

Accepted Manuscript

Principal component analysis in an asymmetric norm

Ngoc M. Tran, Petra Burdejová, Maria Ospienko, Wolfgang K. Härdle

PII: S0047-259X(17)30396-2

DOI: <https://doi.org/10.1016/j.jmva.2018.10.004>

Reference: YJMVA 4427

To appear in: *Journal of Multivariate Analysis*

Received date: 4 July 2017

Please cite this article as: N.M. Tran, et al., Principal component analysis in an asymmetric norm, *Journal of Multivariate Analysis* (2018), <https://doi.org/10.1016/j.jmva.2018.10.004>

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.



Principal component analysis in an asymmetric norm

Ngoc M. Tran^{a,b}, Petra Burdejová^{c,*}, Maria Ospienko^c, Wolfgang K. Härdle^{c,d}

^aDepartment of Mathematics, University of Texas at Austin, USA

^bInstitute for Applied Mathematics, University of Bonn, Germany

^cHumboldt-Universität zu Berlin, C.A.S.E. - Center for Applied Statistics and Economics, Unter den Linden 6, Berlin, Germany

^dSim Kee Boon Institute for Financial Economics, Singapore Management University, 90 Stamford Road, Singapore 178903

Abstract

Principal component analysis (PCA) is a widely used dimension reduction tool in high-dimensional data analysis. In risk quantification in finance, climatology and many other applications, however, the interest lies in capturing the tail variations rather than variation around the mean. To this end, we develop Principal Expectile Analysis (PEC), which generalizes PCA for expectiles. It can be seen as a dimension reduction tool for extreme-value theory, where fluctuations in the τ -expectile level of the data are approximated by a low-dimensional subspace. We provide algorithms based on iterative least squares, derive bounds on their convergence time, and compare their performance through simulations. We apply the algorithms to a Chinese weather dataset and fMRI data from an investment decision study.

Keywords: Asymmetric norm, Dimension reduction, Expectile, Growth data, Quantile, Risk attitude, Temperature

1. Introduction

Principal component analysis (PCA) is a classical dimension reduction tool widely used in fields such as genetics, environmental studies and demographics. The monographs of Jolliffe [11] and Ramsay and Silverman [20] contain many examples. The basic principle is to find a basis for a k -dimensional affine linear subspace that best approximates the data. If the data points are finite-dimensional vectors, the basis vectors are called principal components or factors. One then views each observation as a residual plus a point in this subspace, which is expressed as a vector in \mathbb{R}^k of coefficients, also called loadings. A classic example is the Canadian temperature dataset in Ramsay and Silverman [20], where temperature curves recorded daily over a year are considered at multiple stations in an area. The premise is that only few factors influence temperature across stations, and that the temperature curve from each station is well approximated on average by a specific linear combination of these factors.

In classical PCA, the optimal k -dimensional subspace is one that minimizes the L_2 -norm of the residuals. When $k = 0$, this is the mean of the data, so that classical PCA decomposes the data around its mean subspace. Much research in the broader field of functional data analysis has focused on the variation around an average pattern; see [6, 10]. In many applications such as risk analysis, however, one is not only interested in variations around the mean, but rather those around the tail of the data. For example, one may be interested in extreme phenomena like droughts, heavy rainfalls, or heat waves. Can the data be decomposed around the 99th quantile, say, and produce some ‘best’ principal component where only 1% of the observations have positive loadings? In the aforementioned temperature data, for instance, this principal component could be interpreted as one that influences locations with extreme temperatures.

Note that the above problem is different from finding the 99th quantile of the loadings in classical PCA. Doing so corresponds to keeping the same PCA-optimal subspace, and translating it so that each component has 1% positive loadings. The principal components are the same; the tail of the data is reflected by the loadings. In our setup, one wants to find a low-dimensional subspace that best approximates the data by some tail measure, say, an appropriate

*Corresponding author

Email address: petra.burdejova@hu-berlin.de (Petra Burdejová)

analogue of the 99th quantile. In this case, the tail of the data is reflected by the principal components. As we will show in Section 4, only in some special cases do these two methods give the same subspace.

In this paper, we generalize multivariate PCA to Principal Expectile Analysis, a method which, for a given expectile level τ , produces k principal expectile components (PECs) that best decompose the data around its τ -expectile. Classical PCA corresponds to the case $\tau = 0.5$. Expectiles, proposed in [19], are natural analogues of quantiles for the mean. While the τ -quantile minimizes asymmetric ℓ_1 -error, the τ -level expectile minimizes asymmetric ℓ_2 -error.

Expectiles enjoy several advantages over quantiles, including computational efficiency; see [23]. It is also more sensitive to extreme values in the data, and thus is preferred in the calculation of risk measures of a financial asset or a portfolio. For instance, value-at-risk (VaR) is commonly used to measure the downside risk, especially in portfolio risk management. Given a predetermined probability level, VaR represents the quantile of the portfolio loss distribution; see [12]. Since VaR merely depends on the probability value and neglects the size of the downside loss, it has been criticized as a risk measure. Alternative measures based on expectiles have been investigated; see, e.g., [5, 15].

Our definitions of PECs are different from the principal directions for quantiles of Fraiman and Pateiro-López [7]. These authors are focused on doing classical PCA for quantile level sets. Related is the work of Kong and Mizera [14], who propose the same approach but without centering the data at its expectation. Since the quantile has to be computed in each direction, the principal quantile directions in these papers can only be explicitly computed in small dimensions in general. Applying the methods of [7] to a dataset amounts to a two-step procedure: first transform the data to its quantile level sets, then do PCA. In contrast, our method generalizes PCA by minimizing the τ -variance in the residuals. It is a one-step procedure with an iterative approximation algorithm that runs well on large datasets; see Sections 5 and 8. Like the definition in [7], one can generalize PEC to functional data. Furthermore, our definition of PEC enjoys many nice properties, some of which are shared by the principal directions; see Section 4.

Though we make computational gains over [7], it is not without loss. In particular, for $\tau \neq 1/2$, our estimators generally do not have the nested subspace property. That is, the best k -dimensional subspace that minimizes the τ -variance in the data may not contain the best $(k - 1)$ -dimensional subspace. This issue was raised in other contexts; see [4, 26] and references therein. We show that this discrepancy leads to three plausible estimators for the principal expectile component, which we term TopDown (TD), BottomUp (BUP), and PrincipalExpectile (PEC). We compare the theoretical and computational properties of these estimators on both simulated and real-life datasets. In general, we find that TopDown has the best mean-squared error in simulations and the algorithm has much better convergence properties than PEC. In contrast, PEC runs somewhat faster, has comparable mean-squared error, and does not require a pre-specified number of components. As the three methods often give similar estimates, we advocate TopDown for small to medium size datasets, and PEC for larger datasets, where computation may be an issue.

Since we initiated this project, Principal Expectiles Analysis has found numerous applications, mainly in quantifying risks. In climate analysis, for example, Burdejová et al. [2] looked for trends and critical changing points in the strength of tropical storms in two different areas over several decades. The analysis considers the wind data observed every six hours represented as functional data for several τ -expectile levels. A proposed test based on principal components shows that there is a significant trend in the shape of the annual pattern of upper wind speed levels of hurricanes. In this setup, PECs yield time varying information of storm strength which lies between ‘typical’ and ‘extreme’ behavior. This approach can be applied to any environmental data that can be represented as annual curves which evolve from year to year, such as daily temperature or log-precipitation curves at specific locations.

We here present three more applications. The first analyzes extreme temperature based on recordings of daily temperature over the last 50 years for 159 stations in China. This is an analogue to the commonly known approach of Ramsay and Silverman [20]. For this dataset, the principal expectile components differ significantly from classical PCs. Furthermore, the PECs are interpretable. The first PEC shows the long-term changes in extreme temperature, while the second PEC reflects temporal seasonal extremes. The second application applies PEC to an fMRI dataset called the RPID (Risk Perception in Investment Decision) experiment. Majer et al. [16] found strong relations between fMRI reactions and diagnosed risk perception. Their results show that one can predict the risk perception parameter of each individual based on the classical principal components of the fMRI data. In contrast, we show that one can obtain even better predictions using PECs with $\tau = 0.6$. In other words, extreme fMRI reactions is a better predictor of more extreme behavior against risk. Our last application concerns the Berkeley Growth Study, in which the height of 39 boys and 54 girls was monitored. Principal component analysis of these growth curves was carried out in [20, 25]. We repeat this analysis for three different τ levels: $\tau = 0.05, 0.5$ and 0.95 . The PECs for $\tau = 0.95$ would represent fluctuations in growth among the tallest children, for instance. Our analysis reveals that height fluctuations differ

significantly across τ for boys, but follow the same pattern for all τ for girls. Compared to the average, both the tall and short group have less variation and the overall change in growth rate is less extreme.

In summary, we provide a solid optimization framework with algorithms to generalize multivariate PCA to principal expectile analysis, which can be seen as a dimension reduction tool for extreme-value theory. Different views of PCA lead to different ways to generalize it, and thus we obtain three different estimators. We give algorithms to estimate them and compare their performance in both synthetic and real-life datasets; we also offer interpretations of these results. Our work also leaves a number of interesting open problems, such as comparisons between quantiles and expectile, generalizations to functional data, and further statistical properties of our estimators.

Our paper is organized as follows. In Section 2 we review quantiles, expectiles and PCA. We then discuss the issues in generalizing PCA to expectiles, and propose a definition for principal expectile components, PrincipalExpectile algorithm and two other variations named TopDown and BottomUp. In Section 4, we report the statistical properties of these estimators. In Section 5, we provide algorithms to compute PEC, TopDown and BottomUp based on iterative weighted least squares, and derive upper bounds on their convergence time. We compare their performance in a simulation study in Section 6. In Section 7, we show applications to three different real-life datasets pertaining to the weather, fMRI and growth curves. The last section summarizes our findings. Further experimental results and discussion on various aspects of the algorithms are given in the Appendix.

2. Background

2.1. Quantiles and expectiles

Take the data dimension p as fixed. For $y \in \mathbb{R}^p$, we define $y_+ = \max(0, y)$, $y_- = \max(0, -y)$ coordinate-wise. For $\tau \in (0, 1)$, let $\|\cdot\|_1$ be the L_1 -norm in \mathbb{R}^p , i.e., $\|y\|_1 = |y_1| + \dots + |y_p|$. The L_1 -norm with weight τ in \mathbb{R}^p is

$$\|y\|_{\tau,1} = \tau\|y_+\|_1 + (1-\tau)\|y_-\|_1 = \sum_{j=1}^p |y_j| \times \{\tau \mathbf{1}(y_j \geq 0) + (1-\tau) \mathbf{1}(y_j < 0)\},$$

where $\mathbf{1}$ is the indicator function. Similarly, let $\|\cdot\|_2$ denote the L_2 -norm in \mathbb{R}^p , $\|y\|_2^2 = y_1^2 + \dots + y_p^2$. The asymmetric L_2 -norm with weight τ in \mathbb{R}^p is

$$\|y\|_{\tau,2}^2 = \tau\|y_+\|_2^2 + (1-\tau)\|y_-\|_2^2.$$

When $\tau = 1/2$, we recover constant multiples of the L_1 and L_2 -norms, respectively. These two families of norms belong to the general class of asymmetric norms with sign-sensitive weights. They stem from approximation theory; see [3]. Among other properties, these norms are convex, and their unit balls restricted to a given orthant in \mathbb{R}^p are weighted simplices for the $\|\cdot\|_{\tau,1}$ norm, and axis-aligned ellipsoids for the $\|\cdot\|_{\tau,2}$ norm. In other words, they coincide with the unit balls of axis-aligned weighted L_1 and L_2 norms.

Let $Y \in \mathbb{R}^p$ be a random variable with cumulative distribution function (cdf) F . The τ -quantile $q_\tau(Y) \in \mathbb{R}^p$ of F_Y is the solution to the following optimization problem:

$$q_\tau(Y) = \operatorname{argmin}_{q \in \mathbb{R}^p} E\|Y - q\|_{\tau,1}.$$

Similarly, the τ -expectile $e_\tau(Y) \in \mathbb{R}^p$ of F_Y is the solution to

$$e_\tau(Y) = \operatorname{argmin}_{e \in \mathbb{R}^p} E\|Y - e\|_{\tau,2}^2.$$

As shown by Cobzaş [3], both $q_\tau(Y)$ and $e_\tau(Y)$ exist and are unique, assuming that $E(Y)$ is finite.

2.2. Classical principal component analysis

There are multiple, equivalent ways of defining classical PCA, which generalize to different definitions of principal components for quantiles and expectiles. We focus on two formulations: minimizing the residual sum of squares, and maximizing the variance captured. For further details, see [11].

Suppose we observe n vectors $Y_1, \dots, Y_n \in \mathbb{R}^p$ with empirical distribution function (edf) F_n . Write Y for the $n \times p$ data matrix. PCA solves for the k -dimensional affine subspace that best approximates Y_1, \dots, Y_n in L_2 -norm. In matrix terms, we are looking for the constant $m^* \in \mathbb{R}^p$ and the matrix E_k^* , the rank- k matrix that best approximates $Y - \iota(m^*)^\top$ in the Frobenius norm, where ι stands for a vector of 1s. That is,

$$(m_k^*, E_k^*) = \underset{m \in \mathbb{R}^p, E \in \mathbb{R}^{n \times p}: \text{rank}(E)=k}{\text{argmin}} \|Y - \iota m^\top - E\|_{1/2,2}^2. \quad (1)$$

As written, m is not well defined: if (m, E) is a solution, then $(m + c, E - \iota c^\top)$ is another solution for any c in the column space of E . Geometrically, this means that we can express the affine subspace $m + E$ with respect to any chosen point m . It is intuitive to choose m to be the best constant in this affine subspace that approximates Y . By a least squares argument, the solution is $m_k^* = E(Y)$. That is, it is independent of k and coincides with the best constant approximation to Y . Thus, it is sufficient to assume $E(Y) = m \equiv 0$, and consider the optimization problem in (1) without the constant term.

Suppose Y is of full rank and the eigenvalues of its covariance matrix are distinct. This is necessary and sufficient for principal components to be unique. Again by least squares argument, for all $k \in \{1, \dots, p-1\}$, one can show that

$$E_k^* \subset E_{k+1}^*, \quad (2)$$

and $E_{k+1}^* \setminus E_k^*$ is the optimal rank-one approximation of $Y - E_k^*$. This has two consequences. First, there exists a natural basis for E_k^* . Indeed, there exists a unique ordered sequence of orthonormal vectors $v_1^*, \dots, v_p^* \in \mathbb{R}^p$ such that $E_1^* = U_1 V_1^\top$, $E_2^* = U_2 V_2^\top$, and so on, where the columns of V_k are the first k v_i^* s. The v_i^* s are called the principal components or factor. For fixed k , V_k is the component, or factor matrix, and U_k is the loading. The second implication of (2) is that one can compute the principal components by a greedy algorithm which solves k iterations of the one-dimensional version of (1).

The one-dimensional version of (1) has another characterization. The first principal component v^* is the unit vector in \mathbb{R}^p which maximizes the variance of the data projected onto the subspace spanned by v^* . That is,

$$v^* = \underset{v \in \mathbb{R}^p, v^\top v = 1}{\text{argmax}} \text{var}\{v^\top Y_i : 1 \leq i \leq n\} = \underset{v \in \mathbb{R}^p, v^\top v = 1}{\text{argmax}} \frac{1}{n} \sum_{i=1}^n (v^\top Y_i - \overline{v^\top Y})^2, \quad (3)$$

where var is the variance of the sequence in the argument, while $\overline{v^\top Y} = (v^\top Y_1 + \dots + v^\top Y_n)/n$ is the mean of the projected data, or equivalently, the projection of the mean \bar{Y} onto the subspace spanned by v . Given that the first principal component is v_1^* , the second principal component v_2^* is the unit vector in \mathbb{R}^p which maximizes the variance of the residual $Y_i - (v_1^*)^\top \bar{Y} - v_1^* (v_1^*)^\top Y_i$, and so on. In this formulation, the data do not need to be pre-centered. The sum $(v_1^*)^\top \bar{Y} + \dots + (v_k^*)^\top \bar{Y}$ is the overall mean \bar{Y} projected onto the subspace spanned by the first k principal components.

For the benefit of comparisons with Theorem 2, let us reformulate (3) as follows. Define

$$C = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})(Y_i - \bar{Y})^\top. \quad (4)$$

Then v^* is the solution to the following optimization problem:

$$\text{maximize } v^\top C v \quad \text{subject to } v^\top v = 1.$$

It is clear from this formulation that this optimization problem has a solution unique up to a sign if and only if C has a unique largest eigenvalue. For this reason, we will implicitly assume that all eigenvalues of C are distinct.

3. Principal expectile analysis

We now generalize the above definitions of PCA to those for expectiles, leading to principal expectile analysis. While we focus exclusively on expectiles in this paper, we note that the generalization for quantiles follows similarly, and algorithms for L_1 matrix factorization can also be adapted to this case.

The two views of PCA, minimizing-least-squares in (1), and maximizing-projected-variance in (4), are no longer equivalent when one optimizes these functions under the asymmetric L_2 -norm. This is because the asymmetric norm is not a projection. The analogue of (1) is the following low-rank matrix approximation problem:

$$(m_k^*, E_k^*) = \underset{m \in \mathbb{R}^p, E \in \mathbb{R}^{n \times p}: \text{rank}(E)=k}{\text{argmin}} \|Y - um^\top - E\|_{\tau,2}^2. \quad (5)$$

Again, we may define m to be the best constant approximation to Y on the affine subspace determined by (m, E) . For a fixed affine subspace, such a constant is unique and is the coordinate-wise τ -expectile of the residuals $Y - E$. However, the expectile is not additive for $\tau \neq 1/2$. Thus in general the column space of E_k^* is not a subspace of the column space E_{k+1}^* , the constant m_k^* depends on k , and is not equal to the τ -expectile $e_\tau(Y)$. In other words, even when E_k^* is a well-defined subspace, it does not come with a natural basis, and hence there are no natural candidates for ‘principal components’.

