 %	10.33 %	0.00 %
NFL: FSVD	32.17 %	52.33 %	15.50 %	0.00 %
HRF: 3DIF	18.50 %	79.67 %	1.83 %	0.00 %
HRF: FSVD	27.67 %	61.33 %	11.00 %	0.00 %

Table 3: Robust: average percentage of the estimated functional factors that detect the true regions; "0 region" means no active region and hence a nonzero values indicates mis-detection.

DLPFC (343 voxels). The factor loadings and the noise level remain the same as in the previous experiments. Both normal and HRF factor loadings are considered. Each data generation is repeated 100 times.

We still implement the 3DIF and FSVD methods to the generated fMRI data. As the average number of voxels now is about eight times of that in the previous simulations, the active regions are trimmed at extreme quantiles. Results evidence a stable performance. Again, 3DIF provides better identification, see Table 3 for the average percentage of the true regions detected by each estimated factor. In the NFL case, 62.67 % of the estimated functional factors are associated with exactly one region, 27 % are mis-detected and 10.33 % are mixed. On the contrary, the alternative method performs worse with less one-to-one match at 52.33 %, more mis-detection at 32.17 % and mixture at 15.5 %. In the HRF case, 3DIF still outperforms the alternative with 79.67 % one-to-one match, 18.50 % mis-detection and 1.83 % mixture, compared to 61.33 %, 27.67 % and 11.00 % by FSVD. Similarly, the 3DIF method provides realistic contiguous regions, while the alternative FSVD detects discrete voxels, see Figure 6 for the contour plot of the risk region lOFC as illustration.

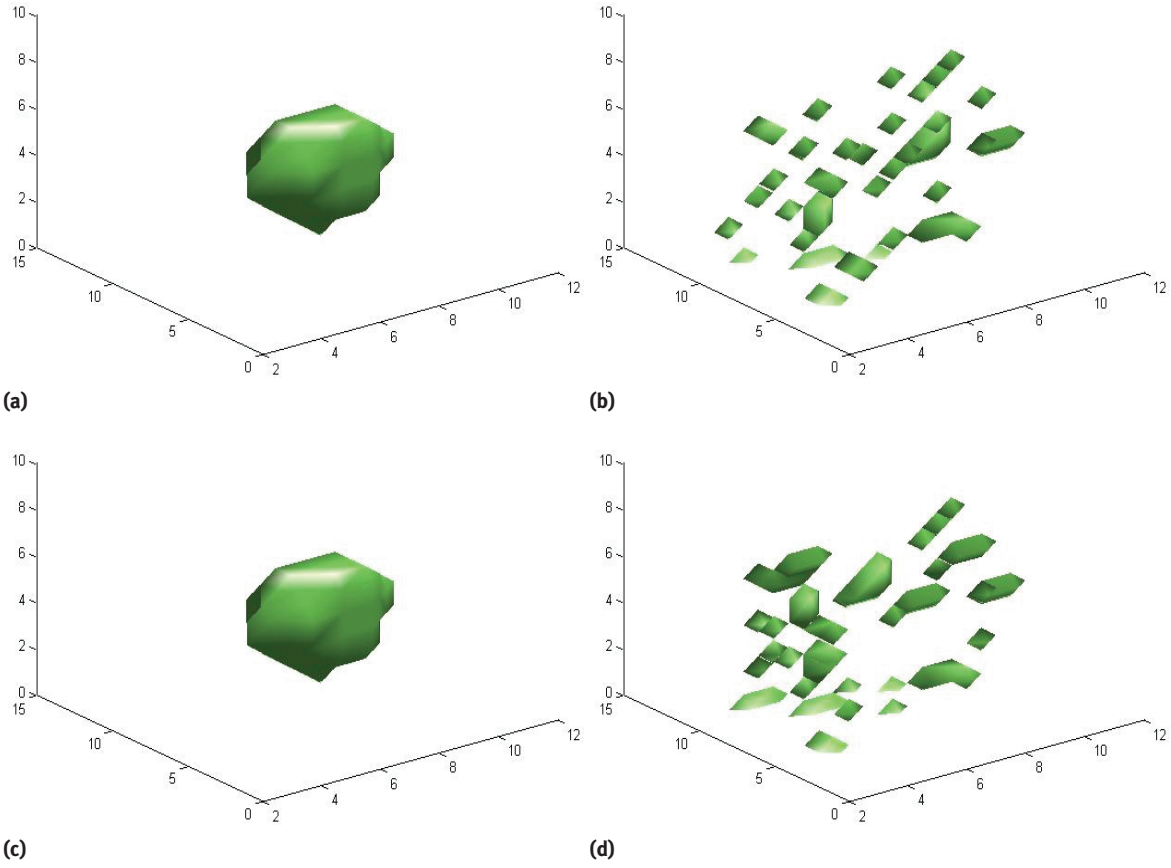


Figure 6: Robust: contour plot of the active region on lOFC. The left column is the estimated region in 3DIF and the right column is the estimated region with FSVD method. The top row is the result for NFL with normal factor loadings and on the bottom is the result for HRF with HRF incorporated in factor loadings.

The simulation study shows that the proposed 3DIF outperforms the alternative functional approach, with better quality of risk related regions detected. The relative good performance is stable for different scenarios with various parameters.

5 Empirical results

We implement the proposed 3DIF method to the fMRI signals data collected in the RPID experiment as described in Section 2, which mimics real-life investment decisions by providing subjects with return streams of investments. We assume that all subjects exhibit homogenous brain structure. In other words, the spatial maps are common for all, while the individual differences are represented by the subject specific scores. We report the detected common risk related regions and compare with several alternative methods. We classify subjects' risk perception based on the extracted subject specific signals, i.e. signature scores, and evaluate the risk classification accuracy with the help of psychological risk-return (PRR) model.

5.1 Computational time

The analyzed fMRI data are high dimensional ($91 \times 109 \times 91 \times 1\,360$ scans = 1,227,575,440) and require large memory (17×1.3 GB). The 3DIF method is implemented on twelve cores ProLiant BL680c G7 server equipped with Intel(R) Xeon(R) CPU E7-4860@2.27 GHz processors and 252 GB memory loading. The main

computation time is spent on computing the tensor integral $\boldsymbol{W} = \iiint \boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{\phi}^\top(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$, which exponentially increases in the number of knots K. Though a large number of knots provides better fit, it extends the computational time. Van Bömmel, Song, Majer, Mohr, Heekeren and Härdle [47] choose the basis function with fourteen knots in the x- and y-axis and twelve knots in the z-axis to balance accuracy and computational time. In our study, we increase the number of knots $K = 16$ in each direction leading to $K^3 = 4096$ basis functions, to further improve the estimation accuracy by utilizing larger computational power. The computation of the triple integral $\boldsymbol{W}$ costs 48 hours. It is worth noting that the value of the triple integral only depends on the B-spline basis functions and hence can be used for other fMRI data analysis. With the value of $\boldsymbol{W}$, the computation of 3DIF only needs 4 hours to complete.

5.2 Alternative methods

For comparison, two alternative methods have been implemented on the same data. Mohr, Biele, Krugel, Li and Heekeren [33] conducted the general linear model (**GLM**) with six design factors on the individual fMRI data. Van Bömmel, Song, Majer, Mohr, Heekeren and Härdle [47] proposed a panel version of the dynamic semiparametric factor model (**PDSFM**) to reanalyze the data. See Section 2 for details of their findings.

In addition, we consider three more methods that have previously been proposed in literature. We implement them to analyze the same data, including singular value decomposition (SVD) – a multivariate statistical technique – in a discrete framework, and two functional data analysis methods functional principal component analysis (FPCA) and functional SVD (FSVD) in a continuous but 1-dimensional space.

SVD: Denote the vectorized fMRI signal data as $Y = [Y_1, Y_2, \ldots, Y_N]$ that has $p \times N$ dimensions with $p = 91 \times 109 \times 91$ and $N = 1360$ in our study, SVD decomposes the discrete data averaged over subjects and constructs common spatial factors of risk-related brain regions $Y = \Gamma\Lambda^{\frac{1}{2}}U^\top$, where Γ is a $p \times N$ orthonormal matrix, Λ is a diagonal matrix and U is an $N \times N$ orthogonal matrix. The ℓ-th spatial factor is constructed with the ℓ-th column of Γ. Compared to the classic principal component analysis (PCA), SVD is computationally efficient and feasible with reduced dimensionality, i.e. decomposing a $p \times N$ sample matrix instead of a $p \times p$ covariance matrix given that $p \gg N$, when dealing with high-dimensional data. It however ignores contiguity nature of the fMRI signals, which leads to discontinued active regions.

FPCA and FSVD: The FPCA method estimates eigenfunctions in a functional framework. Similar to the proposed 3DIF method, the vectorized data is smoothed but using 1D basis functions and one performs eigendecomposition for the covariance operator. Denote the covariance operator by V we have $V\xi = \lambda\xi$, where ξ represents the eigenfunction corresponding the eigenvalue λ, see [39, 40]. The FPCA approach, though guarantees the contiguity of risk related brain regions, is subject to the curse of dimensionality. Zipunnikov, Caffo, Yousem, Davatzikos, Schwartz and Crainiceanu [54] proposed FSVD, which implements SVD to the smoothed functional data instead of the discrete raw data to balance the tradeoff between high dimensionality and computational efficiency. Nevertheless, the two functional data analysis methods requests pre-processing vectorization, which may misrepresent the raw spatial structure of the fMRI data.

5.3 Risk related regions $\widehat{\boldsymbol{\xi}}_\ell$

The 3D Image FPCA (3DIF) technique is utilized to capture the fundamental spatial maps under risk decisions. We identify the common spatial factors and use them to represent the brain regions with significant activity during the RPID experiment. One question remains on how to choose the number of spatial factors, denoted by L. The larger the number of spatial factors, the better the in-sample accuracy of the fitted model. On the other hand, too large L leads to over-fitting and poor out-of-sample performance. The selection of the number of factors may rest on the explained variation for different model specification. Table 4 presents the explained variance averaged over the seventeen subjects for different number of factors. It shows that 86 % variation in the data is attributed to the first spatial factor when using 3DIF, which can be interpreted as the typical

	L				
	1	**2**	**4**	**6**	**20**
3DIF	86.03 %	88.93 %	90.05 %	92.78 %	94.34 %
FSVD	96.50 %	96.57 %	96.65 %	96.74 %	97.07 %
FPCA	70.06 %	81.62 %	87.85 %	92.82 %	95.27 %
SVD	96.67 %	96.73 %	96.80 %	96.89 %	97.21 %

Table 4: Explained variance by different number of spatial factors.

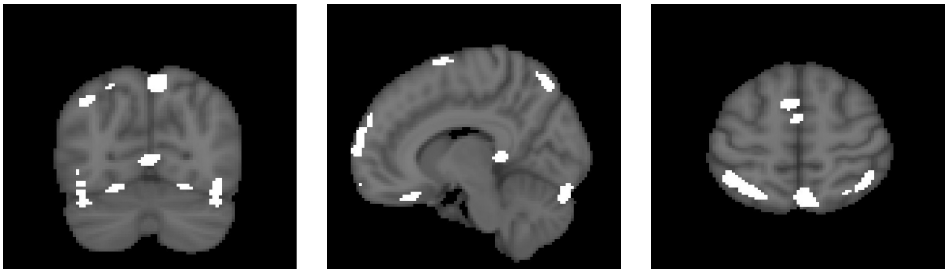
brain activity during the RPID experiment. Alternatively, the dominant component explains 96.50 % variation in FSVD, 70 % in FPCA and 98.67 % in SVD. Though numerically important, the first spatial factor has less psychological meaning and is irrelevant to any important risk related regions documented in literature. On the contrary, the inclusion of subsequent factors allows more useful information captured and simultaneously enables the detection of important risk related regions. For example, aINS is in modest size relative to visual or audial cortex but highly relevant to risk perception and investment decisions. Thus, $L = 20$ is chosen in our study. In this case, 94 % of variation is explained by the 3DIF method, which is lower than the alternatives. However, it is worth mentioning that higher variance is explained by the 3DIF spatial factor associated with important risk related regions. For example, the 3DIF factor for lOFC ($\widehat{\xi}_5$) explains 2.73 % (the difference between 92.78 % for $L = 6$ and 90.05 % for $L = 4$), while FSVD ($\widehat{\xi}_5$) and SVD ($\widehat{\xi}_5$) both contribute 0.09 % and FPCA ($\widehat{\xi}_3$) provide 6.23 %. We will continue the performance comparison of the data-driven methods in the risk classification analysis.