Fix k and consider the problem of computing m_k^* and E_k^* . Write a rank- k matrix E as $E = UV^\top$, where $U \in \mathbb{R}^{n \times k}$, $V \in \mathbb{R}^{p \times k}$. Adjoin U with an all-1 column to form \tilde{U} , and adjoin m to the corresponding column of V to form \tilde{V} . Thus $um^\top + E = \tilde{U}\tilde{V}^\top$. Eq. (5) is an unconstrained minimization problem over the pair of (adjoined) matrices (\tilde{U}, \tilde{V}) with minimization objective

$$J(\tilde{U}, \tilde{V}, W) = \|Y - \tilde{U}\tilde{V}^\top\|_{\tau,2}^2 = \sum_{i=1}^n \sum_{j=1}^p w_{ij} \left(Y_{ij} - m_j - \sum_{\ell} U_{i\ell} V_{\ell j} \right)^2,$$

where the weights w_{ij} are sign-dependent: $w_{ij} = \tau$ if $Y_{ij} - m_j - \sum_{\ell} U_{i\ell} V_{\ell j} > 0$, $w_{ij} = 1 - \tau$ otherwise.

To define principal expectile components, one can equip E_k^* with two types of basis, which we call TopDown and BottomUp. In TopDown, one first finds E_k^* . Then for $j \in \{1, \dots, k-1\}$, one finds E_j , the best j -dimensional subspace approximation to $Y - m_k^*$, subject to $E_{j-1} \subset E_j \subset E_k^*$. This defines a nested sequence of subspace $E_1 \subset \dots \subset E_{k-1} \subset E_k^*$, and hence a basis for E_k^* , such that E_j is an approximation of the best j -dimensional subspace approximation to $Y - m_k^*$ contained in E_k^* . In BottomUp, one first finds E_1^* . Then for $j \in \{2, \dots, k\}$, one finds (m_j, E_j) , the optimal j -dimensional affine subspace approximation to Y , subjected to $E_{j-1} \subset E_j$. In each step we re-estimate the constant term. Again, we obtain a nested sequence of subspaces $E_1^* \subset \dots \subset E_k^*$, and constant terms m_1, \dots, m_k , where (m_j, E_j) is the best affine j -dimensional subspace approximation to Y .

When $\tau = 1/2$, i.e., when doing usual PCA, both definitions correctly recover the principal components. For $\tau \neq 1/2$, they can produce different output. Interestingly, both in simulations and in practice, these outputs are not very different; see Sections 6 and 7. See Section 5 for a formal description of the TopDown and BottomUp algorithms and computational bounds on their convergence time.

Generalization of (3) is more fruitful, both theoretically and computationally. First we need a weighted definition of the variance. Let $Y \in \mathbb{R}$ be a random variable with cdf F . Its τ -variance is

$$\text{var}_\tau(Y) = \mathbb{E}\|Y - e_\tau\|_{\tau,2}^2 = \min_{e \in \mathbb{R}} \mathbb{E}\|Y - e\|_{\tau,2}^2,$$

where $e_\tau = e_\tau(Y)$ is the τ -expectile of Y . When $\tau = 1/2$, this reduces to the usual definition of variance. The direct generalization of (3) is

$$v_\tau^* = \underset{v \in \mathbb{R}^p, v^\top v = 1}{\text{argmax}} \text{var}_\tau(v^\top Y_1, \dots, v^\top Y_n) = \underset{v \in \mathbb{R}^p, v^\top v = 1}{\text{argmax}} \frac{1}{n} \sum_{i=1}^n (v^\top Y_i - \mu_\tau)^2 w_i$$

where $\mu_\tau \in \mathbb{R}$ is the τ -expectile of the sequence of n real numbers $v^\top Y_1, \dots, v^\top Y_n$, and

$$w_i = \begin{cases} \tau & \text{if } \sum_{j=1}^p Y_j v_j > \mu_\tau, \\ 1 - \tau & \text{otherwise.} \end{cases}$$

Definition 1. Suppose we observe $Y_1, \dots, Y_n \in \mathbb{R}^p$. The first principal expectile component (PEC) v_τ^* is the unit vector in \mathbb{R}^p that maximizes the τ -variance of the data projected on the subspace spanned by v_τ^* . That is, v_τ^* solves (6).

Like in classical PCA, the other components are defined based on the residuals, and thus by definition, they are orthogonal to the previously found components. Therefore one obtains a nested sequence of subspace which captures the tail variations of the data.

The τ -variance measures the spread of the data relative to the τ -expectile e_τ . For τ very close to 1, for instance, observations above e_τ are assigned the very high weight τ , while those below receive very little weight. Similarly, for τ very close to 0, observations below e_τ carry most of the weight. In other words, the τ -variance is dominated by the observations more extreme than e_τ . Thus, PEC, the direction that maximizes the τ -variance of the projected data can be interpreted as the direction with the most ‘extreme’ behavior in the loadings.

3.1. Relation to principal quantile directions

Though we only focus on expectiles, for ease of comparison with [7] and the related literature, we give the quantile analogue of our definition of PEC. By replacing the $\|\cdot\|_{\tau,2}^2$ norm with the $\|\cdot\|_{\tau,1}$ norm, one can define the analogue of principal component for quantiles. The analogue of τ -variance is the τ -deviation

$$\text{Dev}_\tau(Y) = \mathbb{E}\|Y - q_\tau(Y)\|_{\tau,1} = \min_{q \in \mathbb{R}} \mathbb{E}\|Y - q\|_{\tau,1}.$$

This leads to the optimization problem

$$v_{\tau,L_1}^* = \operatorname{argmax}_{v \in \mathbb{R}^p: \sum_j |v_j|=1} \text{Dev}_\tau(v^\top Y_1, \dots, v^\top Y_n).$$

One can define the first principal quantile component (PQC) v_{τ,L_1}^* as the L_1 -unit vector in \mathbb{R}^p that maximizes the τ -deviation captured by the data projected on the subspace spanned by v_{τ,L_1}^* . In comparison, Fraiman and Pateiro-López [7] define the first principal quantile direction ψ to be the one that maximizes the L_2 norm of the τ -quantile of the centered data, projected in the direction ψ . That is, ψ is the solution of

$$\max_{v \in \mathbb{R}^p: v^\top v=1} \|v^\top q_\tau(Y - \mathbb{E}Y)\|_{1/2,2}.$$

Kong and Mizera [14] proposed the same definition but without centering Y at $\mathbb{E}Y$.

4. Statistical properties of PEC

We now show that our definition of PEC satisfies many important properties, such as being compatible to orthogonal transformations of the data, and coinciding with classical PC for elliptically symmetric distributions; see Proposition 2. More important, we show that the empirical estimator in (6) is consistent under some mild uniqueness assumptions akin to the unique leading eigenvalue assumption in classical PCA.

Proposition 1 (Properties of τ -variance). *Let $Y \in \mathbb{R}$ be a random variable. For $\tau \in (0, 1)$, the following statements hold:*

- (i) $\text{var}_\tau(Y + c) = \text{var}_\tau(Y)$ for $c \in \mathbb{R}$;
- (ii) $\text{var}_\tau(sY) = s^2 \text{var}_\tau(Y)$ for $s \in \mathbb{R}, s > 0$;
- (iii) $\text{var}_\tau(-Y) = \text{var}_{1-\tau}(Y)$.

Proof. The first two follow directly from corresponding properties for e_τ . We will prove that last assertion. Recall that $e_\tau(-Y) = -e_{1-\tau}(Y)$. Thus

$$\text{var}_\tau(-Y) = \mathbb{E}\| -Y - e_\tau(-Y) \|_{\tau,2}^2 = \mathbb{E}\| -Y - e_{1-\tau}(Y) \|_{\tau,2}^2 = \mathbb{E}\| Y - e_{1-\tau}(Y) \|_{1-\tau,2}^2 = \text{var}_{1-\tau}(Y).$$

Hence the argument is complete. \square

As a corollary, we see that PECs are sign-sensitive in general, unless the distribution of Y is symmetric or $\tau = 1/2$.

Corollary 1. For $\tau \in (0, 1)$ and random variable $Y \in \mathbb{R}^p$, suppose v_τ^* is a first τ -PEC of Y . If $\tau \neq 1/2$, then $-v_\tau^* = v_{1-\tau}^*$, i.e., $-v_\tau^*$ is also a first $(1 - \tau)$ -PEC of Y . Furthermore, if the distribution of Y is symmetric about 0, i.e., $Y =_{\mathcal{L}} -Y$, or if $\tau = 1/2$, then $-v_\tau^*$ is also a first τ -PEC of Y .

Proof. By Proposition 1, $\text{var}_\tau(v_\tau^{*\top} Y) = \text{var}_{1-\tau}\{(-v_\tau^{*\top})Y\}$. Thus if v_τ^* solves (6) for τ , then so does $(-v_\tau^*)^*$ for $1 - \tau$. If the distribution of Y is symmetric about 0, then $\text{var}_\tau(v_\tau^{*\top} Y) = \text{var}_{1-\tau}\{v_\tau^{*\top}(-Y)\} = \text{var}_\tau(v_\tau^{*\top} Y)$. In this case $-v_\tau^* = v_{1-\tau}^*$ is another τ -PEC of Y . \square

Proposition 2. [Properties of principal expectile component] Let $Y \in \mathbb{R}^p$ be a random variable and $v_\tau^*(Y)$ its unique first principal expectile component as given in Definition 1.

- (i) For any constant $c \in \mathbb{R}^p$, $v_\tau^*(Y + c) = v_\tau^*(Y)$. In words, the PEC is invariant under data translations.
- (ii) If $B \in \mathbb{R}^{p \times p}$ is an orthogonal matrix, then $v_\tau^*(BY) = Bv_\tau^*(Y)$. In words, the PEC respects changes of basis.
- (iii) If the distribution of Y is elliptically symmetric about some point $c \in \mathbb{R}^p$, i.e., there exists an invertible $p \times p$ real matrix A such that $BA^{-1}(Y - c) =_{\mathcal{L}} A^{-1}(Y - c)$ for all orthogonal matrices B , then $v_\tau^*(Y) = v_{1/2}^*(Y)$. In this case, the PEC coincides with the classical PC regardless of τ .
- (iv) If the distribution of Y is spherically symmetric about some point $c \in \mathbb{R}^p$, i.e., $B(Y - c) =_{\mathcal{L}} Y - c$ for all orthogonal matrix B , then all directions are principal.

Proof. By the first part of Proposition 1,

$$\text{var}_\tau\{v^\top(Y_i + c) : i \in \{1, \dots, n\}\} = \text{var}_\tau\{v^\top Y_i + v^\top c : i \in \{1, \dots, n\}\} = \text{var}_\tau\{v^\top Y_i : i \in \{1, \dots, n\}\}.$$

This proves the first statement. For the second, note that

$$\text{var}_\tau\{v^\top B Y_i : i \in \{1, \dots, n\}\} = \text{var}_\tau\{(B^\top v)^\top Y_i : i \in \{1, \dots, n\}\}.$$

Thus if v_τ^* is the first τ -PEC of Y , then $(B^\top)^{-1}v_\tau^*$ is the first τ -PEC of BY . But B is orthogonal, i.e., $(B^\top)^{-1} = B$. Hence Bv_τ^* is the τ -PEC of BY . This proves the second statement.

For the third statement, by Statement 1, we can assume $c \equiv 0$. Thus $Y = AZ$, where $BZ =_{\mathcal{L}} Z$ for all orthogonal matrices B . Write A in its singular value decomposition $A = UDV$, where D is a diagonal matrix with positive values $D_{ii} = d_i$ for $i \in \{1, \dots, p\}$, and U and V are $p \times p$ orthogonal matrices. Choosing $B = V^{-1}$ gives

$$v_\tau^*(Y) = v_\tau^*(UDZ) = Uv_\tau^*(DZ).$$

Now, by Proposition 1, since $d_j \geq 0$ for all $j \in \{1, \dots, p\}$,

$$\text{var}_\tau(v^\top DZ) = \text{var}_\tau\left(\sum_{j=1}^p d_j Z_j v_j\right) = \sum_j v_j^2 d_j^2 \text{var}_\tau(Z_j).$$

Since $\sum_j v_j^2 = 1$, $\text{var}_\tau(v^\top DZ)$ lies in the convex hull of the p numbers $d_j^2 \text{var}_\tau(Z_j)$ for $j \in \{1, \dots, p\}$. Therefore, it is maximized by setting v to be the unit vector along the axis j with maximal $d_j^2 \text{var}_\tau(Z_j)$.

But $Z =_{\mathcal{L}} BZ$ for all orthogonal matrices B and so $Z_j =_{\mathcal{L}} Z_k$. Therefore, $\text{var}_\tau(Z_j) = \text{var}_\tau(Z_k)$ for all indices $j, k \in \{1, \dots, p\}$. Thus $\text{var}_\tau(v^\top DZ)$ is maximized when v is the unit vector along the axis j with maximal d_j . This is precisely the axis with maximal singular value of A , and hence is also the direction of the (classical) principal component of DZ . This proves the claim. The last statement follows immediately from the third statement. \square

We now prove consistency of local maximizers of (6). The main theorem in this section is the following.

Theorem 1. Fix $\tau > 0$. Let Y be a random variable in \mathbb{R}^p with finite second moment and distribution function F . Suppose $v^* = v_\tau^*$ is a unique global solution to (6) corresponding to Y . Suppose we observe n iid copies of Y , with

empirical distribution function F_n . Let Y_n be a random variable whose cdf is F_n . Then for sufficiently large n , for any sequence of global solutions v_n^* of (6) corresponding to Y_n , we have, as $n \rightarrow \infty$,

$$v_n^* \xrightarrow{F\text{-a.s.}} v^* \text{ in } \mathbb{R}^p.$$

For the proof, we first need the following lemma.

Lemma 1. *Under the assumptions of Theorem 1, uniformly over all $v \in \mathbb{R}^p$ with $v^\top v = 1$, and uniformly over all $\tau \in (0, 1)$,*

$$\text{var}_\tau(Y_n^\top v) \xrightarrow{F\text{-a.s.}} \text{var}_\tau(Y^\top v).$$

Proof. Since Y_n is the empirical version of Y and the set of all unit vectors $v \in \mathbb{R}^p$, $v^\top v = 1$ is compact, it follows from the Cramér–Wold theorem that $Y_n^\top v \rightsquigarrow Y^\top v$ uniformly over all such unit vectors $v \in \mathbb{R}^p$. It then follows that e_τ and var_τ , which are completely determined by the distribution function, also converge F -a.s. uniformly over all v . \square

Proof of Theorem 1. Let \mathbb{S}^{p-1} denote the unit sphere in \mathbb{R}^p . Equip \mathbb{R}^p with the Euclidean norm $\|\cdot\|$. Define the map $V_Y : \mathbb{S}^{p-1} \rightarrow \mathbb{R}$, $V_Y(v) = \text{var}_\tau(Y^\top v)$. Fix $\epsilon > 0$. We will prove that there exists $\delta > 0$ such that the global minimum of V_{Y_n} is necessarily within δ -distance of v^* .

Since V_Y is continuous, \mathbb{S}^{p-1} is compact, and v^* is unique, there exists a sufficiently small $\delta > 0$ such that

$$|V_Y(v) - V_Y(v^*)| < \epsilon \quad \Rightarrow \quad \|v - v^*\| < \delta$$

for $v \in \mathbb{S}^{p-1}$. In particular, if $\|v - v^*\| > \delta$, then $V_Y(v^*) + \epsilon < V_Y(v)$. By Lemma 1, $V_{Y_n} \rightarrow V_Y$ as $n \rightarrow \infty$ uniformly over \mathbb{S}^{p-1} . In particular, there exists a large N such that for every integer $n > N$, $|V_{Y_n}(v) - V_Y(v)| < \epsilon/6$ for all $v \in \mathbb{S}^{p-1}$. Thus for $v \in \mathbb{S}^{p-1}$ such that $\|v - v^*\| > \delta$, $V_{Y_n}(v) - V_Y(v^*) > \epsilon - \epsilon/6 = 5\epsilon/6$. Meanwhile, since V_Y is continuous, one can choose $\epsilon' = \epsilon/6$, and thus obtain δ' such that $\|v - v^*\| < \delta' \Rightarrow |V_Y(v) - V_Y(v^*)| < \epsilon/6$. Then, for v such that $\|v - v^*\| < \delta'$,

$$V_{Y_n}(v) - V_Y(v^*) \leq |V_{Y_n}(v) - V_Y(v)| + |V_Y(v) - V_Y(v^*)| < \epsilon/6 + \epsilon/6 = \epsilon/3.$$

So far we have shown that if $\|v - v^*\| > \delta$, then $V_{Y_n}(v)$ is at least $5\epsilon/6$ larger than $V_Y(v^*)$. Meanwhile, if $\|v - v^*\| < \delta'$, then $V_{Y_n}(v)$ is at most $\epsilon/3$ larger than $V_Y(v^*)$. Thus the global minimum v_n^* of V_{Y_n} necessarily satisfies $\|v_n^* - v^*\| < \delta$. This completes the proof. \square

4.1. PEC as constrained PCA

To compute the principal expectile component v_τ^* , one needs to optimize the right-hand side of (6) over all unit vectors v . Although this is a differentiable function in v , optimizing it is a difficult problem, since μ_τ also depends on v and does not have a closed form solution. However, Theorem 2 below shows that in certain situations, for given weights w_i , not only μ_τ but also v_τ^* do have closed form solutions. In particular, in this setting, PEC is the constrained classical PC of a weighted version of the covariance matrix of the data, centered at a constant possibly different from the mean. This theorem forms the backbone of our iterative algorithm for computing PEC discussed in Section 5.