Figure 7 displays the identified risk related brain regions by using the proposed 3DIF method, the alternative 1D functional data analysis methods FSVD and FPCA, and the multivariate technique SVD. All detect the risk related brain regions including parietal cortex (PC), lateral orbifrontal cortex (lOFC) and ventrolateral prefrontal cortex (VLPFC). The three regions have been documented in literature and also by [33] analyzing the same data with GLM. However only the proposed 3DIF method successfully finds anterior insula (aINS) that is associated with value processing, risk and uncertainty. Moreover, the 3DIF method detects the activation of medial orbifrontal cortex (mOFC) as documented in [47] when analyzing the same data using PDSFM. The mOFC has been interpreted to be related to evaluation and contrast of various choices [45]. The FPCA method provides over-smoothed regions, though continuous, due to the extremely high dimensionality larger than 220,000 after vectorization. Table 5 summarizes the region detection for the same data by various methods. The proposed 3DIF method and the GLM [33] both identified five regions, where four of them are consistent. The alternative FSVD, FPCA and SVD found three regions and the PDSFM [47] obtained two.

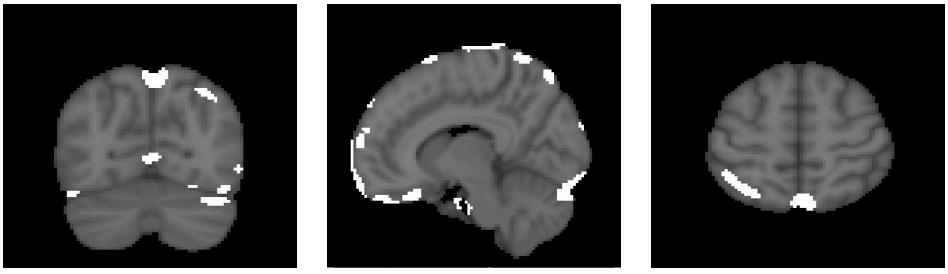
Figure 8 displays details of the detected regions by 3DIF. The relevant spatial factors are $\widehat{\xi}_\ell(x_1, x_2, x_3)$ with $\ell = 3, 4, 5, 12, 18, 19$. In particular, $\widehat{\xi}_3$ and $\widehat{\xi}_{12}$ are located in PC and attributed to risk related processes and selective attention (see [6, 41]); $\widehat{\xi}_4$ is related to the VLPFC region that stands for value processing. The regions mOFC and lOFC picked up by $\widehat{\xi}_5$ that are associated with evaluating and contrasting of different choice options [45]. The aINS region is captured by $\widehat{\xi}_{18}$ and related to risk and uncertainty [18], and the DLPFC area is highlighted by $\widehat{\xi}_{19}$. Figures 9–11 display the detected risk related brain regions by the alternative approaches. The identified regions of lOFC and VLPFC in Figures 9–11are similar due to the nearby coordinates of the regions. The center coordinates of the identified lOFC is (61, 94, 31) and of the VLPFC is (30, 94, 36).

	PC	**VLPFC**	**lOFC**	**aINS**	**DLPFC**	**mOFC**	**MPFC**
3DIF	✓	✓	✓	✓	✓		
GLM	✓	✓		✓	✓		✓
PDSFM	✓					✓	
FSVD	✓	✓	✓				
FPCA	✓	✓	✓				
SVD	✓	✓	✓				

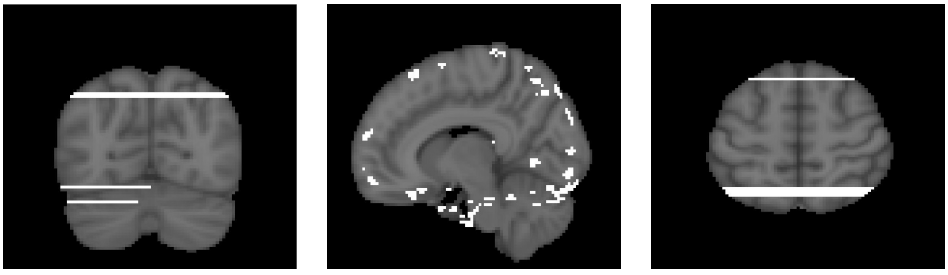
Table 5: Detected risk related brain regions for the same fMRI data of the RPID experiments in [33].



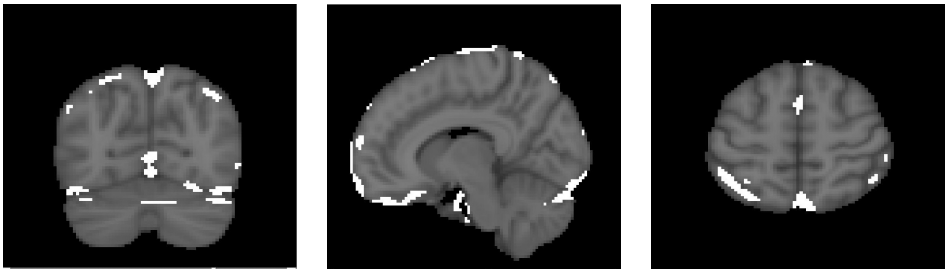
(a) 3DIF.



(b) FSVD.

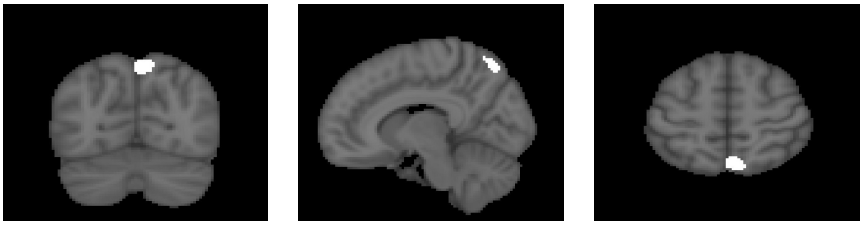


(c) FPCA.

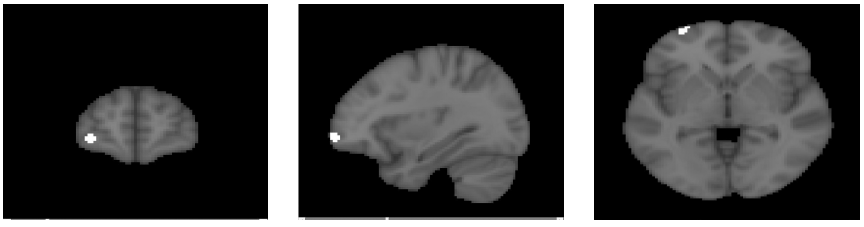


(d) SVD.