Theorem 2. *Consider (6). Suppose we are given the true weights w_i , which are either τ or $1 - \tau$. Let $\tau_+ = \{i \in \{1, \dots, n\} : w_i = \tau\}$ be the set of observations Y_i with ‘positive’ labels, and $\tau_- = \{i \in \{1, \dots, n\} : w_i = 1 - \tau\}$ be its complement. Let n_+ and n_- be the sizes of the respective sets. Define an estimator $\hat{e}_\tau \in \mathbb{R}^p$ of the τ -expectile via*

$$\hat{e}_\tau = \frac{\tau \sum_{i \in \tau_+} Y_i + (1 - \tau) \sum_{i \in \tau_-} Y_i}{\tau n_+ + (1 - \tau) n_-}. \quad (6)$$

Define

$$C_\tau = \frac{\tau}{n} \left\{ \sum_{i \in \tau_+} (Y_i - \hat{e}_\tau)(Y_i - \hat{e}_\tau)^\top \right\} + \frac{1 - \tau}{n} \left\{ \sum_{i \in \tau_-} (Y_i - \hat{e}_\tau)(Y_i - \hat{e}_\tau)^\top \right\}. \quad (7)$$

Then v_τ^* is the solution to the following optimization problem:

$$\text{maximize } v^\top C_\tau v \quad \text{subject to } v^\top Y_i > v^\top \hat{e}_\tau \quad \Leftrightarrow \quad i \in \tau_+, v^\top v = 1. \quad (8)$$

Proof. Since the weights are the true weights coming from the true principal expectile component v_τ^* , clearly v_τ^* satisfies the constraint in (8). Now suppose v is another vector in this constraint set. By definition of \hat{e}_τ , $v^\top \hat{e}_\tau$ is exactly μ_τ , the τ -expectile of the sequence of n real numbers $v^\top Y_1, \dots, v^\top Y_n$. Therefore, the quantity we need to maximize in (6) reads

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n (v^\top Y_i - \mu_\tau)^2 w_i &= \frac{\tau}{n} \sum_{i \in \tau_+} (v^\top Y_i - v^\top \hat{e}_\tau)^2 + \frac{1-\tau}{n} \sum_{i \in \tau_-} (v^\top Y_i - v^\top \hat{e}_\tau)^2 \\ &= \frac{\tau}{n} \sum_{i \in \tau_+} v^\top (Y_i - \hat{e}_\tau)(Y_i - \hat{e}_\tau)^\top v + \frac{1-\tau}{n} \sum_{i \in \tau_-} v^\top (Y_i - \hat{e}_\tau)(Y_i - \hat{e}_\tau)^\top v = v^\top C_\tau v. \end{aligned}$$

Thus the optimization problem above is indeed an equivalent formulation of (6), which was used to define v_τ^* . Finally, the last observation follows by comparing the above with the optimization formulation for PCA; see the paragraph after (4). Indeed, when $\tau = 1/2$, $\hat{e}_{1/2} = \bar{Y}$, $C_{1/2} = C$, and we recover the classical PCA. \square

5. Algorithms

5.1. Principal expectile components

Suppose the conditions of Theorem 2 are satisfied, so that finding PEC is the problem of solving a constrained PCA given in (8), but with unknown weights depending on the true principal direction. Since \hat{e}_τ is a linear function in Y_i , (8) defines a system of linear constraints in the entries of Y_i and v_τ^* . Thus for each fixed sign sets (τ_+, τ_-) , there exist (not necessarily unique) local optima $v_\tau^*(\tau_+, \tau_-)$. There are 2^n possible sign sets, one of which corresponds to the global optimum v_τ^* that we need. It is clear that finding the global optimum v_τ^* by enumerating all possible sign sets is intractable. However, in many situations, the constraint in (8) is inactive. That is, the largest eigenvector of C_τ satisfies (8) for free. In such situations, we call v_τ^* a stable solution. Just like classical PCA, stable solutions are unique for matrices C_τ with unique principal eigenvalue. More importantly, we have an efficient algorithm for finding stable solutions, if they exist.

Definition 2. For some given sets of weights $w = (w_i)$, define $e_\tau(w)$ via (6), $C_\tau(w)$ via (7). Let $v_\tau(w)$ be the largest eigenvector of $C_\tau(w)$. If $v_\tau(w)$ satisfies (8), we say that $v_\tau(w)$ is a locally stable solution with weights w .

To find locally stable solutions, one can solve (3) using iterative re-weighted least squares: first initialize the w_i s, compute estimators $\mu_\tau(w)$ and $v_\tau(w)$ ignoring the constraint (8), update the weights via (3), and iterate. At each step of this algorithm, one finds the principal component of a weighted covariance matrix with some approximate weight. Since there are only finitely many possible weight sets, the algorithm is guaranteed to converge to a locally stable solution if it exists. In particular, if the true solution to (3) is stable, then for appropriate initial weights, the algorithm will find this value. We call this algorithm `PrincipalExpectile`.

We now describe the details of this algorithm for the case $k = 1$, i.e., the algorithm for computing the first principal expectile component only. To obtain higher order components, one iterates the algorithm over the residuals $Y_i - \hat{v}_1(\hat{v}_1^\top Y_i + \hat{\mu}_1)$, where $\hat{\mu}_1$ is the τ -expectile of the loadings $\hat{v}_1^\top Y_i$.

For n observations Y_1, \dots, Y_n , there are at most 2^n possible labels for the Y_i s, and hence the algorithm has in total 2^n possible values for the w_i s. Thus either Algorithm 1 converges to a point which satisfies the properties of the optimal solution that Theorem 2 prescribes, or it iterates infinitely over a cycle of finitely many possible values of the w_i s. In particular, the true solution is a fixed point, and thus fixed points always exist. In practice, we find that the algorithm converges very quickly, though occasionally it can get stuck in a finite cycle of values; see the Appendix. In this case, one can jump to a different starting point and restart the algorithm. Choosing a good starting value is important in ensuring convergence. In practice we employ two initialization methods: random and incremental. With the first method, one selects a random fraction (say, 10%) of the weights and initialize it to $1 - \tau$, while the rest to τ . One repeats PEC estimation for m different initialization settings, take the average curve, and then use the weights induced by this curve for initialization. With the second method, one performs a sequence of PEC computations for a sequence of τ starting with $\tau = 1/2$, and set the initial weight to be that induced by the previous run of the algorithm for a slightly smaller τ . Since the τ -variance is a continuous function in τ , the incremental method is theoretically

Algorithm 1 PrincipalExpectile

```

1: Input: Data  $Y \in \mathbb{R}^{n \times p}$ .
2: Output: A vector  $\hat{v}$ , an estimator of the first principal expectile component of  $Y$ .
3: procedure PRINCIPALEXPECTILE( $Y$ )
4:   Initialize the weights  $w_i^{(0)}$ .
5:   Set  $t = 0$ .
6:   repeat
7:     Let  $\tau_+^{(t)}$  be the set of indices  $i$  such that  $w_i^{(t)} = \tau$ , and  $\tau_-^{(t)}$  be the complement.
8:     Compute  $e_\tau^{(t)}$  as in Eq. (6) with sets  $\tau_+^{(t)}, \tau_-^{(t)}$ .
9:     Compute  $C_\tau^{(t)}$  as in Eq. (7) with sets  $\tau_+^{(t)}, \tau_-^{(t)}$ .
10:    Set  $v^{(t)}$  to be the largest eigenvector of  $C_\tau^{(t)}(C_\tau^{(t)})^\top$ .
11:    Set  $\mu_\tau^{(t)}$  to be the  $\tau$ -expectile of  $(v^{(t)})^\top Y_i$ .
12:    Update  $w_i$ : set  $w_i^{(t+1)} = \tau$  if  $(v^{(t)})^\top Y_i > \mu_\tau^{(t)}$ , and set  $w_i^{(t+1)} = 1 - \tau$  otherwise.
13:    Set  $t = t + 1$ .
14:   until  $w_i^t = w_i^{(t+1)}$  for all  $i$ .
15: return  $\hat{v} = v^{(t)}$ 
16: end procedure

```

sound, and produces a deterministic estimate for each choice of increment sequence. On simulated data, we found that both methods produce the same estimate; see the Appendix.

5.2. TopDown and BottomUp

We now describe how iterative weighted least squares can be adapted to implement TopDown and BottomUp. We start with a description of the asymmetric weighted least squares (LAWS) algorithm of Newey and Powell [19]. The basic algorithm outputs a subspace without the affine term, and needs to be adapted. See [9] for a variation with smoothing penalty and spline basis.

Algorithm 2 Asymmetric weighted least squares (LAWS)

```

1: Input: Data  $Y \in \mathbb{R}^{n \times p}$ , positive integer  $k < p$ .
2: Output:  $\hat{E}_k^*$ , an estimator of  $E_k^*$ , expressed in product form  $\hat{E}_k^* = \hat{U}\hat{V}^\top$ , where  $\hat{U} \in \mathbb{R}^{n \times k}$ ,  $\hat{V} \in \mathbb{R}^{p \times k}$ ;  $\hat{U}, \hat{V}$  are unique up to multiplication by an invertible matrix.
3: procedure LAWS( $Y, k$ )
4:   Set  $V^{(0)}$  to be some rank- $k$   $p \times k$  matrix.
5:   Set  $W^{(0)} \in \mathbb{R}^{n \times p}$  to be 1/2 everywhere.
6:   Set  $t = 0$ .
7:   repeat
8:     Update  $U$ : Set  $U^{(t+1)} = \operatorname{argmin}_{U \in \mathbb{R}^{n \times k}} J(U, V^{(t)}, W^{(t)})$ .
9:     Update  $W$ : Set  $W_{ij}^{(t+1)} = \tau$  if  $Y_{ij} - \sum_l U_{il}^{(t+1)} V_{lk}^{(t)} > 0$ ,  $W_{ij}^{(t+1)} = 1 - \tau$  otherwise.
10:    Update  $V$ : Set  $V^{(t+1)} = \operatorname{argmin}_{V \in \mathbb{R}^{k \times p}} J(U^{(t+1)}, V, W^{(t+1)})$ .
11:    Update  $W$ : Set  $W_{ij}^{(t+1)} = \tau$  if  $Y_{ij} - \sum_l U_{il}^{(t+1)} V_{lk}^{(t+1)} > 0$ ,  $W_{ij}^{(t+1)} = 1 - \tau$  otherwise.
12:    Set  $t = t + 1$ .
13:   until  $U^{(t+1)} = U^{(t)}, V^{(t+1)} = V^{(t)}, W^{(t+1)} = W^{(t)}$ .
14: return  $\hat{E}_k = U^{(t)}(V^{(t)})^\top$ 
15: end procedure

```

Proposition 3. *The LAWS algorithm is well defined, and is a gradient descent algorithm. Thus it converges to a critical point of the optimization problem (1).*

Proof. First note that the steps in the algorithm are well defined. For fixed W and V , $J(U, V, W)$ is a quadratic in the entries of U . Thus the global minimum on line 8 has an explicit solution; see [9, 24]. A similar statement applies to line 9. Note that $J(U, V, W)$ is not jointly convex in U and V , but as a function in U for fixed V , it is a convex, continuously differentiable, piecewise quadratic function. The statement holds for $J(U, V, W)$ as a function in V for fixed U . Hence lines 8 and 9 is one step in a Newton–Raphson algorithm on $J(U, V, W)$ for fixed V . Similarly, lines 10 and 11 is one step in a Newton–Raphson algorithm on $J(U, V, W)$ for fixed U . Thus the algorithm is a coordinate-wise gradient descent on a coordinate-wise convex function, and hence it converges. \square

If some columns of U or V are pre-specified, one can run LAWS and not update these columns in lines 8 and 10. Thus one can use LAWS to find the optimal affine subspace by writing $im^\top + E = \tilde{U}\tilde{V}$ with the first column of \tilde{U} constrained to be ι . Similarly, we can use this technique to solve the constrained optimization problems:

Find a rank- k approximation E_k whose span contains a given subspace of dimension $r < k$.

Solution: Constrain the first r columns of $V^{(0)}$ to be a basis of the given subspace.

Find a rank- k approximation whose span lies within a given subspace of dimension $r > k$.

Solution: Let $B \in \mathbb{R}^{n \times r}$ be a basis of the given subspace. Then the optimization problem becomes

$$\min_{U \in \mathbb{R}^{n \times k}, V \in \mathbb{R}^{p \times k}} \|Y - BUV^\top\|_{\tau, 2}^2.$$

One can then apply the LAWS algorithm with variables U and V .

Find a rank- k approximation whose span contains a given subspace of dimension $r < k$, and is contained in a given subspace of dimension $R > k$.

Solution: Combine the previous two solutions.

Algorithm 3 TopDown

- 1: *Input:* Data $Y \in \mathbb{R}^{n \times p}$, positive integer $k < p$.
 - 2: *Output:* \hat{E}_k^* , an estimator of E_k^* , expressed in product form $\hat{E}_k^* = \hat{U}\hat{V}^\top$, where $\hat{U} \in \mathbb{R}^{n \times k}$, $\hat{V} \in \mathbb{R}^{p \times k}$ are unique.
 - 3: **procedure** TOPDOWN(Y, k)
 - 4: Use LAWS(Y, k) to find \hat{m}_k^*, \hat{E}_k^* . Write $\hat{E}_k^* = UV^\top$ for some orthonormal basis U .
 - 5: Use LAWS to find \hat{U}_1 , the vector which spans the optimal subspace of dimension 1 contained in U .
 - 6: Use LAWS to find \hat{U}_2 , where (\hat{U}_1, \hat{U}_2) spans the optimal subspace of dimension 1 contained in U and contains the span of \hat{U}_1 .
 - 7: Repeat the above step until obtains \hat{U} .
 - 8: Obtain \hat{V} through the constraint $\hat{E}_k^* = \hat{U}\hat{V}^\top$.
 - 9: **return** $\hat{m}_k^*, \hat{E}_k^*, \hat{U}, \hat{V}^\top$
 - 10: **end procedure**
-

Algorithm 4 BottomUp

- 1: *Input:* Data $Y \in \mathbb{R}^{n \times p}$, positive integer $k < p$.
 - 2: *Output:* \hat{E}_k^* , an estimator of E_k^* , expressed in product form $\hat{E}_k^* = \hat{U}\hat{V}^\top$, where $\hat{U} \in \mathbb{R}^{n \times k}$, $\hat{V} \in \mathbb{R}^{p \times k}$ are unique.
 - 3: **procedure** BOTTOMUP(Y, k)
 - 4: Use LAWS to find \hat{E}_1^* . Let \hat{U}_1 be the basis vector.
 - 5: Use LAWS to find \hat{U}_2 such that (\hat{U}_1, \hat{U}_2) is the best bivariate approximation to Y , subject to containing \hat{U}_1 .
 - 6: Repeat the above step until obtains \hat{U} . We obtain \hat{V} and \hat{E}_k^* in the last iteration.
 - 7: **return** $\hat{E}_k^*, \hat{U}, \hat{V}^\top$
 - 7: **end procedure**
-

With these tools, we now define the two algorithms, TopDown and BottomUp. The TopDown algorithm requires the weights w_{ij} and the loadings on previous principal components to be re-evaluated when finding the next principal component. A variant of the algorithm would be to keep the weights w_{ij} , in which case the algorithm is still well defined though it will produce a different basis matrix \hat{U} , since the estimators are no longer optimal in the $\|\cdot\|_{\tau,2}^2$ norm.

5.3. Performance bounds of TopDown and BottomUp

We now show that the dependence on k only grows polylog in n . Thus both TopDown and BottomUp are fairly efficient algorithms even for large k .

Theorem 3. *For fixed V of dimension k , LAWS requires at most $O\{\ln(p)^k\}$ iterations, $O\{npk^2 \ln(p)^k\}$ flops to estimate U .*

In other words, if V has converged, LAWS needs at most $O\{npk^2 \ln(p)^k\}$ flops to estimate U . The role of U and V are interchangeable if we transpose Y . Thus if U has converged, LAWS needs at most $O\{npk^2 \ln(n)^k\}$ to estimate V . We do not have a bound for the number of iterations needed until convergence. In practice this seem to be of order log of n and p . For the proof of Theorem 3, we need the following two lemmas.

Lemma 2. *If $Y_1, \dots, Y_n \in \mathbb{R}$ are n real numbers, then LAWS finds their τ -expectile e_τ in $O\{\ln(n)\}$ iterations.*

Proof. Given the weights w_1, \dots, w_n , that is, given which Y_i 's are above and below e_τ , the τ -expectile e_τ is a linear function in the Y_i as we saw in (6). As shown in Proposition 3, LAWS is equivalent to a Newton–Raphson algorithm on a piecewise quadratic function. Since the Y_i s are ordered, it takes $O\{\ln(n)\}$ to learn their true weights. Thus the algorithm converges in $O\{\ln(n)\}$ iterations. \square

Lemma 3. *An affine line in \mathbb{R}^p can intersect at most $2p$ orthants.*

Proof. Recall that an orthant of \mathbb{R}^p is a subset of \mathbb{R}^p where the sign of each coordinate is constrained to be either nonnegative or nonpositive. There are 2^p orthants in \mathbb{R}^p . Let $f(\lambda) = Y + \lambda v$ be our affine line, $\lambda \in \mathbb{R}, Y, v \in \mathbb{R}^p$. Let $\text{sgn} : \mathbb{R}^p \rightarrow \{\pm 1\}^p$ denote the sign function. Now, $\text{sgn}\{f(0)\} = \text{sgn}(Y)$, $\text{sgn}\{f(\infty)\} = \text{sgn}(v)$, and $\text{sgn}\{f(\lambda)\}$ is a monotone increasing function in λ . As $\lambda \rightarrow \infty$, $\text{sgn}\{f(\lambda)\}$ goes from $\text{sgn}(Y)$ to $\text{sgn}(v)$ one bit flip at a time. Thus there are at most p flips, i.e., the half-line $f(\lambda)$ for $\lambda \in [0, \infty)$ intersects at most p orthants. By a similar argument, the half-line $f(\lambda)$ for $\lambda \in (-\infty, 0)$ intersects at most p other orthants. This concludes the proof. \square

Corollary 2. *An affine subspace of dimension k in \mathbb{R}^p can intersect at most $O(p^k)$ orthants.*

Proof. Fix any basis, say ψ_1, \dots, ψ_k . By Lemma 3, ψ_1 can intersect at most $2p$ orthants. For each orthant of ψ_1 , varying along ψ_2 can yield at most another $2p$ orthants. The proof follows by induction. This is a rather liberal bound, but it is of the correct order for k small relative to p . \square

Proof of Theorem 3. By Corollary 2, it is sufficient to consider the case $k = 1$. Fix V of dimension 1. Since U, V are column matrices, we write them in lower case letters u, v . Solving for each u_i is a separate problem; thus we have n separate optimization problem, and it is sufficient to prove the claim for each $i \in \{1, \dots, n\}$.