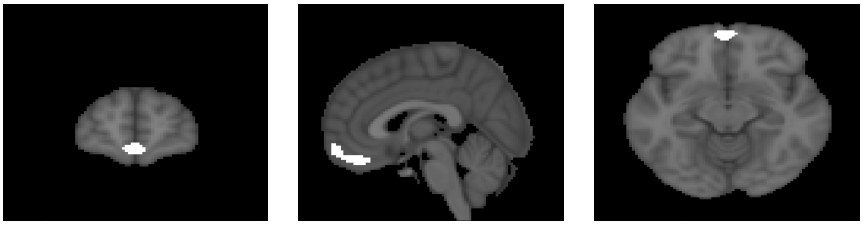
Figure 7: Detected risk-related brain regions by the first twenty eigenfunctions using (a) the 3DIF and alternative methods including (b) FSVD, (c) FPCA and (d) SVD.



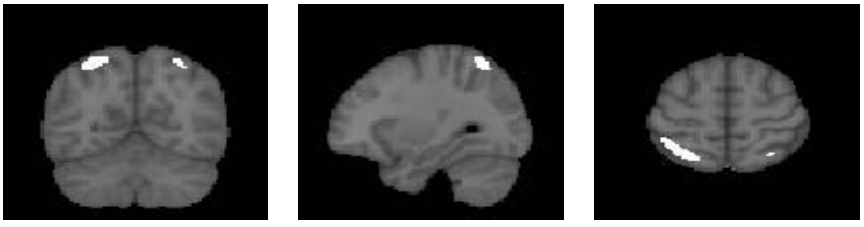
(a) Parietal Cortex.



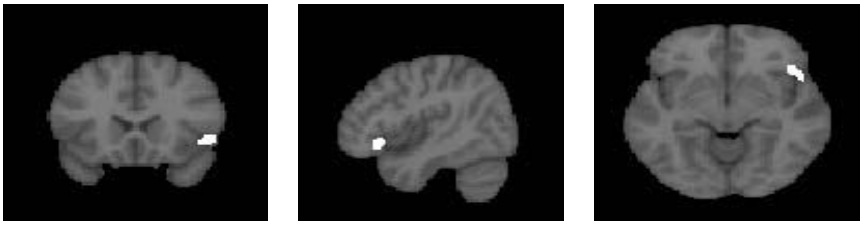
(b) VLPFC.



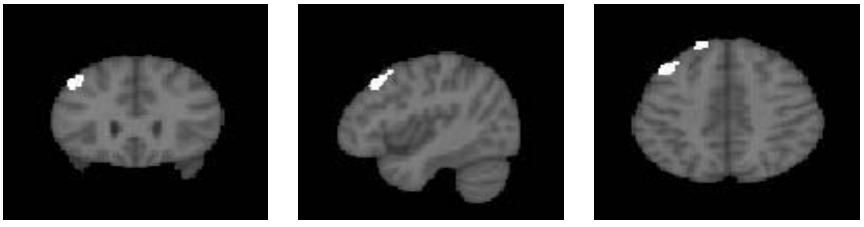
(c) lOFC.



(d) Parietal Cortex.

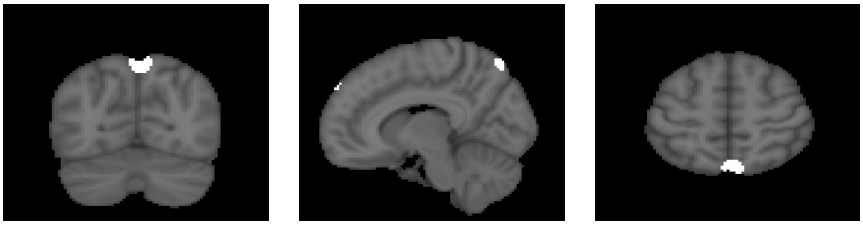


(e) aINS.

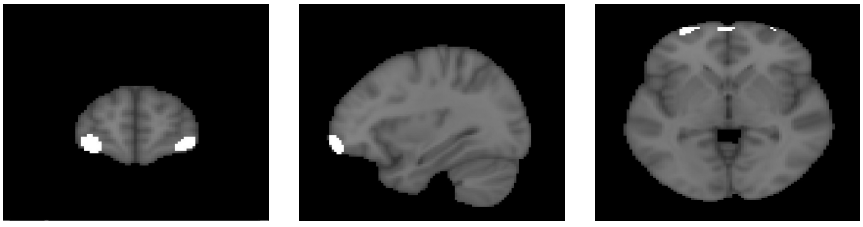


(f) DLPFC

Figure 8: 3DIF: Selected identified risk related regions $\widehat{\xi}_\ell$, $\ell = 3, 4, 5, 12, 18, 19$. (a) Estimated function $\widehat{\xi}_{12}$ in Parietal Cortex; (b) $\widehat{\xi}_4$ in VLPFC; (c) $\widehat{\xi}_5$ in lOFC; (d) $\widehat{\xi}_3$ in Parietal Cortex; (e) $\widehat{\xi}_{18}$ in aINS; (f) $\widehat{\xi}_{19}$ in DLPFC.

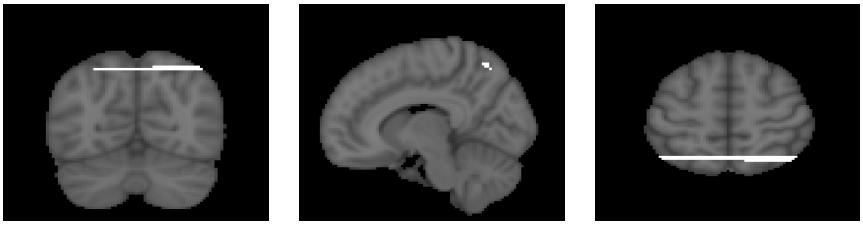


(a) Parietal Cortex.

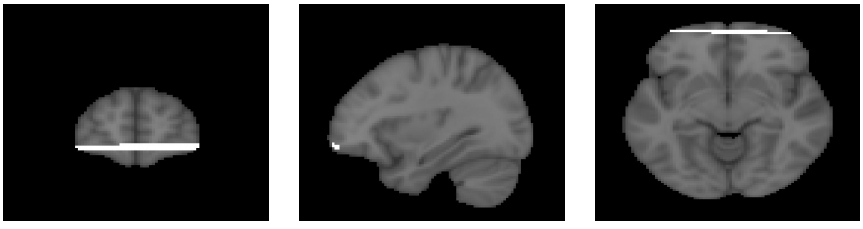


(b) VLPFC & lOFC.