Fix $i \in \{1, \dots, n\}$. As u_i varies, $Y_i - m_i - u_i v$ defines a line in \mathbb{R}^p . The weight vector (w_{i1}, \dots, w_{ip}) only depends on which coordinates are the orthant of \mathbb{R}^p in which $Y_i - m_i - u_i v$ is in. The latter is equivalent to determining the weights of the p points $(Y_i - m_i)/v_i$. By Lemma 2, it takes $O\{\ln(p)\}$ for LAWS to determine the weights correctly. Thus LAWS takes at most $O\{\ln(p)\}$ iterations to converge, since each iteration involves estimating w , then v . Each iteration solves a weighted least squares, thus takes $O(npk^2)$. Hence for fixed v , LAWS can estimate u after at most $O\{npk^2 \ln(p)\}$ flops for $k = 1$. This concludes the proof for fixed v . By considering the transposed matrix Y , we see that the roles of u and v are interchangeable. The conclusion follows similarly for fixed u . \square

6. Simulation

To study the finite-sample properties of the proposed algorithms we perform a simulation study. We follow the simulation setup of Guo et al. [9], i.e., we simulate the data Y_{ij} for all $i \in \{1, \dots, n\}$, $j \in \{1, \dots, p\}$ as

$$Y_{ij} = \mu(t_j) + f_1(t_j)\alpha_{1i} + f_2(t_j)\alpha_{2i} + \varepsilon_{ij}, \quad (9)$$

where the t_j s are equidistant on $[0, 1]$, $\mu(t) = 1 + t + \exp\{-(t - 0.6)^2/0.05\}$ is the mean function, $f_1(t) = \sqrt{2} \sin(2\pi t)$ and $f_2(t) = \sqrt{2} \cos(2\pi t)$ are principal component curves, and ε_{ij} is a random noise. We consider two settings, each with five error scenarios:

1. $\alpha_{1i} \sim \mathcal{N}(0, 36)$ and $\alpha_{2i} \sim \mathcal{N}(0, 9)$ are both iid and the ε_{ij} s are (1) iid $\mathcal{N}(0, \sigma_1^2)$, (2) iid $t_{(5)}$, (3) independent $\mathcal{N}[0, \mu(t_j)\sigma_1^2]$, (4) iid $\ln \mathcal{N}(0, \sigma_1^2)$ and (5) iid sums of two uniforms $\mathcal{U}(0, \sigma_1^2)$ with $\sigma_1^2 = 0.5$.
2. $\alpha_{1i} \sim \mathcal{N}(0, 16)$ and $\alpha_{2i} \sim \mathcal{N}(0, 9)$ are both iid and the ε_{ij} s are (1) iid $\mathcal{N}(0, \sigma_2^2)$, (2) iid $t_{(5)}$, (3) independent $\mathcal{N}[0, \mu(t_j)\sigma_2^2]$, (4) iid $\ln \mathcal{N}(0, \sigma_2^2)$ and (5) iid sums of two uniforms $\mathcal{U}(0, \sigma_2^2)$ with $\sigma_2^2 = 1$.

Note that the settings imply different ratios of coefficient-to-coefficient-to-noise variations. In Setting 1, Scenario (1), we have a ratio 36:9:0.5, whereas in Setting 2, Scenario (1) we have 16:9:1. Apart from standard Gaussian errors, we also consider ‘‘fat tailed’’ errors in Scenario (2), heteroscedastic in (3) and skewed errors in (4). We study the performance of the algorithms for three sample sizes: (i) small $n = 20$, $p = 100$; (ii) medium $n = 50$, $p = 150$; (iii) large $n = 100$, $p = 200$.

For every combination of parameters we repeat the simulations 500 times and record the mean computing times, the mean of the average mean squared error (MSE), its standard deviation, and convergence ratio for each algorithm. We label the run of the algorithm as unconverged whenever after 30 iterations and 50 restarts from a random starting point the algorithms fail to converge.

We compare computational times and MSEs of the three methods TopDown (TD), BottomUp (BUP) and PrincipalExpectile (PEC) in the Appendix. In general, PEC is the fastest but it has a lower convergence rate than TopDown (TD) and BottomUp (BUP). From the MSEs, we conclude that whenever the error distribution is fat-tailed or skewed, or for small samples, PEC is likely to produce more reliable results in terms of its MSE, whereas for errors close to normal and moderate or large samples, TD is likely to produce smaller MSEs.

7. Empirical study

We apply the proposed algorithms to three different datasets. In Section 7.1 we investigate the fMRI data from the Risk Perception in Investment Decisions (RPID) study. Since the technical details of the experiment are complex and beyond the scope of this research, we provide only a short summary of the experiment and refer the reader to [16–18] for more details about the experiment and to [1] for the analysis of fMRI data in general. In Section 7.2 we analyze the daily temperature dataset over multiple Chinese stations. Finally, we perform PEC-analysis on the child growth data in Section 7.3.

In these empirical studies, we keep the multivariate data in their original form. We also note that a complete data analysis using PEC requires one to address many issues common to classical PCA such as smoothing, choice of number of components and cross-validation of estimated parameters, to name a few; see the classic monographs [20, 21]. In-depth analysis of these steps is beyond the scope of the present work. Rather, the results in this section are meant to demonstrate how PECs can be applied and interpreted in various settings.

7.1. Application to fMRI data

Risk Perception in Investment Decisions (RPID) Study performed an experiment over 19 individuals. Each participant was asked 256 investment questions, where past returns were presented and participants had to make a choice whether they would invest in a bond with 5% fixed return or the displayed investment. Individual responses reflect the risk attitude of every participant. Following the common Markowitz mean-variance approach, one can evaluate the this risk attitude (for details, see [18]) and assign the corresponding values between -0.1 and 1.1 reflecting individual risk perception. We show the values in the right panel of Figure 2 on vertical axes. Higher values represent the higher

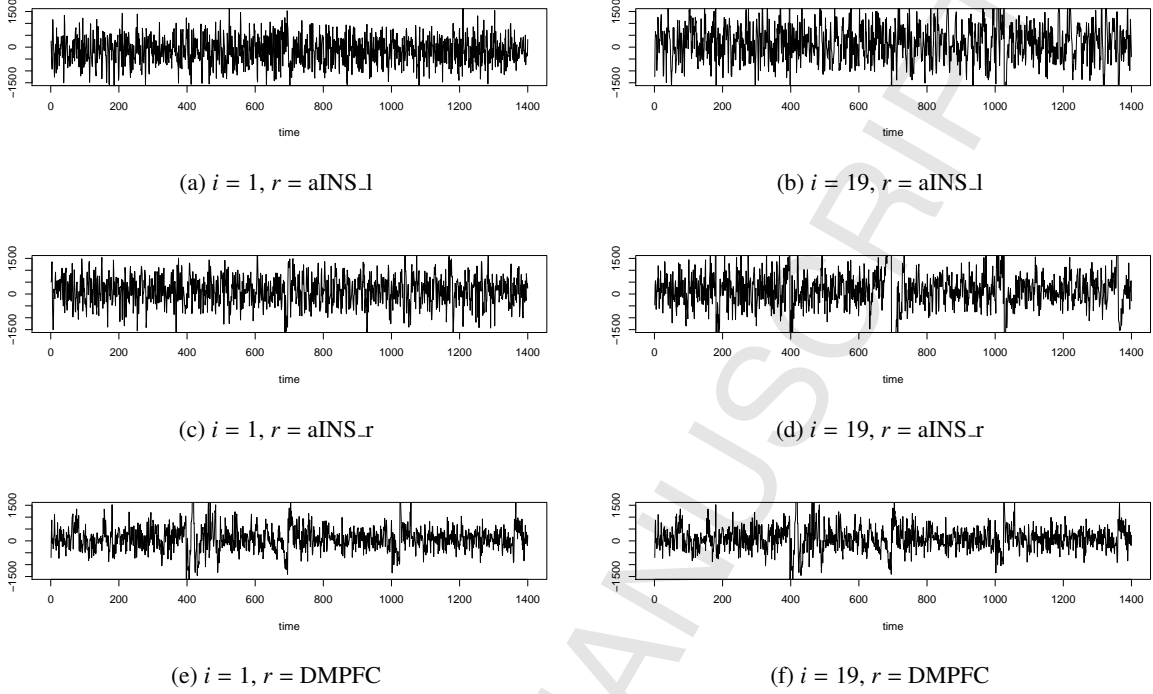


Figure 1: Loadings for the first principal expectile component for active regions of individual No 1 (left) and No 19 (right).

risk aversion. Individual No 19 is considered as the most risk seeking and individual No 1 as the most risk averse participant in the sample.

The aim of the experiment was to study if individual risk perception can be interpreted and recovered by brain activities. With functional magnetic resonance imaging (fMRI) one can measure such neural activity by the blood oxygen level-dependent (BOLD) signal.

Regarding the settings of our dataset, scans of voxels were taken every two seconds, resulting in high-dimensional data for each individual. Majer et al. [16] identified three brain regions (clusters) which are activated during the experiment: the anterior insula (aINS; left and right) and the dorsomedial prefrontal cortex (DMPFC). From a statistical point of view the scan of all voxels in certain brain areas can be considered as a very noisy multi-dimensional time series of round 300–400 voxels for every individual. In order to capture the variability in these series of every region, we use principal expectile components.

Following the notation from Section 4, denote $Y_t^{(r)}$ the response, $N^{(r)}$ -dimensional vector, obtained at specific region, $r = \text{aINS}_{left}, \text{aINS}_{right}, \text{DMPFC}$ at time $t \in \{1, \dots, 1400\}$, where $N^{(r)}$ is a number of voxels in a specific region r . Further let $\phi_{\tau, (r)}^k$ be the k th principal expectile component (PEC) at level τ and $\psi_{\tau, (r)}^k$ be the corresponding projections, also known as loadings. PECs provide us with dimension reduction; each region dynamics is now captured by univariate time series of loadings. The loadings of all three regions for individuals No 1 and No 19 at level $\tau = 0.6$ are presented in Figure 1.

Since the response function usually achieves its peak only shortly after stimulus, i.e., portfolio question, we focus on average loadings after stimulus. The average loadings of three active regions are considered as the regressors for explanation of this risk attitude. For comparison with the results of previous work, we follow Majer et al. [16] and model the relationship of the risk attitude att_i and brain reactions via linear regression, which provides the simplest, yet quite accurate, comparison. More precisely for individuals $i \in \{1, \dots, 19\}$ and any fixed τ -level, we have

$$att_i = \beta_0 + \sum_{k=1,2, r=\text{aINS}_l, \text{aINS}_r, \text{DMPFC}} \beta_{k,r} \overline{\psi_{\tau, (r), i}^k} + \varepsilon_i. \quad (10)$$

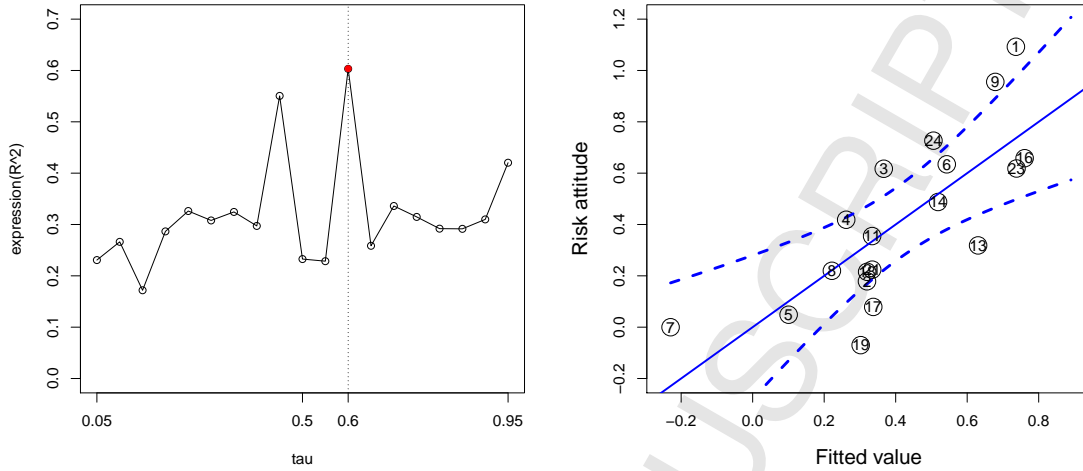


Figure 2: Coefficients of determination for all considered τ -level (left) and the regression fit for model (10) at $\tau = 0.6$ (right). The horizontal axis represents \hat{att}_i , the best linear combination of regressors $\psi_{\tau,(r),i}^k$.

We performed the PrincipalExpectile algorithm for $\tau \in \{0.05, 0.1, \dots, 0.9, 0.95\}$. It is interesting to see that the best result with respect to coefficient of determination R^2 is obtained not for $\tau = 0.5$, but rather for $\tau = 0.6$. We report the coefficients of determination for all considered τ -level in Figure 2, together with the regression fit for model (10) at level $\tau = 0.6$. The use of $\tau = 0.6$ over-performs the traditional use of $\tau = 0.5$.

7.2. Application to Chinese Weather Data

We apply the algorithms BottomUp, TopDown and PrincipalExpectile to Chinese temperature data using daily average temperature data of 159 weather stations in mainland China for the years 1957 to 2009 provided by the Chinese Meteorological Administration via its website. We did not pre-smooth the data. The original data averaged over years for every country are presented in Figure 3.

We run the algorithms to estimate principal expectile components for the weather stations at each of the τ -levels 10%, 50% and 90% with respect to days of a year from 1 to 365. Our analysis for the 50% expectile corresponds to the classical PCA. We estimate the first two principal component functions. The estimation results of the three proposed algorithms are rather similar. In Figure 4, we present the estimated principal component functions for $\tau = 0.1$ and $\tau = 0.9$.

We see that the three algorithms give really similar results. However, one can be more interested in differences between the τ levels. Thus in Figure 5 we show the differences of PEC component at level $\tau = 0.9$ (red), $\tau = 0.1$ (blue) respectively, and PEC component at level $\tau = 0.5$, which corresponds to the ordinary principal component, i.e., $\phi_{0.9}^k - \phi_{0.5}^k$, respectively $\phi_{0.1}^k - \phi_{0.5}^k$. We observe that both components differ from the ordinary principal component. Moreover, we plot also differences for $\tau = 0.8$ (dashed gray) and $\tau = 0.7$ (solid gray) to show that in case of the second component, the difference increases with higher levels of τ .

The obtained first and second components indicate the changes in the temperature distribution from lighter to heavier tails and the other way around within a typical year. A positive score on the first component would mean heavier than average tails in winter and lighter than average tails during the rest of the year. Similarly, a positive score on the second component would indicate lighter than average tails of the temperature distribution in winter months, and heavier in summer.

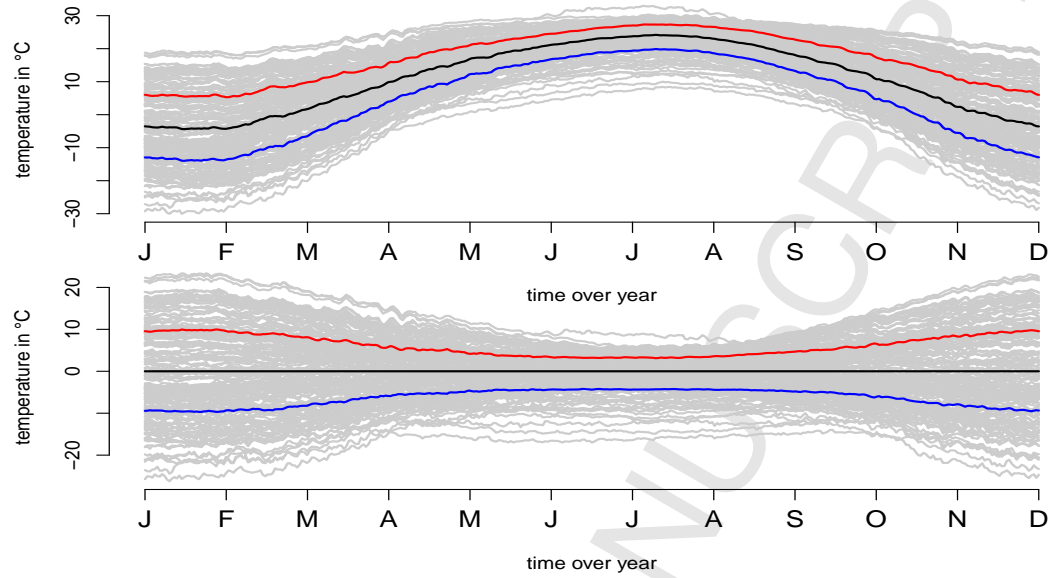


Figure 3: Observed averaged daily temperature on 159 stations (upper panel) and decentred data (lower panel) with expectiles for level $\tau \in \{0.9, 0.5, 0.1\}$.

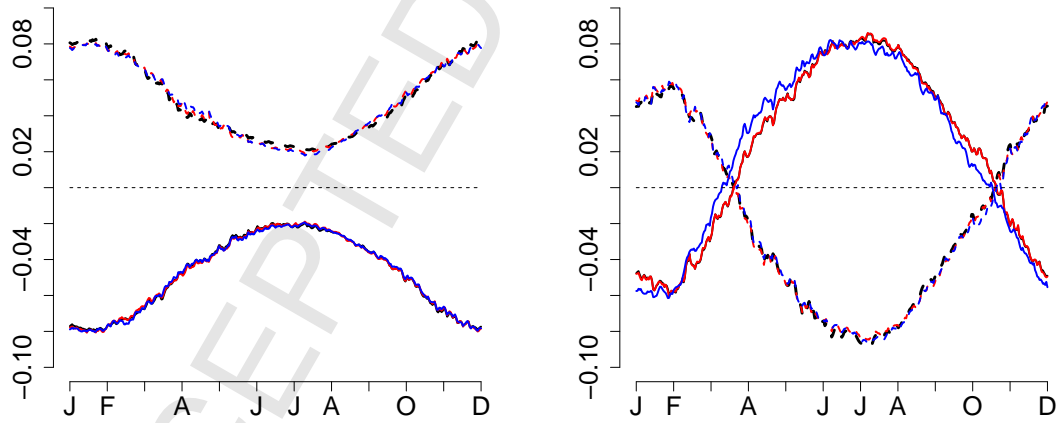


Figure 4: The estimated first PEC (left) and second PEC (right) for $\tau = 0.1$ (dashed) and $\tau = 0.9$ (solid, multiplied by -1) computed with three proposed algorithms TopDown (red), BottomUp (black) and PrincipeExpectile (blue).