Figure 9: FSVD: Selected identified risk related regions. (a) Estimated function $\widehat{\xi}_{10}$ in Parietal Cortex; (b) $\widehat{\xi}_5$ in VLPFC and lOFC.

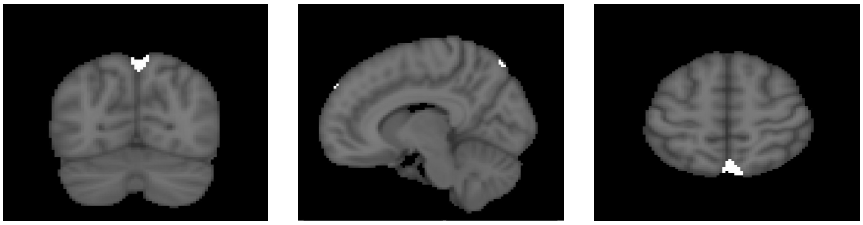


(a) Parietal Cortex.

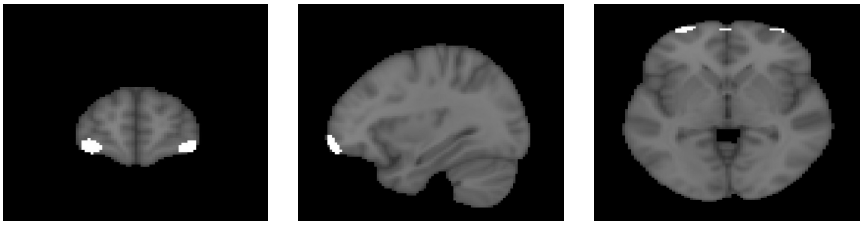


(b) VLPFC & lOFC.

Figure 10: FPCA: Selected identified risk related regions. (a) Estimated function $\widehat{\xi}_2$ in Parietal Cortex; (b) $\widehat{\xi}_3$ in VLPFC and lOFC.



(a) Parietal Cortex.



(b) VLPFC & lOFC.

Figure 11: SVD: Selected identified risk related regions by SVD. (a) Estimated function $\widehat{\xi}_{10}$ in Parietal Cortex; (b) $\widehat{\xi}_5$ in VLPFC and lOFC.

5.4 Subject specific signature scores $Z_{\ell,t}^{(j)}$

The dynamic behaviors of the individual brain activities are represented by the subject specific signature $Z_{\ell,t}^{(j)}$ with $j = 1, \ldots, 17$, $\ell = 1, \ldots, 20$, and $t = 1, \ldots, 1360$. Given the risk related regions common for all subjects, the individual risk perception and attitude during decision making under risk are reflected by the series of the activation. An interesting question is whether the extracted subject specific signature scores properly reflect the risk preference of individual. Among others, for the active brain regions that have been found to be related to risk and uncertainty, the respective signature scores are expected to carry explanatory power for the heterogeneity of individual risk preferences. Understanding those variations requires a careful investigation and is presented in the following risk classification study.

5.4.1 Risk attitudes

Mohr, Biele, Krugel, Li and Heekeren [33] quantify the risk preference of the seventeen subjects in the same experiment with the help of psychological risk-return (PRR) model

$$V_j(x) = \mu_j(x) - \beta_j \sigma_j(x),$$

where $V_j(x)$ is the value of investment x by subject j, $\mu_j(x)$ is the respective expected return, $\sigma_j(x)$ is the perceived risk, and β_j is a subject specific weight coefficient and reflects the risk attitude of subject j. Given the displayed streams of returns in the RPID experiment and the subjects' answers to the two tasks, i.e. subjective expected return and perceived risk, the risk weight β_j is estimated in a logistic regression framework. In total, seven subjects ($j = 2, 5, 6, 8, 10, 11, 17$) are categorized as weakly risk averse with the risk weight $\beta_j < 5$, and the remaining ten subjects are classified as strongly risk averse, with higher risk attitudes. The dichotomization and derived risk attitudes β_j are presented in Figure 12.

5.4.2 Risk classification

The aim of risk classification analysis is to investigate the possible relation between neural processes underlying investment decisions and subjects' risk preferences. A classification method is proposed to predict

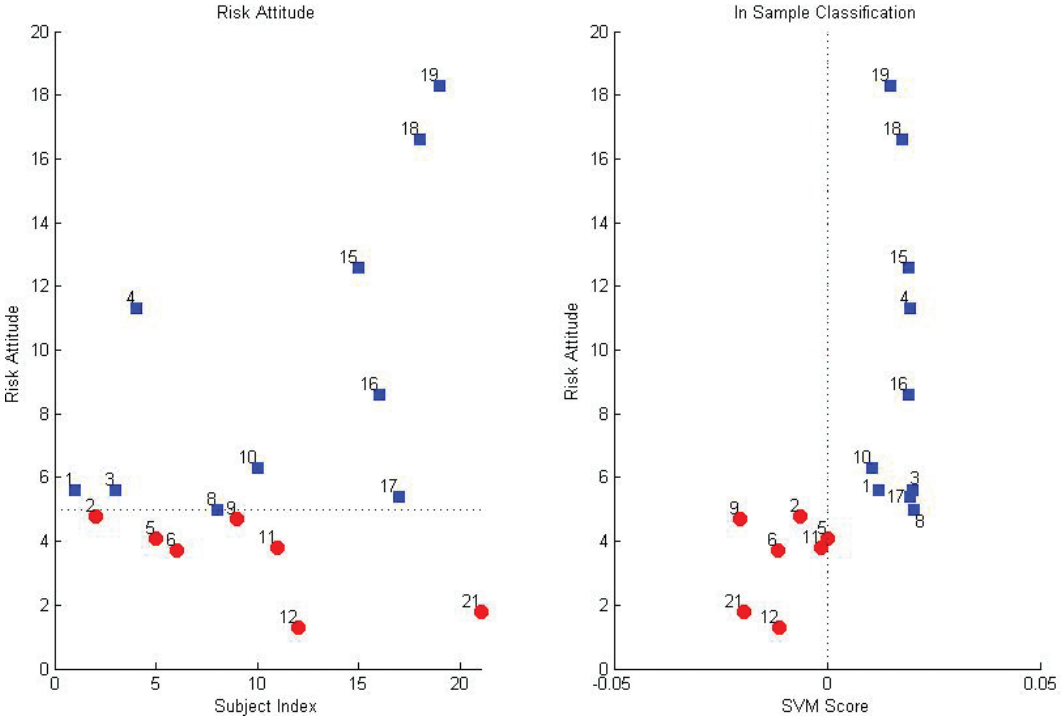


Figure 12: Risk attitudes and SVM scores of seventeen subjects. Subjects with risk attitude ≤ 5 are marked as red circles, otherwise as blue squares.