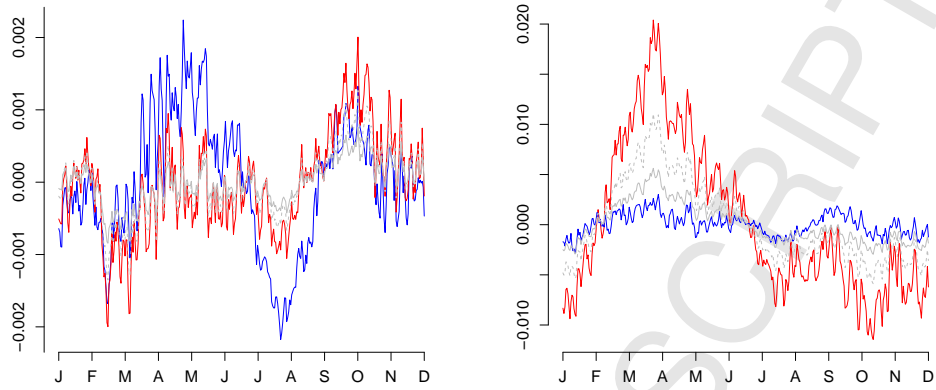


Figure 5: The differences of estimated PECs for $\tau = 0.1$ (blue) and $\tau = 0.9$ (red, multiplied by -1) from estimated PEC for $\tau = 0.5$, computed with PrincipleExpectile algorithm. Differences for the first component are shown on the left, and for the second component on the right.

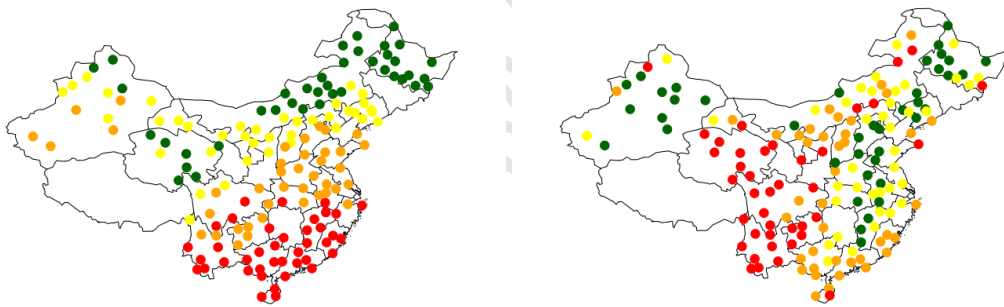


Figure 6: The scores for first (left) and second (right) PECs computed by Principle expectile algorithm for $\tau = 0.9$.

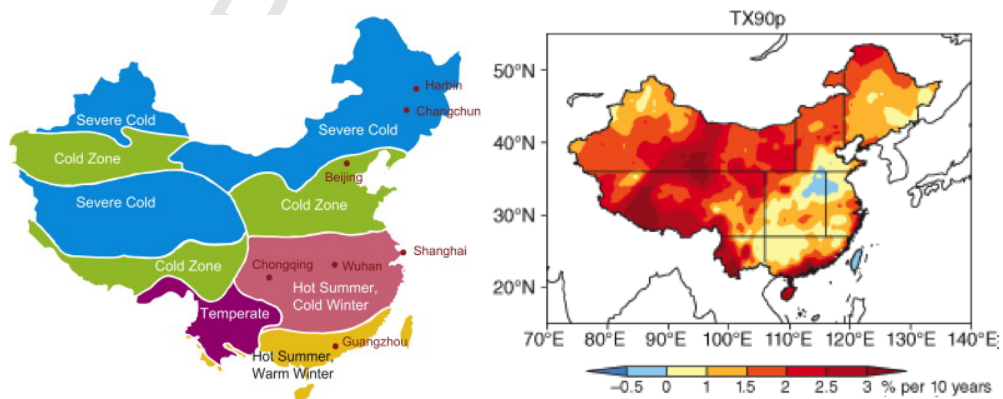


Figure 7: Map of Chinese climate zones by Gao et al. [8] (left) and distribution of trends in temperature percentile index TX90p for the period 1961–2010 by Zhou et al. [27] (right).

For better interpretation we also provide the map of scores in Figure 6. The scores of the first principal expectile component correspond to the climate regions; see Figure 7, explaining the short-term periodic behavior precisely. The scores of the second principal expectile component correspond to the increasing trend in extremes observed for different areas, shown in Figure 7 via temperature index TX90p. TX90p - Warm days indicator is a percentage of time when daily max temperature is higher than 90th percentile. It is one of the 27 core indicators for temperature and perspiration recommended by the World Meteorological Organization - Expert Team on Climate Change Detection and Indices (WMO - ETCCDI); see [13]. The scores of the second component do not necessary coincide with the climate regions but with areas of TX90p index, which explains more long-term behavior and trend.

7.3. Application to Berkeley Growth Data

The Berkeley Growth Study data consist of the height measurements of 39 boys and 54 girls. Children were monitored until the age of 18 and the observations are not equally spaced. Four measurements were taken while the child is one year old, followed by annual measurements from two to eight years and later the height was measured biannually. Data are publically available as in R-package `fda`; see Ramsay et al. [22].

Introductory analysis of such growth curves was already presented in Ramsay and Silverman [20], where the differences between the boy and girl curves were considered. Later Yu et al. [25] showed the application of functional principal component analysis as a tool for the modeling and prediction of growth data. We enlarge the analysis for all τ -levels and computed the principal expectile components for Berkeley growth data with respect to the gender as well.

In Figure 8 we present the first (upper row) and second (bottom row) Principal Expectile Components computed for all children (right column) as well as the subsets of girls (left column) and boys (middle column). In order to point out the difference between the τ -levels, we present PECs for τ equal to 0.05 (blue), 0.95 (red) and τ equal to 0.5 (black) that correspond to the traditional principal components.

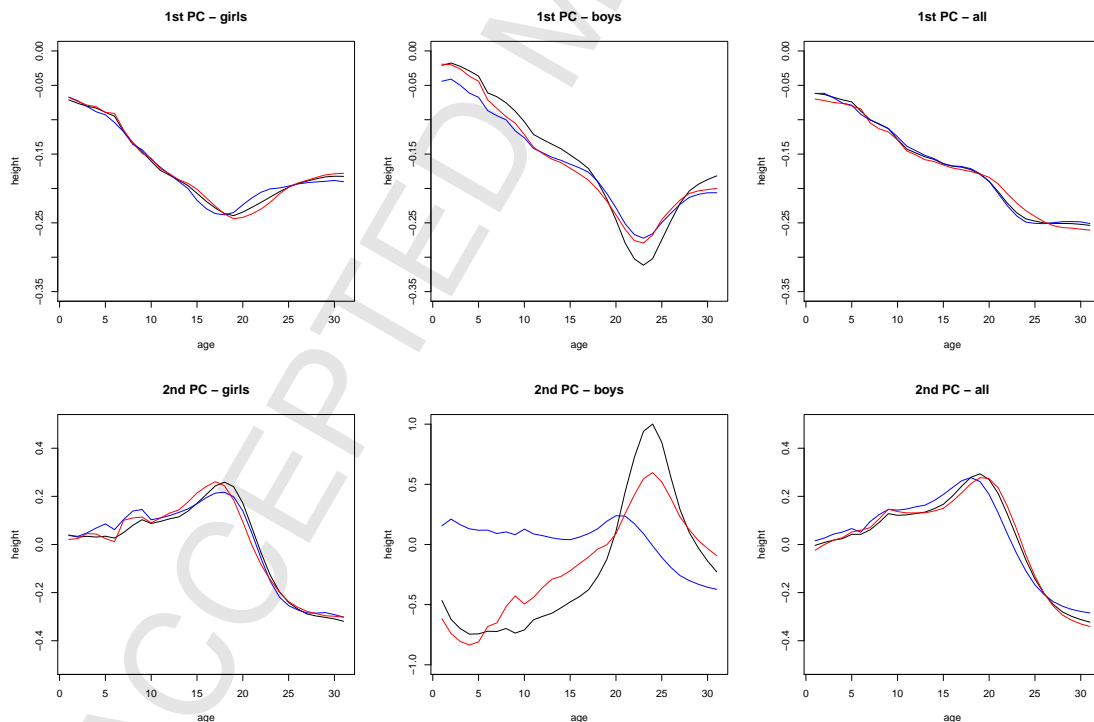


Figure 8: First PEC (first row) and second PEC (second row) for Berkeley Growth data with respect to gender: girls (right), boys (middle), all together (left).

The different pattern for girls and boys should be definitely taken into consideration. We also see that in case of boys compared to the mean (black curve), there are less variations in both the 95th (red curve) and the 5th (blue curve) expectile around the 20–25 age range. This is especially true for the second principal component at τ -level 0.5, suggesting that in the bottom 5th percentile, for the first 20 years, the variation around the 5th expectile can be mostly explained by the first PEC.

8. Summary

We proposed two definitions of principal components in an asymmetric norm and provided consistent algorithms based on iterative least squares. We derived the upper bounds on their convergence time as well as other useful properties of the resulting principal components in an asymmetric norm.

The algorithms TopDown and BottomUp minimize the projection error in a τ -asymmetric norm, and PrincipalExpectile algorithm maximizes the τ -variance of the low-dimensional projection. The latter algorithm was shown to share ‘nice’ properties of PCA such as invariance under translations and changes of basis. Moreover, it coincides with classical PCA for elliptically symmetric distributions. In simulations, PrincipalExpectile and TopDown both have very satisfactory performance in terms of MSE. From a theoretical viewpoint TD is preferable because the estimated subspace is optimal in the desired norm. As shown in experiments, TD outperforms BUP both in computation time and mean-squared error of the estimates. From a computational viewpoint, PE is faster and significantly easier to implement. In addition, it shows robustness to ‘fat-tails’ and skewed data. Therefore, we see that these methods are complementary. For medium-sized data, we would advocate TD. For larger data where efficiency may be prioritized, we would recommend PE.

We applied the algorithm to fMRI data to analyze the possibility of better explanation of individual risk attitude by brain reactions. We have shown that one can achieve better results with a help of higher τ -level rather than by commonly used $\tau = 0.5$. We also applied the algorithms to a Chinese weather dataset with a view to analyzing weather extremes and long-term behavior. Analogously to principal components in [20], we estimated the first two principal expectile component functions of the temperature as functions of days over a year. The resulting component functions indicate relative changes in the tails of the temperature distribution from light to heavier and vice versa. Our results clarify the meaning of the first component as a seasonal component explaining short-term variance of climate areas, while the second component corresponds to the long-term changes. Last, we performed the PrincipalExpectile algorithm for Berkeley Growth data measurements of children’s height. We observed that there are variations in heights across ages that also differ with respect to τ -level. Mainly for boys the variation is explained by the first two components and the variation occurs in ages 20–25. Among the 95th expectile (tall boys), we see the same pattern of variation as for the mean, i.e., 50th expectile, but in smaller magnitude. Among the 5th expectile (short boys), we see the same pattern of variation, but even smaller magnitude, almost flat.

The proposed algorithms appear to be a good way to study multivariate data extremes. They are easy to compute, relatively fast and their results are easy to interpret.

Appendix

Table 1 and 2 show the runtimes of the simulations. PrincipalExpectile (PEC) is the fastest algorithm but it has a relative low convergence rate: for all sample sizes only around 80% of the algorithm runs converged. In 20% cases the algorithm kept iterating between two sets of weights which possibly indicates an adverse sample geometry, i.e., that two eigenvalues of the scaled covariance matrix are too close to each other. TD, on the contrary, converges almost always in medium and large sample sizes.

Table 1: Average time in seconds for convergence of the algorithms by 500 simulations

Sample τ /sec	Small			Medium			Large		
	BUP	TD	PEC	BUP	TD	PEC	BUP	TD	PEC
0.900	1.15	0.70	0.57	2.87	1.59	1.39	7.44	4.02	2.71
0.950	1.52	1.13	0.55	3.94	2.68	1.57	10.34	6.88	3.03
0.975	2.47	2.32	0.56	5.49	4.62	1.56	14.37	10.96	3.54

Table 2: Nonconvergence rates of the algorithms by 500 simulation runs

Sample τ /rate	Small			Medium			Large		
	BUP	TD	PEC	BUP	TD	PEC	BUP	TD	PEC
0.900	0.11	0.00	0.24	0.07	0.00	0.23	0.03	0.00	0.20
0.950	0.17	0.00	0.22	0.13	0.00	0.26	0.11	0.00	0.21
0.975	0.25	0.03	0.21	0.22	0.01	0.25	0.22	0.00	0.24

The results in terms of MSEs are presented in Tables 3–4 for both simulation settings. Only the magnitude of the average MSE differs; there is no substantial qualitative difference in relative performance of the algorithms. BUP performs the worst among the three algorithms in terms of its MSE in all scenarios. TD and PEC are comparable in terms of their MSEs. PEC is robust against skewness and fat tails in the error distribution since it produces the lowest MSEs in Scenarios (2) and (4). Yet TD tends to outperform PEC slightly in medium and large samples for errors close to iid normal or normal heteroscedastic; for small sample sizes PEC outperforms TD in all scenarios but (5).

Figures 9 – 10 illustrate the difference in the quality of component estimation for the 95% expectile when coefficient-to-coefficient-to-noise variation ratio changes (setting 1 versus setting 2, respectively). The results are shown for the error scenario (1) and small sample size. We observe that as the ratio changes from 36:9:0.5 (Setting 1, Figure 9) to 16:9:1 (Setting 2, Figure 10) the variability of the estimators of both component functions increases. The overall mean of the estimators remains very close to the true component functions.

Finally, we comment on the run time. The computation time always includes the initialization step. For these simulations, for PEC, we use random weight initialization (and not incremental initialization). This is to ensure a fair comparison with TD and BUP. Should one compute with incremental initialization, then the runtime should be multiplied by the number of increments in the sequence. However, this also means that the algorithm outputs all PECs for a large range of τ , and therefore it is not a fair comparison with TD/BUP (which only output for one specific τ).

When τ is very close to 1 or 0, which observation is assigned the rarer weight has a large influence on the estimator. The objective function $J(U, V, W)$ becomes very sensitive to weight changes, and the Newton–Raphson algorithm can take longer to converge. The three algorithms run more slowly when τ is more extreme. PEC is faster compared to TD and BUP. While all of these algorithms are iterative and involve solving a family of optimization problems (one for each weight w), TD/BUP does this internal optimization through an iterative coordinate descent procedure (embedded in LAWS). In comparison, PEC just needs to solve for the largest eigenvector of a particular matrix (Step 10), for which many efficient algorithms with efficient implementations exist. Therefore, PEC runs much faster.

Table 3: Average MSE and its standard deviation in brackets by 500 simulation runs for the simulation Setting 1.

scenario	τ	$n = 20, p = 100$			$n = 50, p = 150$			$n = 100, p = 200$		
		BUP	TD	PEC	BUP	TD	PEC	BUP	TD	PEC
(1)	0.900	0.2762 (0.1997)	0.1216 (0.0097)	0.1123 (0.0111)	0.1339 (0.1099)	0.0538 (0.0033)	0.0632 (0.0029)	0.0698 (0.0552)	0.0297 (0.0015)	0.0459 (0.0014)
	0.950	0.3619 (0.2199)	0.1568 (0.0123)	0.1334 (0.0181)	0.2323 (0.2076)	0.0705 (0.0045)	0.0727 (0.0044)	0.1312 (0.1415)	0.0394 (0.0020)	0.051 (0.0019)
	0.975	0.5064 (0.2977)	0.2053 (0.0154)	0.1601 (0.0276)	0.3583 (0.2989)	0.0944 (0.0060)	0.0874 (0.0075)	0.2157 (0.2314)	0.0536 (0.0027)	0.0594 (0.0035)
(2)	0.900	0.7092 (0.2382)	0.5421 (0.1096)	0.3147 (0.0685)	0.3382 (0.1223)	0.2714 (0.0727)	0.1494 (0.0117)	0.1866 (0.0522)	0.1548 (0.0217)	0.0932 (0.0050)
	0.950	1.105 (0.4453)	0.7847 (0.1646)	0.3854 (0.0988)	0.5789 (0.2664)	0.4440 (0.1675)	0.1819 (0.0192)	0.3316 (0.1144)	0.2680 (0.0575)	0.1101 (0.0075)
	0.975	1.6066 (0.7968)	1.1158 (0.2106)	0.4709 (0.1413)	0.9956 (0.6936)	0.7033 (0.2629)	0.2309 (0.0341)	0.5780 (0.2227)	0.4641 (0.1175)	0.1358 (0.0132)
(3)	0.900	0.4146 (0.2413)	0.2300 (0.0195)	0.2215 (0.0236)	0.1829 (0.1070)	0.1019 (0.0065)	0.1270 (0.0066)	0.0962 (0.0510)	0.0562 (0.0029)	0.0942 (0.0032)
	0.950	0.6261 (0.6313)	0.2966 (0.0246)	0.2792 (0.0369)	0.3538 (1.1684)	0.1335 (0.0088)	0.1622 (0.0097)	0.1603 (0.1135)	0.0746 (0.0039)	0.1208 (0.0045)
	0.975	0.8051 (0.4516)	0.3885 (0.0312)	0.3516 (0.0527)	0.4879 (0.3736)	0.1789 (0.0118)	0.2109 (0.0167)	0.2665 (0.2234)	0.1016 (0.0052)	0.1568 (0.0077)
(4)	0.900	0.9162 (0.2432)	0.8041 (0.1532)	0.2226 (0.0588)	0.4854 (0.1093)	0.4510 (0.0597)	0.1077 (0.0089)	0.2876 (0.0498)	0.2763 (0.0247)	0.0697 (0.0042)
	0.950	1.4972 (0.4494)	1.2869 (0.2337)	0.2725 (0.0713)	0.9127 (0.4895)	0.8092 (0.1187)	0.1296 (0.0142)	0.5585 (0.1595)	0.5280 (0.0554)	0.0812 (0.0069)
	0.975	2.3371 (1.0034)	1.9727 (0.2835)	0.3331 (0.0979)	1.5522 (0.7483)	1.3387 (0.1999)	0.1629 (0.0248)	1.2223 (1.4707)	0.9421 (0.1110)	0.0995 (0.0117)
(5)	0.900	0.0343 (0.0224)	0.0091 (0.0007)	0.0368 (0.0013)	0.0298 (0.0261)	0.0038 (0.0002)	0.0315 (0.0004)	0.0244 (0.0238)	0.0021 (0.0001)	0.0296 (0.0002)
	0.950	0.1225 (1.1145)	0.0110 (0.0008)	0.0409 (0.0020)	0.0351 (0.0398)	0.0044 (0.0003)	0.0345 (0.0007)	0.0285 (0.0254)	0.0023 (0.0004)	0.0322 (0.0004)
	0.975	0.0776 (0.3266)	0.0135 (0.0011)	0.0474 (0.0034)	0.0455 (0.0658)	0.0052 (0.0003)	0.0397 (0.0012)	0.0360 (0.0309)	0.0027 (0.0001)	0.0366 (0.0006)

Table 4: Average MSE and its standard deviation in brackets by 500 simulation runs for the simulation Setting 2.

scenario	τ	$n = 20, p = 100$			$n = 50, p = 150$			$n = 100, p = 200$		
		BUP	TD	PEC	BUP	TD	PEC	BUP	TD	PEC
(1)	0.900	0.4484 (0.2671)	0.2436 (0.0195)	0.1988 (0.0238)	0.2053 (0.1273)	0.1077 (0.0066)	0.1002 (0.0058)	0.1109 (0.0924)	0.0595 (0.0030)	0.0660 (0.0027)
	0.950	0.7021 (0.4611)	0.314 (0.0246)	0.2418 (0.0386)	0.3681 (0.3066)	0.1411 (0.0090)	0.119 (0.0091)	0.2075 (0.2346)	0.0788 (0.0039)	0.0761 (0.0039)
	0.975	0.9218 (0.5578)	0.4116 (0.0312)	0.2945 (0.0546)	0.5957 (0.4751)	0.1890 (0.0121)	0.1483 (0.0152)	0.3364 (0.3565)	0.1074 (0.0053)	0.0925 (0.0067)
(2)	0.900	0.7424 (0.2933)	0.5427 (0.1099)	0.3186 (0.0762)	0.3560 (0.1695)	0.2716 (0.0728)	0.1502 (0.0123)	0.2047 (0.1886)	0.1549 (0.0218)	0.0935 (0.0050)
	0.950	1.1483 (0.5078)	0.7855 (0.1643)	0.3920 (0.1096)	0.6656 (0.6719)	0.4437 (0.1658)	0.1832 (0.0185)	0.3805 (0.3563)	0.2684 (0.0581)	0.1103 (0.0075)
	0.975	1.7083 (0.8614)	1.1095 (0.1744)	0.4805 (0.1493)	1.1714 (0.9716)	0.7048 (0.2652)	0.2342 (0.0323)	0.6974 (0.5981)	0.4648 (0.1192)	0.1368 (0.0126)
(3)	0.900	0.6616 (0.2625)	0.4613 (0.0392)	0.4093 (0.0486)	0.2993 (0.1163)	0.2041 (0.0131)	0.2200 (0.0134)	0.1684 (0.1880)	0.1126 (0.0058)	0.1540 (0.0066)
	0.950	1.0027 (0.5055)	0.5948 (0.0495)	0.5229 (0.0802)	0.4979 (0.3671)	0.2675 (0.0177)	0.2875 (0.0215)	0.3031 (0.4360)	0.1494 (0.0077)	0.2042 (0.0090)
	0.975	1.465 (0.8018)	0.7811 (0.0627)	0.6719 (0.1154)	0.8605 (0.8004)	0.3587 (0.0237)	0.3831 (0.0338)	0.5173 (0.6708)	0.2036 (0.0103)	0.2724 (0.0156)
(4)	0.900	5.4073 (2.1503)	5.2042 (1.9812)	1.0318 (0.9534)	3.3226 (1.1548)	3.2871 (1.0106)	0.4075 (0.1258)	2.0358 (0.6044)	2.0686 (0.5259)	0.2295 (0.1632)
	0.950	8.7171 (2.8223)	8.0696 (2.3418)	1.4256 (1.4550)	6.5227 (1.9576)	6.2094 (1.5846)	0.5143 (0.1540)	4.5541 (1.4193)	4.4481 (1.0287)	0.2939 (0.3150)
	0.975	13.419 (5.1223)	11.635 (1.6721)	2.0054 (2.2733)	11.202 (4.0968)	9.8804 (1.8550)	0.7372 (0.5037)	8.9280 (2.4679)	8.3663 (2.7240)	0.3889 (0.3161)
(5)	0.900	0.1135 (0.0755)	0.0365 (0.0027)	0.0572 (0.0041)	0.0923 (0.0878)	0.0153 (0.0009)	0.0394 (0.0011)	0.0561 (0.0628)	0.0083 (0.0004)	0.0333 (0.0005)
	0.950	0.1430 (0.1214)	0.0440 (0.0034)	0.0651 (0.0060)	0.1197 (0.1033)	0.0177 (0.0010)	0.0434 (0.0018)	0.0896 (0.0938)	0.0093 (0.0005)	0.0356 (0.0008)
	0.975	0.2489 (0.6091)	0.0540 (0.0042)	0.0769 (0.0099)	0.1538 (0.1272)	0.0209 (0.0013)	0.0499 (0.0031)	0.1145 (0.1042)	0.0107 (0.0006)	0.0396 (0.0013)

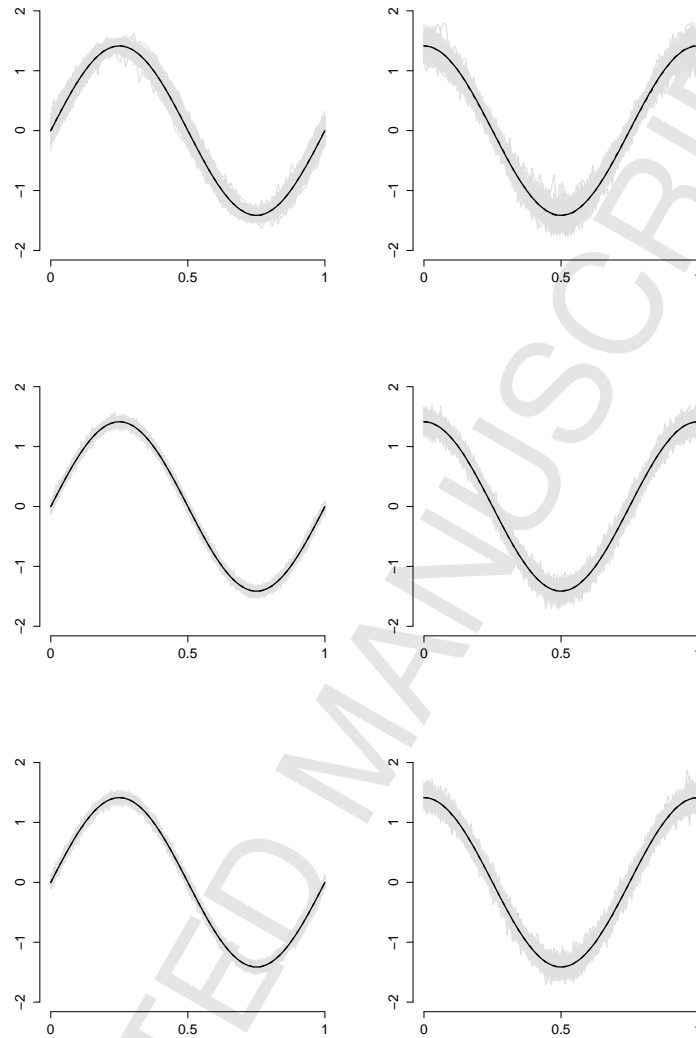


Figure 9: Estimated component functions (solid gray) from 500 simulation runs for simulation Setting 1, Scenario 1 small sample size and 95% expectile. The rows from the top to the bottom show respectively results produced by BUP, TD and PEC. Left panel corresponds to the first component function, right panel – to the second. The true functions are shown as solid black curves. The overall mean across simulation runs is shown as dashed black curve. The later can not be distinguished from the true curve.

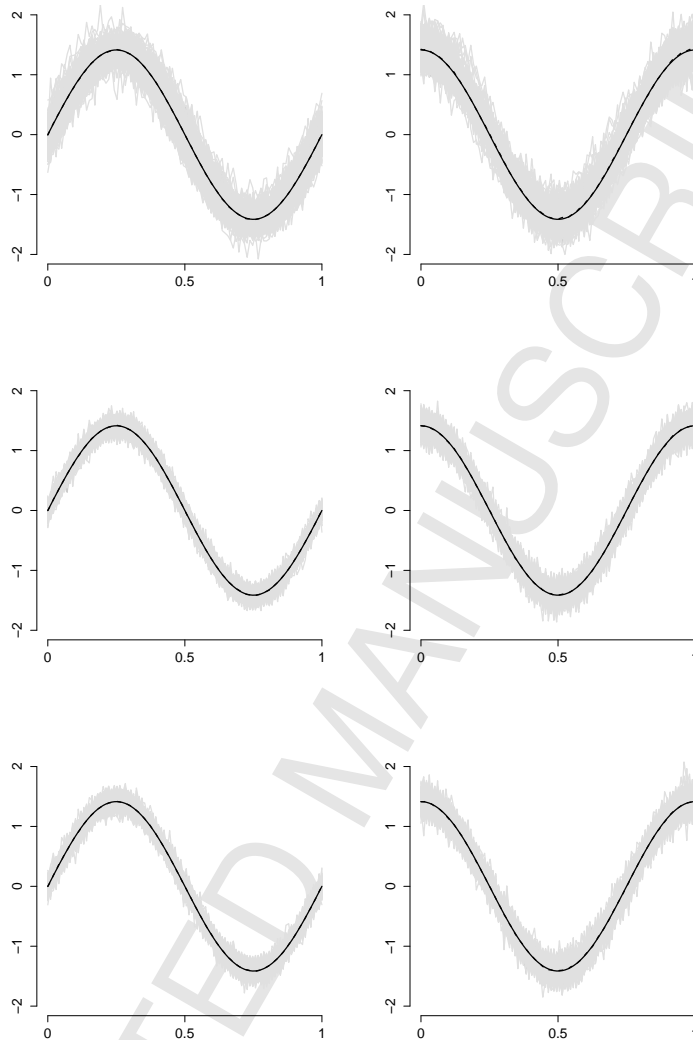


Figure 10: Estimated component functions (gray) from 500 simulation runs for simulation Setting 2, Scenario 1 small sample size and 95% expectile. The rows from the top to the bottom show respectively results produced by BUP, TD and PEC. Left panel corresponds to the first component function, right panel – to the second. The true functions are shown as solid black curves. The overall mean across simulation runs is shown as dashed black curve. The later can not be distinguished from the true curve.

Initialization of weights for PEC algorithm

As the first step of the estimation procedure of PEC algorithm, one has to set up initialization weights. We performed simulations to measure how sensitive the final weight is to initialization. In the simulation below (setup with $n = 20$ and $p = 100$ for Scenario 1, error type 1; see Section 6), we computed PEC for $\tau = 0.9$ with different initialization settings. Each time, we randomly initialized a fraction k of the weights to be 0.9, and the rest to be 0.1. For each value of $k \in \{0.1, \dots, 0.9\}$, we performed 100 runs and plotted the mean squared error of the residual; see Figure 11. It appears that the change the ratio of 0.1s to 0.9s in the initialization step does not affect performance.

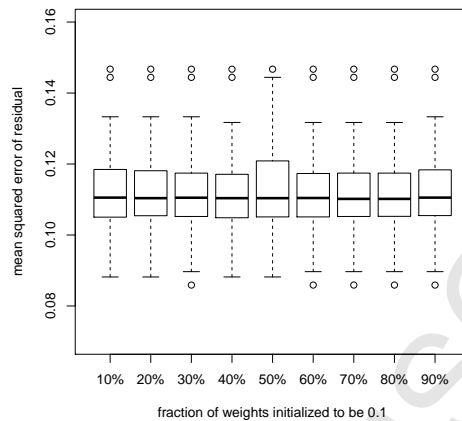


Figure 11: The mean squared error of the residual with randomly initialized a fraction k of the weights from 10% to 90%.

Finally, we compared the random initialization setup above with the incremental algorithm. In the same setup, to compute $\tau = 0.9$, we computed a sequence of PECs for τ from 0.5 to 0.9, in increments of size 0.1, 0.05 and 0.01. With an increment of size 0.1, for instance, this means we computed PEC for $\tau = 0.5$ (which is the usual PCA), used weights from this output to initialize the computation for $\tau = 0.6$, and then again used these weights to initialize the computation for $\tau = 0.7$, and so on. Figure 12 shows the boxplots of 20 simulation runs for each setting of the incremental algorithm. These plots show that the increment sequence does not affect the final result significantly. Furthermore, they largely agree with the random initialization method. In conclusion, either random or incremental initialization works well in practice and yields similar results.

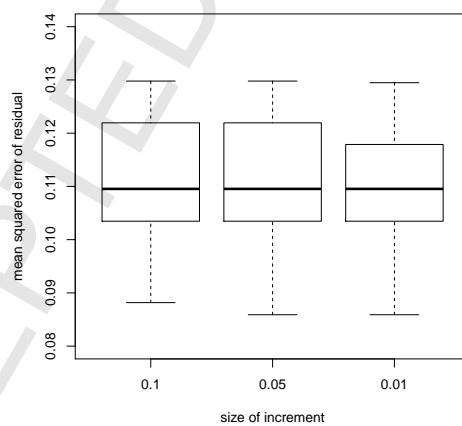


Figure 12: The mean squared error of the residual with incrementally computed algorithm.

References

- [1] F.G. Ashby, *Statistical Analysis of fMRI Data*, The MIT Press, Cambridge, MA, 2011.
- [2] P. Burdejová, W.K. Härdle, P. Kokoszka, Q. Xiong, Change point and trend analysis of annual expectile curves of tropical storms, *Econometrics and Statistics* 1 (2017) 101–117.
- [3] Ş. Cobzaş, *Functional Analysis in Asymmetric Normed Spaces*, Birkhäuser/Springer Basel AG, Basel, 2013.
- [4] J. Damon, J. Marron, Backwards principal component analysis and principal nested relations, *J. Math. Imaging Vis.* 50 (2014) 107–114.

- [5] A. Daouia, S. Girard, G. Stupfler, Estimation of Tail Risk based on Extreme Expectiles, 2016. Working paper.
- [6] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis: Theory and Practice, Springer, New York, 2006.
- [7] R. Fraiman, B. Pateiro-López, Quantiles for finite and infinite dimensional data, *J. Multivariate Anal.* 108 (2012) 1–14.
- [8] Y. Gao, J. Xu, S. Yang, X. Tang, Q. Zhou, J. Ge, T. Xu, R. Levinson, Cool roofs in China: Policy review, building simulations, and proof-of-concept experiments, *Energy Policy* 74 (2014) 190–214.
- [9] M. Guo, L. Zhou, W.K. Härdle, J. Huang, Functional data analysis of generalized regression quantiles, *Statistics and Computing* 25 (2015) 189–202.
- [10] L. Horváth, P. Kokoszka, Inference for Functional Data with Applications, Springer, New York, 2012.
- [11] I.T. Jolliffe, Principal Component Analysis, Springer, New York, 2004.
- [12] P. Jorion, Risk management lessons from long-term capital management, *Europ. Fin. Manag.* 6 (2000) 277–300.
- [13] A.M. Klein Tank, F.W. Zwiers, X. Zhang, Guidelines on Analysis of extremes in a changing climate in support of informed decisions for adaptation, World Meteorological Organization, 2009.
- [14] L. Kong, I. Mizera, Quantile tomography: Using quantiles with multivariate data, *Statist. Sinica* 22 (2012) 1589–1610.
- [15] C.-M. Kuan, J.-H. Yeh, Y.-C. Hsu, Assessing value at risk with care, the conditional autoregressive expectile models, *J. Econometrics* 150 (2009) 261 – 270.
- [16] P. Majer, P.N.C. Mohr, H. R. Heekeren, W.K. Härdle, Portfolio decisions and brain reactions via the cead method, *Psychometrika* (2015) 1–23.
- [17] P.N. Mohr, G. Biele, L.K. Krugel, S.-C. Li, H.R. Heekeren, Neural foundations of risk-return trade-off in investment decisions, *NeuroImage* 49 (2010) 2556 – 2563.
- [18] P.N. Mohr, I.E. Nagel, Neural foundations of risk-return trade-off in investment decisions, *J. Neurosci.* 30 (2010) 7755 – 7757.
- [19] W. Newey, J. Powell, Asymmetric least squares estimation and testing, *Econometrica* (1987) 819–847.
- [20] J.O. Ramsay, B.W. Silverman, Functional Data Analysis, Springer, New York, 2005.
- [21] J.O. Ramsay, B.W. Silverman, Applied Functional Data Analysis: Methods and Case Studies, Springer, New York, 2007.
- [22] J.O. Ramsay, H. Wickham, S. Graves, G. Hooker, FDA: Functional Data Analysis, 2014. R Package Version 2.4.3.
- [23] S. Schnabel, Expectile smoothing: New perspectives on asymmetric least squares — An application to life expectancy, Ph.D. thesis, Universiteit Utrecht, Utrecht, The Netherlands, 2011.
- [24] N. Srebro, T. Jaakkola, Weighted low-rank approximations, In: ICML'03 Proceedings of the Twentieth International Conference on International Conference on Machine Learning, AAAI Press, pp. 720–727.
- [25] H. Yu, H. Xuming, T. Jian, S. Ning-Zhong, Modeling and prediction of children's growth data via functional principal component analysis, *Science in China, Ser. A, Mathematics, Physics, Astronomy / Chinese Academy of Sciences* 52 (2009) 1342–1350.
- [26] L. Zhang, S. Lu, J. Marron, Nested nonnegative cone analysis, *Computat. Statist. Data Anal.* 88 (2015) 100–110.
- [27] B. Zhou, Y. Xu, J. Wu, S. Dong, Y. Shi, Changes in temperature and precipitation extreme indices over China: Analysis of a high-resolution grid dataset, *Internat. J. Climatol.* 36 (2016) 1051–1066.