individual's risk attitude without any information on his or her decision behavior. Instead, the classification is performed solely on the extracted signature scores. The RPID consists of three types of tasks, we here only utilize the decision task, where subject chooses between risky investment return or sure fixed 5 % return, and thus his risk attitude contributes to the perceived value of the displayed return streams and plays a key role in the decision process. The other two tasks, i.e. subjective expected return and perceived risk, have been employed in the PRR model to provide a benchmark and will be used to verify the classification accuracy. Moreover, the analysis is performed for each subject based on six signature scores $Z^{(j)}_{\ell,t}$, $\ell = 3, 4, 5, 12, 18, 19$, of the active brain regions that have been found to be related to risk and uncertainty.

Each subject was exposed to 27 decision tasks and had to make a choice within the next 7 seconds in the RPID experiment. To investigate the brain reactions to the investment decision task of different groups being strongly/weakly risk averse, three consequent observations after the s-th stimulus at scan t_s are considered, covering the decision making period over 7.5 seconds. The three signature scores are demeaned by the score at the stimulus time point Z^{j}_{ℓ,t_s} to capture the peak of the HRF. We compute the average to stand for the average reaction to stimulus s

$$\overline{\Delta}\widehat{Z}^{(j)}_{\ell,t_s} = \frac{1}{3}\sum_{\tau=1}^{3}\widehat{Z}^{(j)}_{\ell,,t_s+\tau} - \widehat{Z}^{(j)}_{\ell,t_s}$$

and the standard deviation of the 27 average reactions as empirical characteristics of subject's risk preference. For each subject, six standard deviations are obtained and will be used in the risk classification analysis. For the alternative FSVD, FPCA and SVD methods, similar procedures are applied to extract the variables for risk classification.

Classification analysis is performed via support vector machines (SVM), see [7, 17]. Subjects are classified based on their six standard deviations of the average reactions to decision task. For the learning part, the strongly risk averse subjects are denoted with 1 and the weakly risk averse subjects with −1. The classification performance is validated by the estimated risk attitudes, see Section 5.4.1.

We first evaluate the in-sample predictive power of the 3DIF method on risk preferences. Figure 12 shows that the seventeen subjects were perfectly classified, with 100 % correction for both strongly and weakly risk

	Overall				Strong				Weak			
k	3DIF	SVD	FSVD	FPCA	3DIF	SVD	FSVD	FPCA	3DIF	SVD	FSVD	FPCA
1	**88 %**	76 %	76 %	76 %	**100 %**	100 %	100 %	90 %	**71 %**	43 %	43 %	57 %
2	**82 %**	76 %	76 %	76 %	**100 %**	100 %	100 %	89 %	55 %	43 %	43 %	**56 %**
3	**79 %**	75 %	75 %	73 %	98 %	**99 %**	**99 %**	87 %	53 %	42 %	42 %	**54 %**
4	**77 %**	74 %	73 %	72 %	95 %	**98 %**	95 %	85 %	51 %	39 %	41 %	**52 %**
5	**74 %**	71 %	70 %	69 %	92 %	**95 %**	91 %	83 %	**50 %**	37 %	39 %	49 %
6	**73 %**	67 %	66 %	66 %	**90 %**	90 %	86 %	81 %	**49 %**	35 %	37 %	46 %

Table 6: SVM classification rate in percentage points by leave-*k*-out for the 3DIF, SVD, FSVD and FPCA methods. The overall refers to the classification rates of all subjects, while strong and weak refer to the classification rates of strongly risk averse subjects and weakly risk averse subjects respectively.

averse groups. The in-sample classification however by utilizing all the information of subjects may involve over-fitting problem. We thus employ the leave-*k*-out cross validation and continue out-of-sample prediction. Samples are iteratively partitioned to two subsets, i.e. training with $N - k$ subjects and validation with k subjects. The prediction for validation is repeatedly performed based on different training sets. The accuracy measurements are averaged among all the predictions. The algorithm can be formulated as follows:

(1) Divide subjects into training set with $N - k$ people and test set with size of k.
(2) Apply the leave-*k*-out cross validation and find the optimal SVM parameters.
(3) Classify the test data.
(4) Repeat (1)–(3) for all different test sets.

Table 6 reports the classification rate (in percentage) by leave-*k*-out cross validation for $k = 1, \ldots, 6$. The classification rate is relatively stable, though it reduces slowly as k increases. The 3DIF method provides consistently better "overall classification" rate than the alternatives, with 73 %–88 % correction using the optimal SVM parameters. The classification accuracy is remarkably improved for the strongly risk averse subjects. The 3DIF and SVD methods are superior in terms of classification accuracy at 90 %–100 %, while 3DIF and FPCA perform better for weakly risk averse individuals at 49 %–71 %. As a comparison, van Bömmel, Song, Majer, Mohr, Heekeren and Härdle [47] have implemented leave-one-out procedure, i.e. $k = 1$, and reached 97 % for strongly risk-averse individuals and 75 % for weakly risk-averse individual. In summary, the analysis implies that the signature scores of the selected risk related regions carry explanatory power for subjects' risk attitudes derived from their choice in the RPID experiment. The risk preferences can be classified by the volatility (standard deviation) of the signature signals with an considerable accuracy. The proposed 3DIF method has consistent reasonable classification power compared to the alternatives.

6 Conclusion

Understanding how people make decisions among risky choices has attracted much attention of researchers in economics, psychology, and neuroscience. While economists evaluate individual's risk preference through mathematical modeling, neuroscientists answer the question by exploring the neural activities in brain. The existing literature has documented the brain regions of PCC, lOFC, mOFC, VLPFC, VMPFC and aNIS to be associated with decision making process under risk. Our study implements a model-free method to further investigate the links between active risk related brain region detection and individual's risk preference.