How to measure the performance of a Collaborative Research Center

Alona Zharova¹  · Janine Tellinger-Rice² · Wolfgang Karl Härdle^{1,3}

Received: 3 March 2018 / Published online: 22 September 2018
© Akadémiai Kiadó, Budapest, Hungary 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 Prügger.

✉ 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 Bergh 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
<i>Relative reception success</i>		
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 Bergh 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
<i>Research activity</i>		
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 ($RA_{SP} = N_{PubSP} * C_{PubSP} / C_{PubCRC}$)	Pastor et al. (2015)
<i>Research productivity</i>		
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)
$ScSP_{Pub}$	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_{Pub}/FC_m)	Abramo and D’Angelo (2011), Abramo et al. (2009)
h	h -index	Hirsch (2005), Bornmann (2013)
<i>Visibility of the CRC</i>		
$AbsC_{Pub}$	Absolute citation count in the light of maximum citation count of a single publication (C_{Pubmax}) and the number of non-cited publications (N_{ncPub})	Wissenschaftsrat (2012)
<i>Reputation</i>		
	List of scientific prizes and awards	Zheng and Liu (2015), Wissenschaftsrat (2012)
<i>Professional activity</i>		
	Editorships	Wissenschaftsrat (2012)
	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
$N_{\text{Pub}}/N_{\text{Staff}}$	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)
$NC_{\text{Pub}}/N_{\text{Staff}}$	Relation of the number of citations of publications (N_{Pub}) to the number of research staff (N_{Staff})	Wissenschaftsrat (2012), Lee (2010)
$N_{\text{Costs}}/N_{\text{Staff}}$	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
<i>Promotion of young researchers</i>		
N_{YR}	Total number of positions for young researchers	Wissenschaftsrat (2012)
$N_{\text{Ph.D.}}$	Total number of defended Ph.D.	Wissenschaftsrat (2012), Diem and Wolter (2013), Grözinger and Leusing (2006), Schmoch and Schubert (2009)
$D_{\text{Ph.D.}}$	Average duration of Ph.D. study	Wissenschaftsrat (2004)
$N_{\text{PubPh.D.}}$	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

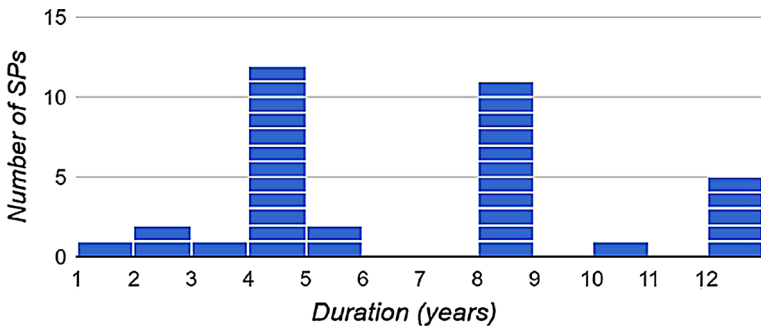


Fig. 1 Distribution of sub-projects (SP) over life span in years

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

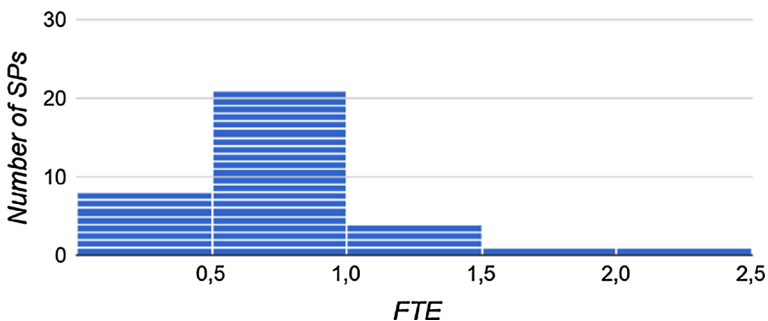


Fig. 2 Distribution of sub-projects according to the number of research staff (in FTE per year)

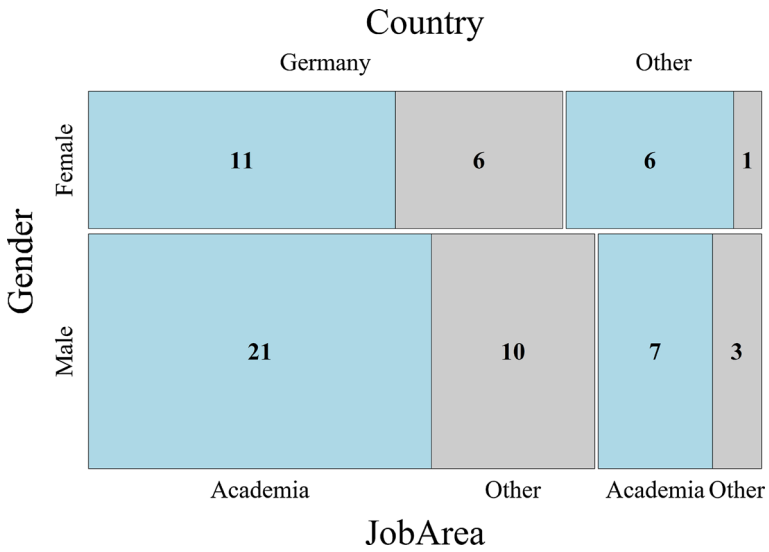


Fig. 4 Mosaic plot of job type, location and gender of 65 CRC members who received their Ph.D. between 2005 and 2016 (as of Dec 2016)

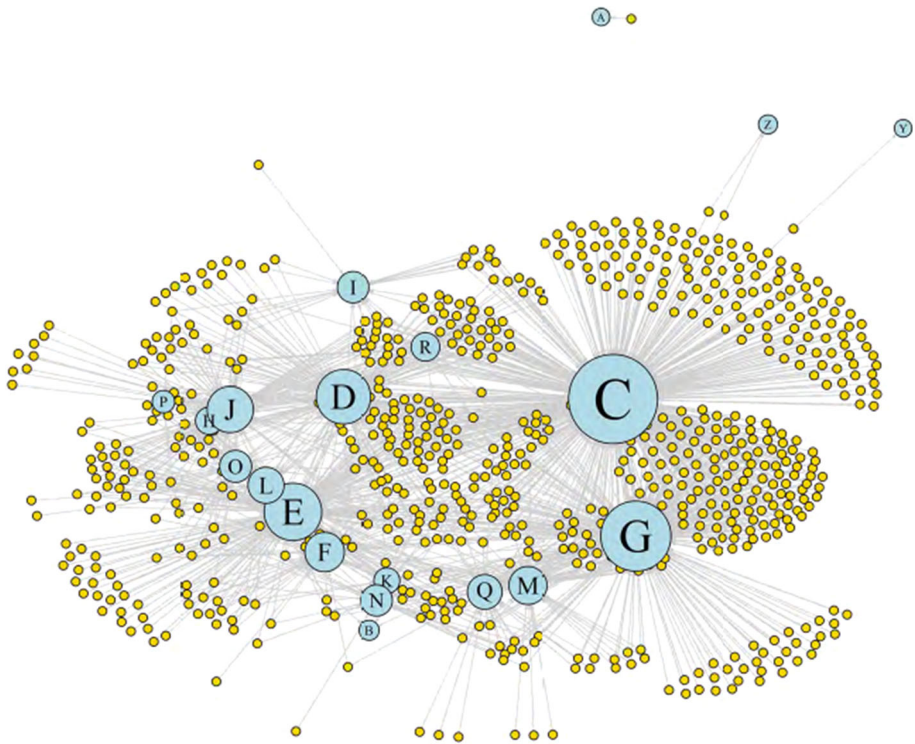


Fig. 5 Network of 760 discussion papers (yellow) and 20 JEL codes (blue) published from 2005 to 2016. (Color figure online)

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

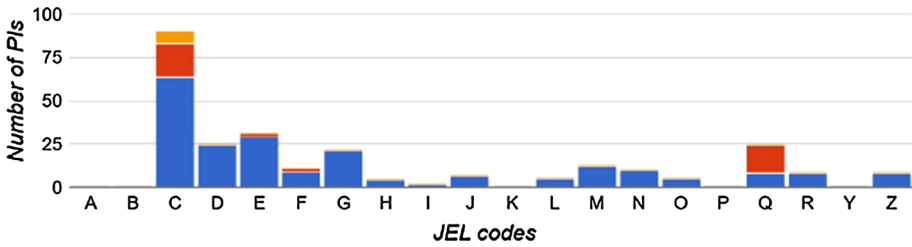


Fig. 6 Cumulative number of PIs (in PI years; full professors—blue, junior professors—red, postdoctoral researchers—orange) from 2005 to 2016 (weighted by the number of research fields and sub-projects) with expertise in corresponding JEL research fields. (Color figure online)

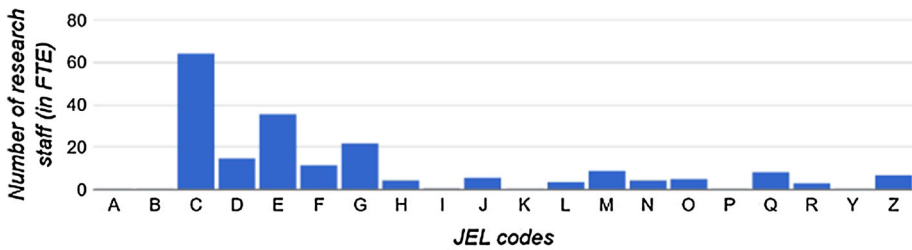


Fig. 7 Cumulative number of research staff in FTE (in staff years; weighted by the number of research fields) from 2005 to 2016 working within corresponding JEL research areas

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):

$$y_i = \beta X_i + u_i, \quad i = 1, \dots, K, \tag{1}$$

where $y_i = (y_{i1}, \dots, y_{iT})^\top$ is a $(1 \times T)$ vector of observations for $t = 1, 2, \dots, T$, $X_i = (x_{i1}^\top, \dots, x_{iT}^\top)^\top$ 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):

$$y_{it} = \beta_1 x_{it1} + \dots + \beta_k x_{itk} + c_i + u_{it}, \quad t = 1, 2, \dots, T, \quad i = 1, 2, \dots, K, \quad (2)$$

where y_{it} includes dependent variables and x_{it} independent variables for individual i at time t , β_1, \dots, β_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):

$$y_{it} = \beta_1 x_{it1} + \dots + \beta_k x_{itk} + \delta_1 + \delta_2 D_2 + \dots + \delta_T D_T + c_i + u_{it}, \quad (3)$$

where D_2, \dots, D_T are time effects and $\delta_1, \dots, \delta_T$ 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:

$$E(y_{it} | x_i, a_i) = a_i \mu(x_{it}, \beta_0), \quad t = 1, 2, \dots, T, \quad (4)$$

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 $\beta_{\log \text{StaffCosts}}$, travel costs $\beta_{\log \text{TravelCosts}}$ 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 $\beta_{\log \text{StaffCosts}}$ and travel costs $\beta_{\log \text{TravelCosts}}$ 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)
$\beta_{\log\text{StaffCosts}}$	1.38** (0.61)	1.62* (0.88)	0.47*** (0.12)	0.43** (0.19)
$\beta_{\log\text{TravelCosts}}$	- 0.94* (0.55)	- 0.34 (0.47)	- 0.22** (0.10)	- 0.04 (0.09)
δ_{2006}	1.61 (1.36)	1.92 (1.61)	0.25 (0.26)	0 (omit.)
δ_{2007}	- 1.20 (1.38)	- 2.55 (2.46)	- 0.30 (0.31)	- 0.98*** (0.25)
δ_{2008}	- 0.95 (1.30)	- 2.03 (2.10)	- 0.23 (0.32)	- 0.97*** (0.36)
δ_{2009}	- 2.05* (1.13)	- 3.16 (1.98)	- 0.54* (0.33)	- 1.20*** (0.23)
δ_{2010}	- 1.93* (1.14)	- 2.13 (2.68)	- 0.51* (0.30)	- 1.03*** (0.31)
δ_{2011}	1.10 (0.70)	0 (omit.)	0.33* (0.20)	0 (omit.)
δ_{2012}	- 2.79* (1.46)	- 3.60* (1.78)	- 0.71** (0.34)	- 1.90*** (0.20)
δ_{2013}	- 2.98** (1.30)	- 3.18 (2.52)	- 0.80** (0.32)	- 1.32*** (0.41)
δ_{2014}	- 1.36 (0.95)	- 1.73 (1.61)	- 0.44 (0.27)	- 0.99*** (0.37)
δ_{2015}	- 2.55** (1.17)	- 1.90 (1.77)	- 0.74** (0.33)	- 1.02*** (0.31)
δ_{2016}	- 0.30 (1.79)	0 (omit.)	- 0.31 (0.36)	- 0.69* (0.41)
$const$	- 2.37 (5.29)	0.05 (10.09)		
$\beta_{nDP_{t-1}}$		0.02 (0.16)		- 0.01* (0.03)
$\beta_{\log\text{StaffCosts}_{t-1}}$		- 0.66 (0.59)		- 0.25 (0.23)
$\beta_{\log\text{TravelCosts}_{t-1}}$		- 0.21 (0.58)		- 0.02 (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 nDP 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 $\beta_{\log\text{StaffCosts}}$ 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 $\beta_{\log\text{TravelCosts}}$ 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 $\log\text{TravelCosts}$ 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.

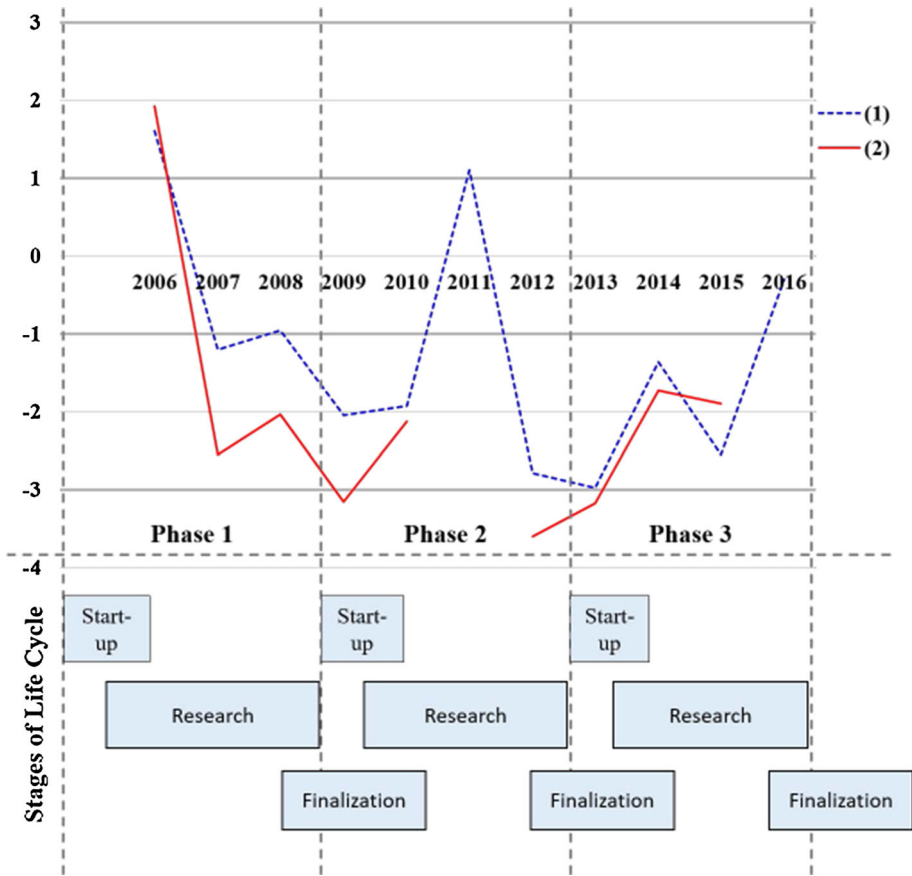


Fig. 8 Estimates of coefficients on the year dummy variables for the time fixed effects (within) regression (models (1) and (2)). The lower part of the figure shows the corresponding stage of the research project life cycle

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

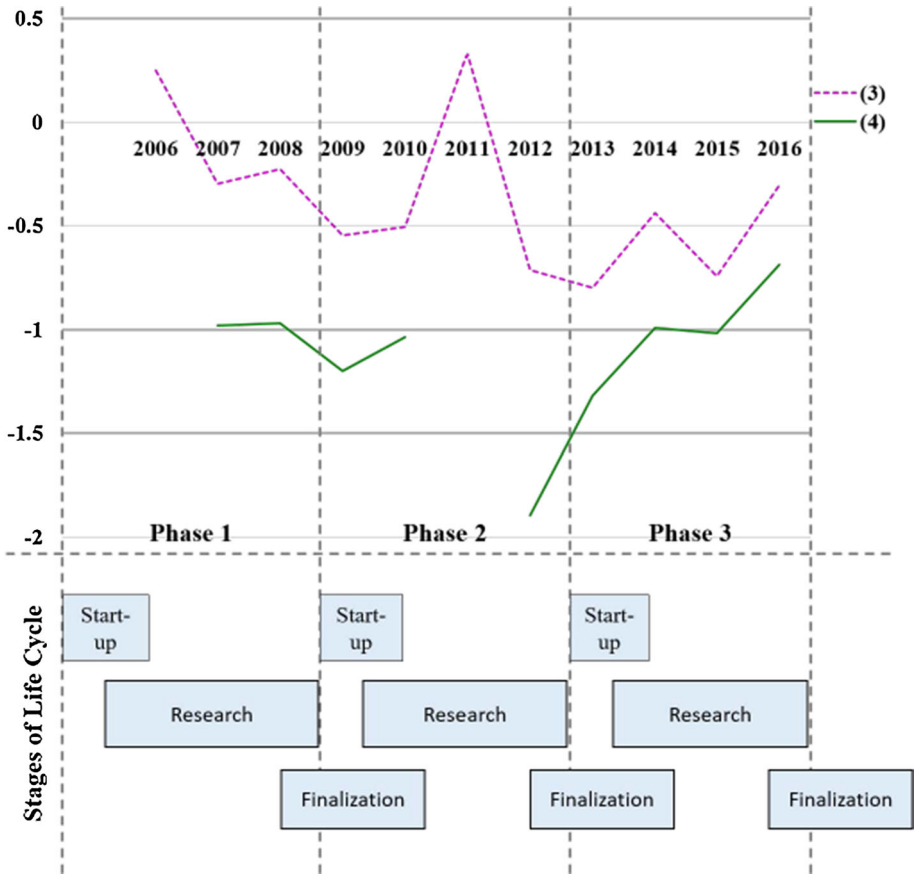


Fig. 9 Estimates of coefficients on the year dummy variables for the fixed effects Poisson regression (models (3) and (4)). The lower part of the figure shows the corresponding stage of the research project life cycle