The proposed 3D Image FPCA (3DIF) methodology is directly applicable to the 3D image data. It avoids spatial information distortion during artificial vectorization or mapping and simultaneously analyzes brain data in the continuous functional domain. Thus, the anatomical brain structure is preserved and efficiently embraced in the estimation procedure. Moreover, it guarantees the contiguity of brain regions rather than discrete voxels. The 3DIF decomposes the fMRI BOLD signals into spatial factors, representing the common spatial maps for all subjects, and the heterogeneity of individual risk preference is explained by subject spe-

cific signature scores. The spatial factors capture the brain regions with the highest variability throughout experiment and consequently represent the activation pattern with a reduced number of factors. The representation precision is controlled by the number of factors L and even subtle effects can be detected. The signature scores mimic activation patterns on subject's risk attitude and correspond to the neural activity of a particular region of interest. As a result, the 3DIF addresses the key limitations of the GLM and the other conventional model-free methods such as PDSFM, FSVD, FPCA and SVD.

The performance is evidenced by our extensive simulation study, where in different setups, region detections and modeling performance were reasonably achieved. Furthermore, our technique outperforms the alternative competitor as the preservation of the spatial brain structure really pays off. In real data analysis, 3DIF detected five risk related regions, which is consistent to the study in [33]. The alternative methods on the other hand only identified limited risk related regions.

Investment decision may be described as a process of evaluating and contrasting of various choices with uncertain outcomes. In this framework the risk preferences are the crucial factor which affects the subjective value of investment. To improve our understanding of the underlying neural activities, we provided the statistical analysis of the extracted signature scores selected in the decision making context. The focus is on the variability in the HRF after the decision stimulus, captured by the score series. The standard deviations derived from the subject-specific responses served as an input in the SVM classifier. We perform both in-sample and out-of-sample risk classifications. In addition to perfect correction for in-sample, the 3DIF provides nice and stable performance for out-of-sample with leave-k-out cross validation, with the best overall classification rate at 73 %–88 %, the 90 %–100 % for strongly risk averse and 49 %–71 % for weakly risk averse. One can conclude that the 3DIF method exhibits better explanatory power for subjects' risk preferences than the alternatives.

Funding: This research was supported by the FRC grant and IDS grant at the National University of Singapore. The authors also acknowledge the support of the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk" and the International Research Training Group (IRTG) 1792 "High-Dimensional Non-Stationary Time Series".

References

[1] C. Amiez, J.-P. Joseph and E. Procyk, Reward encoding in the monkey anterior cingulate cortex, *Cerebral Cortex* **16** (2006), no. 7, 1040–1055.
[2] A. H. Andersen, D. M. Gash and M. J. Avison, Principal component analysis of the dynamic response measured by fMRI: A generalized linear systems framework, *Magn. Resonance Imag.* **17** (1999), no. 6, 795–815.
[3] R. B. Barsky, M. S. Kimball, F. T. Juster and M. D. Shapiro, Preference parameters and behavioral heterogeneity: An experimental approach in the health and retirement survey, Technical report, National Bureau of Economic Research, 1995.
[4] R. Baumgartner, L. Ryner, W. Richter, R. Summers, M. Jarmasz and R. Somorjai, Comparison of two exploratory data analysis methods for fMRI: Fuzzy clustering vs. principal component analysis, *Magn. Resonance Imag.* **18** (2000), no. 1, 89–94.
[5] R. M. W. J. Beetsma and P. C. Schotman, Measuring risk attitudes in a natural experiment: Data from the television game show lingo, *Econom. J.* **111** (2001), no. 474, 821–848.
[6] M. Behrmann, J. J. Geng and S. Shomstein, Parietal cortex and attention, *Current Opinion Neurobiol.* **14** (2004), 212–217.
[7] C. Cortes and V. Vapnik, The nature of statistical learning theory, *Mach. Learn.* **20** (2005), 273–297.
[8] H. D. Critchley, R. N. Melmed, E. Featherstone, C. J. Mathias and R. J. Dolan, Brain activity during biofeedback relaxation, *Brain* **124** (2001), no. 5, 1003–1012.
[9] D. Degras and M. A. Lindquist, A hierarchical model for simultaneous detection and estimation in multi-subject fMRI studies, *NeuroImage* **98** (2014), 61–72.
[10] D. Fetherstonhaugh, P. Slovic, S. Johnson and J. Friedrich, Insensitivity to the value of human life: A study of psychophysical numbing, *J. Risk Uncertainty* **14** (1997), no. 3, 283–300.
[11] K. J. Friston, A. P. Holmes, K. J. Worsley, J. B. Poline, C. Frith and R. S. J. Frackowiak, Statistical parametric maps in functional imaging: A general linear approach, *Human Brain Mapping* **2** (1995), 189–210.
[12] G. H. Glover, Deconvolution of impulse response in event-related bold fMRI, *Neuroimage* **9** (1999), no. 4, 416–429.

[13] G. H. Golub and C. Reinsch, Handbook Series Linear Algebra: Singular value decomposition and least squares solutions, *Numer. Math.* **14** (1970), no. 5, 403–420.
[14] J. Grinband, J. Hirsch and V. P. Ferrera, A neural representation of categorization uncertainty in the human brain, *Neuron* **49** (2006), no. 5, 757–763.
[15] J. Grinband, T. D. Wager, M. Lindquist, V. P. Ferrera and J. Hirsch, Detection of time-varying signals in event-related fMRI designs, *Neuroimage* **43** (2008), no. 3, 509–520.
[16] D. J. Hand and W. E. Henley, Statistical classification methods in consumer credit scoring: A review, *J. Roy. Statist. Soc. Ser. A* **160** (1997), no. 3, 523–541.
[17] W. K. Härdle and L. Simar, *Applied Multivariate Statistical Analysis*, 4th ed., Springer, Heidelberg, 2015.
[18] H. R. Heekeren, S. Marrett and L. G. Ungerleider, The neural systems that mediate human perceptual decision making, *Nat. Rev. Neurosci.* **9** (2008), 467–479.
[19] R. Heller, D. Stanley, D. Yekutieli, N. Rubin and Y. Benjamini, Cluster-based analysis of FMRI data, *NeuroImage* **33** (2006), no. 2, 599–608.
[20] S. A. Huettel, A. W. Song and G. McCarthy, Decisions under uncertainty: Probabilistic context influences activation of prefrontal and parietal cortices, *J. Neurosci.* **25** (2005), no. 13, 3304–3311.
[21] S. A. Huettel, C. J. Stowe, E. M. Gordon, B. T. Warner and M. L. Platt, Neural signatures of economic preferences for risk and ambiguity, *Neuron* **49** (2006), no. 5, 765–775.
[22] J. W. Kable and P. W. Glimcher, The neural correlates of subjective value during intertemporal choice, *Nature Neurosci.* **10** (2007), no. 12, 1625–1633.
[23] D. Kahneman and A. Tversky, Prospect theory: An analysis of decision under risk, *Econometrica* **47** (1979), 263–291.
[24] S. W. Kennerley, A. F. Dahmubed, A. H. Lara and J. D. Wallis, Neurons in the frontal lobe encode the value of multiple decision variables, *J. Cognitive Neurosci.* **21** (2009), no. 6, 1162–1178.
[25] B. Knutson, J. Taylor, M. Kaufman, R. Peterson and G. Glover, Distributed neural representation of expected value, *J. Neurosci.* **25** (2005), no. 19, 4806–4812.
[26] C. M. Kuhnen and B. Knutson, The neural basis of financial risk taking, *Neuron* **47** (2005), no. 5, 763–770.
[27] S.-H. Lai and M. Fang, A novel local pca-based method for detecting activation signals in fMRI, *Magn. Resonance Imag.* **17** (1999), no. 6, 827–836.
[28] G. F. Loewenstein, E. U. Weber, C. K. Hsee and N. Welch, Risk as feelings, *Psychol. Bull.* **127** (2001), no. 2, 267–286.
[29] C. J. Long, E. N. Brown, C. Triantafyllou, I. Aharon, L. L. Wald and V. Solo, Nonstationary noise estimation in functional MRI, *NeuroImage* **28** (2005), no. 4, 890–903.
[30] H. Markowitz, Portfolio selection, *J. Finance* **7** (1952), no. 1, 77–91.
[31] B. A. Mellers, Choice and the relative pleasure of consequences, *Psychol. Bull.* **126** (2000), no. 6, 910–924.
[32] P. N. C. Mohr, G. Biele and H. R. Heekeren, Neural processing of risk, *J. Neurosci.* **30** (2010), no. 19, 6613–6619.
[33] P. N. C. Mohr, G. Biele, L. K. Krugel, S.-C. Li and H. R. Heekeren, Neural foundations of risk-return trade-off in investment decisions, *NeuroImage* **49** (2010), 2556–2563.
[34] M. P. Paulus, C. Rogalsky, A. Simmons, J. S. Feinstein and M. B. Stein, Increased activation in the right insula during risk-taking decision making is related to harm avoidance and neuroticism, *NeuroImage* **19** (2003), no. 4, 1439–1448.
[35] H. Plassmann, J. O'Doherty and A. Rangel, Orbitofrontal cortex encodes willingness to pay in everyday economic transactions, *J. Neurosci.* **27** (2007), no. 37, 9984–9988.
[36] J. W. Pratt, Risk aversion in the small and in the large, *Econometrica* **44** (1964), 122–136.
[37] K. Preuschoff, P. Bossaerts and S. R. Quartz, Neural differentiation of expected reward and risk in human subcortical structures, *Neuron* **51** (2006), no. 3, 381–390.
[38] K. Preuschoff, S. R. Quartz and Peter Bossaerts, Human insula activation reflects risk prediction errors as well as risk, *J. Neurosci.* **28** (2008), no. 11, 2745–2752.
[39] C. Radhakrishna Rao, Some statistical methods for comparison of growth curves, *Biometrics* **14** (1958), no. 1, 1–17.
[40] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*, 2nd ed., Springer, New York, 2005.
[41] A. Rangel, C. Camerer and P. R. Montague, A framework for studying the neurobiology of value-based decision making, *Nat. Rev. Neurosci.* **9** (2008), 545–556.
[42] E. T. Rolls, C. McCabe and J. Redoute, Expected value, reward outcome, and temporal difference error representations in a probabilistic decision task, *Cerebral Cortex* **18** (2008), no. 3, 652–663.
[43] W. F. Sharpe, Capital asset prices: A theory of market equilibrium under conditions of risk, *J. Finance* **19** (1964), no. 3, 425–442.
[44] D. Schunk and C. Betsch, Explaining heterogeneity in utility functions by individual differences in decision modes, *J. Econom. Psychol.* **27** (2006), no. 3, 386–401.
[45] P. N. Tobler, J. P. O'Doherty, R. J. Dolan and W. Schultz, Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems, *J. Neurophysiol.* **97** (2007), 1621–1632.
[46] S. M. Tom, C. R. Fox, C. Trepel and R. A. Poldrack, The neural basis of loss aversion in decision-making under risk, *Science* **315** (2007), no. 5811, 515–518.

[47] A. van Bömmel, S. Song, P. Majer, P. N. C. Mohr, H. R. Heekeren and W. K. Härdle, Risk patterns and correlated brain activities. Multidimensional statistical analysis of fMRI data in economic decision making study, *Psychometrika* **79** (2014), no. 3, 489–514.
[48] T. Vincent, L. Risser and P. Ciuciu, Spatially adaptive mixture modeling for analysis of fMRI time series, *IEEE Trans. Medical Imag.* **29** (2010), no. 4, 1059–1074.
[49] R. Viviani, G. Gron and M. Spitzer, Functional principal component analysis of fMRI data, *Human Brain Mapping* **24** (2005), 109–129.
[50] J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, 1953.
[51] E. U. Weber, The utility of measuring and modeling perceived risk, in: *Choice, Decision, and Measurement: Essays in Honor of R. Duncan Luce*, Lawrence Erlbaum Associates, Mawah (1997), 45–56.
[52] E. U. Weber and E. J. Johnson, Decisions under uncertainty: Psychological, economic, and neuroeconomic explanations of risk preference, in: *Neuroeconomics: Decision Making and the Brain*, Academic Press, New York (2008), 127–144.
[53] K. J. Worsley, C. H. Liao, J. Aston, V. Petre, G. H. Duncan, F. Morales and A. C. Evans, A general statistical analysis for fMRI data, *Neuroimage* **15** (2002), no. 1, 1–15.
[54] V. Zipunnikov, B. Caffo, D. M. Yousem, C. Davatzikos, B. S. Schwartz and C. Crainiceanu, Functional principal component model for high-dimensional brain imaging, *NeuroImage* **58** (2011), no. 3, 772–784.