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.
- Borrmann, 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 Publikationsstrategie. 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). Drittmittelinverbund. 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*, *102*(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 orbitofrontal cortex (IOFC), 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 orbitofrontal cortex (lOFC) and medial orbitofrontal 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 orbitofrontal 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

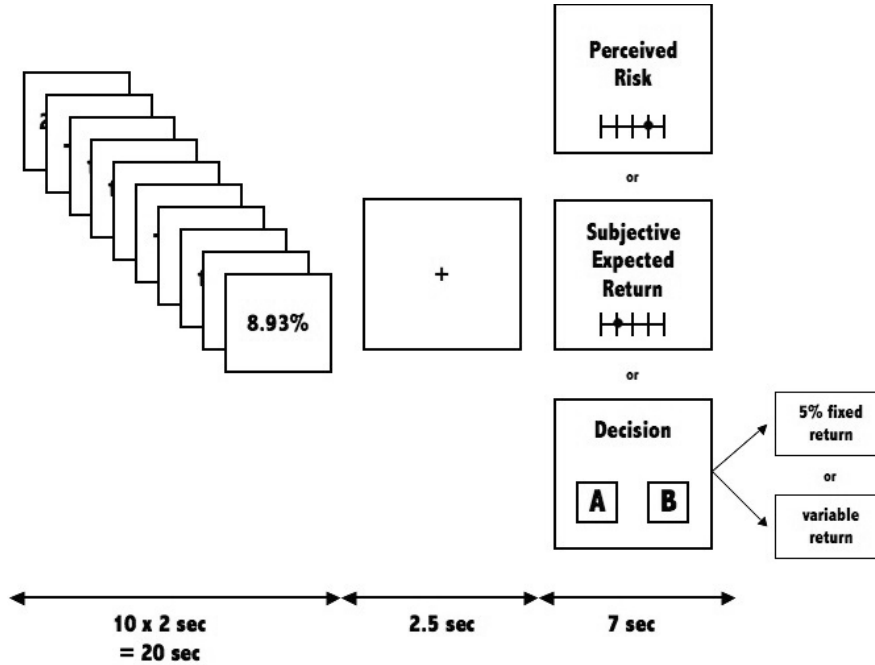


Figure 1: Graphic illustration of one trail of the RPID experiment, see [33].

(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 $t = 1, \dots, N$ for subject $j = 1, \dots, J$ 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 $f_t(\mathbf{x})$ with $\mathbf{x} = (x_1, x_2, x_3)$, one can represent it in a vector format

$$\mathbf{f}(\mathbf{x}) = \mathbf{C}\boldsymbol{\phi}(\mathbf{x}), \quad (3.1)$$

where \mathbf{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

$$\boldsymbol{\phi}(\mathbf{x}) = [\phi_1(x_1, x_2, x_3), \phi_2(x_1, x_2, x_3), \dots, \phi_{K^3}(x_1, x_2, x_3)]^\top$$

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

$$G(\mathbf{x}, \mathbf{s}) = \text{Cov}\{f(\mathbf{x}), f(\mathbf{s})\}$$

and its sample estimator

$$\widehat{G}(\mathbf{x}, \mathbf{s}) = N^{-1} \sum_{t=1}^N f_t(\mathbf{x})f_t(\mathbf{s}). \quad (3.2)$$

The covariance operator V is defined as

$$Vf = \int \int \int_{\mathcal{P}_1 \mathcal{P}_2 \mathcal{P}_3} G(\cdot, \mathbf{x})f(\mathbf{x}) \, d\mathbf{x}.$$

Similarly to the orthogonal decomposition in the multivariate PCA, we have for the 3D image functional data

$$V\xi = \int \int \int_{\mathcal{P}_1 \mathcal{P}_2 \mathcal{P}_3} G(\cdot, \mathbf{x})\xi(\mathbf{x}) \, d\mathbf{x} = \lambda\xi(\mathbf{x}),$$

where $\xi(\mathbf{x})$ and λ denote the eigenfunction on the 3D domain and the eigenvalue respectively. The eigenvalues are real and non-negative $\lambda_1 > \lambda_2 > \dots \geq 0$. 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

$$\widehat{G}(\mathbf{s}, \mathbf{x}) = N^{-1} \boldsymbol{\phi}^\top(\mathbf{s}) \mathbf{C}^\top \mathbf{C} \boldsymbol{\phi}(\mathbf{x}),$$

and the orthogonal decomposition equation as

$$\iiint N^{-1} \boldsymbol{\phi}^\top(\mathbf{s}) \mathbf{C}^\top \mathbf{C} \boldsymbol{\phi}(\mathbf{x}) \boldsymbol{\phi}^\top(\mathbf{x}) \mathbf{b} \, d(\mathbf{x}) = \lambda \boldsymbol{\phi}^\top(\mathbf{s}) \mathbf{b},$$

where the eigenfunction $\xi = \boldsymbol{\phi}^\top \mathbf{b}$ with \mathbf{b} being a coefficient vector. Define

$$\mathbf{W} = \iiint \boldsymbol{\phi}(\mathbf{x}) \boldsymbol{\phi}^\top(\mathbf{x}) \, d\mathbf{x}.$$

By solving

$$N^{-1} \mathbf{W}^{\frac{1}{2}} \mathbf{C}^\top \mathbf{C} \mathbf{W}^{\frac{1}{2}} \mathbf{u} = \lambda \mathbf{u}, \quad (3.3)$$

where $\mathbf{u} = \mathbf{W}^{\frac{1}{2}} \mathbf{b}$ and the coefficient vector \mathbf{b} satisfies $\mathbf{b}_i^\top \mathbf{W} \mathbf{b}_i = 1$ and $\mathbf{b}_i^\top \mathbf{W} \mathbf{b}_j = 0$, 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 :

$$\bar{Y}_t(x_1, x_2, x_3) = J^{-1} \sum_{j=1}^J Y_t^{(j)}(x_1, x_2, x_3), \quad t = 1, \dots, N.$$

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 $\tilde{\xi}_\ell(x_1, x_2, x_3)$ with $\ell = 1, \dots, L$, 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:

$$Y_t^{(j)}(x_1, x_2, x_3) = \sum_{\ell=1}^L Z_{\ell,t}^{(j)} \tilde{\xi}_\ell(x_1, x_2, x_3) + \varepsilon_t^{(j)}(x_1, x_2, x_3), \quad (3.4)$$

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 :

$$\min_{Z_{\ell,t}^{(j)}} \sum_{x_1, x_2, x_3} \left\{ Y_t^{(j)}(x_1, x_2, x_3) - \sum_{\ell=1}^L Z_{\ell,t}^{(j)} \tilde{\xi}_\ell(x_1, x_2, x_3) \right\}^2.$$

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 $\tilde{\xi}_\ell(x_1, x_2, x_3)$ 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, IOFC, 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 $[27, 29] \times [89, 91] \times [38, 40]$, IOFC at $[54, 56] \times [97, 99] \times [30, 32]$,

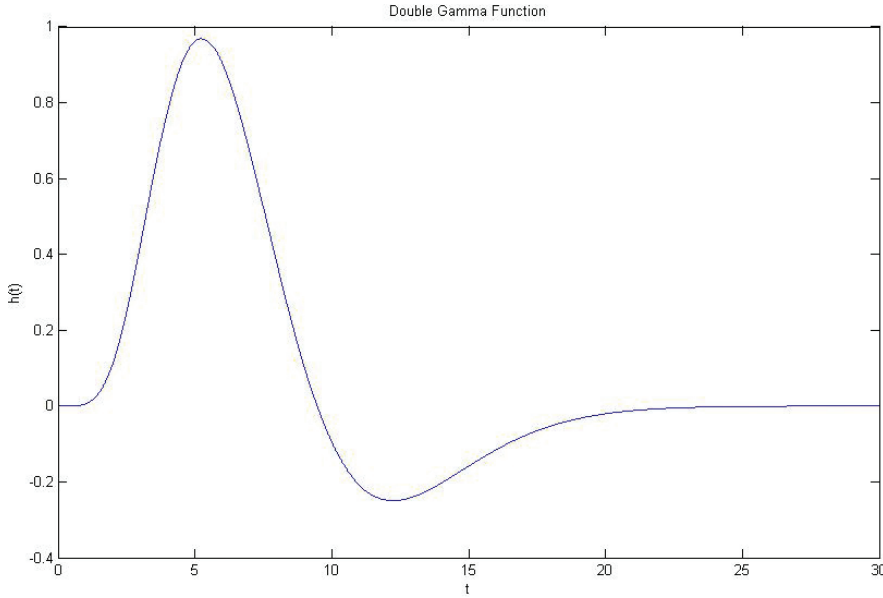


Figure 2: Visualization of the double gamma function.

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:

$$f_t^{(\text{NFL})}(x_1, x_2, x_3) = \sum_{\ell=1}^5 Z_{\ell t} \xi_{\ell}(x_1, x_2, x_3) + \varepsilon_t(x_1, x_2, x_3),$$

$$f_t^{(\text{HRF})}(x_1, x_2, x_3) = \sum_{\ell=1}^5 \{Z_{\ell t} + h(t)\} \xi_{\ell}(x_1, x_2, x_3) + \varepsilon_t(x_1, x_2, x_3),$$

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 $t = 1, \dots, 1000$. 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

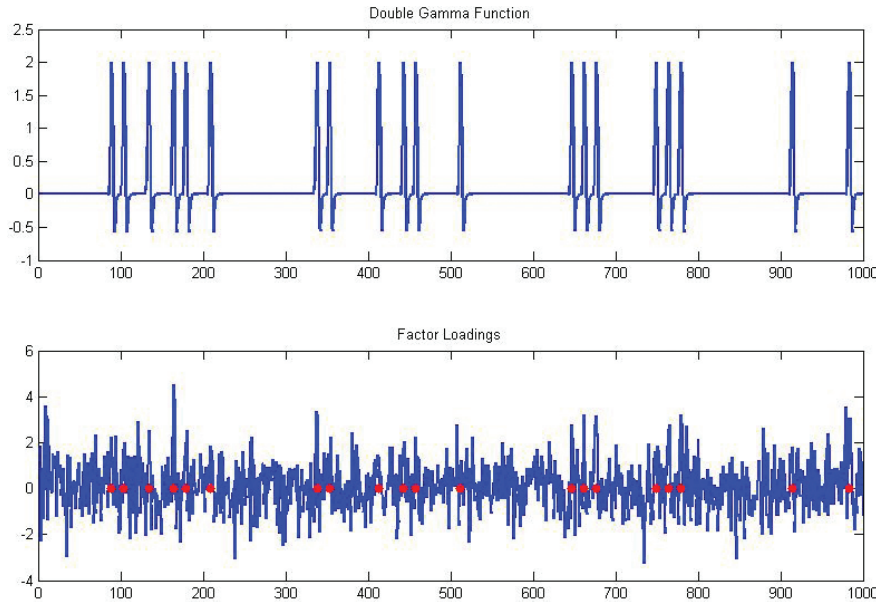


Figure 3: Simulated factor loadings. On top is the double gamma function. The bottom is the simulated factor loadings, which are the sum of the double gamma function and the normal random loadings. The red dots highlight time points when the stimulus are triggered.

$K = 16$ 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 IOFC 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 IOFC 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 IOFC (512 voxels), $[62, 66] \times [74, 78] \times [37, 39]$ for aINS (75 voxels), and $[64, 70] \times [73, 79] \times [51, 57]$ for

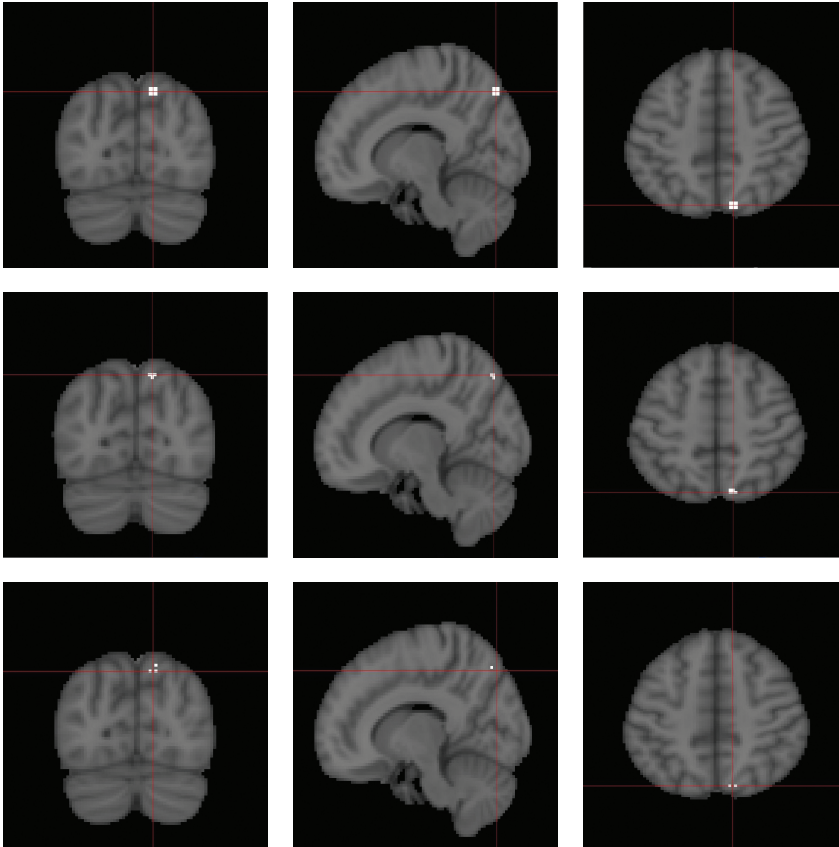


Figure 4: Functional factors on IOFC. From top to bottom are the generated (true) region, the estimated region with 3DIF and the estimated region with the FSVD method.

	Factor						Total
	1	2	3	4	5	6	
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.

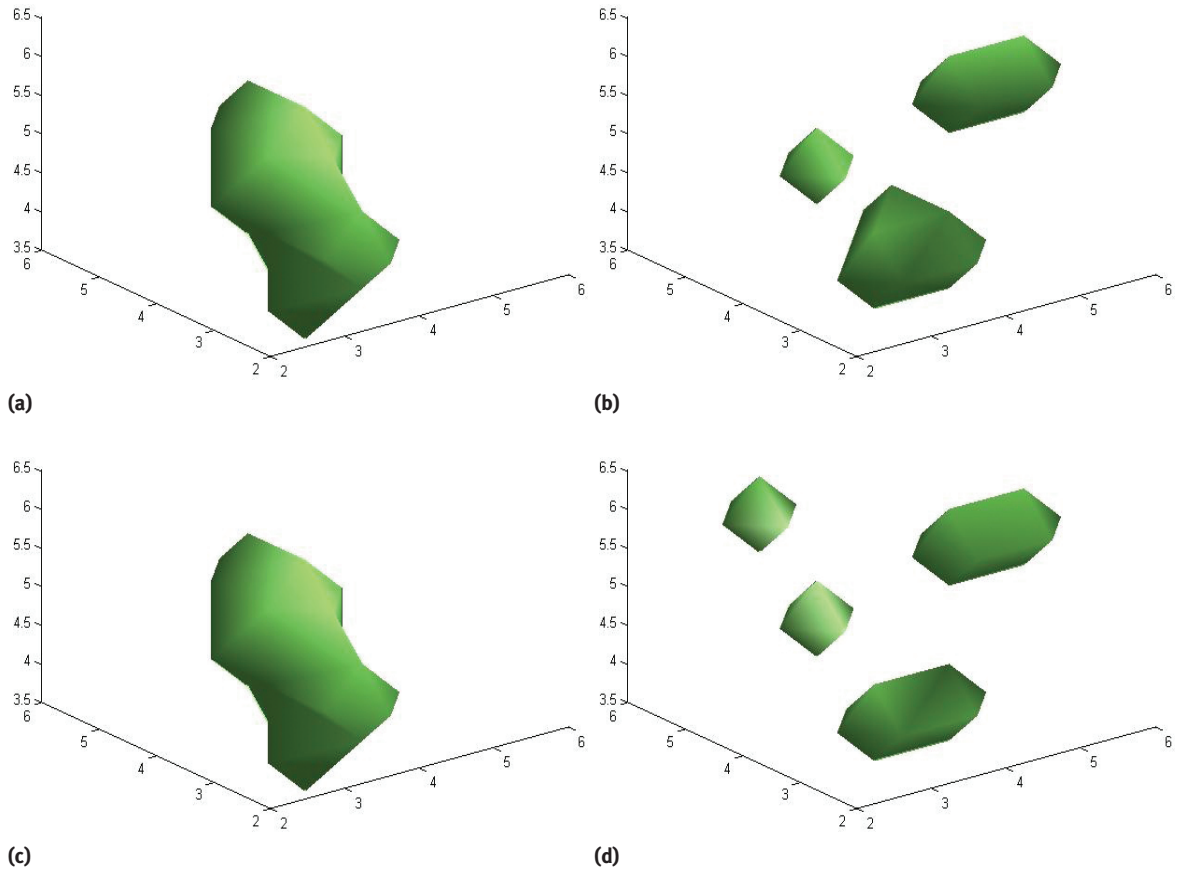


Figure 5: Contour plot of the estimated active region IOFC in NFL (top) and HRF (bottom) cases. On the left is the estimated region with 3DIF and on the right is the estimated region with FSVD.

	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 IOFC as illustration.

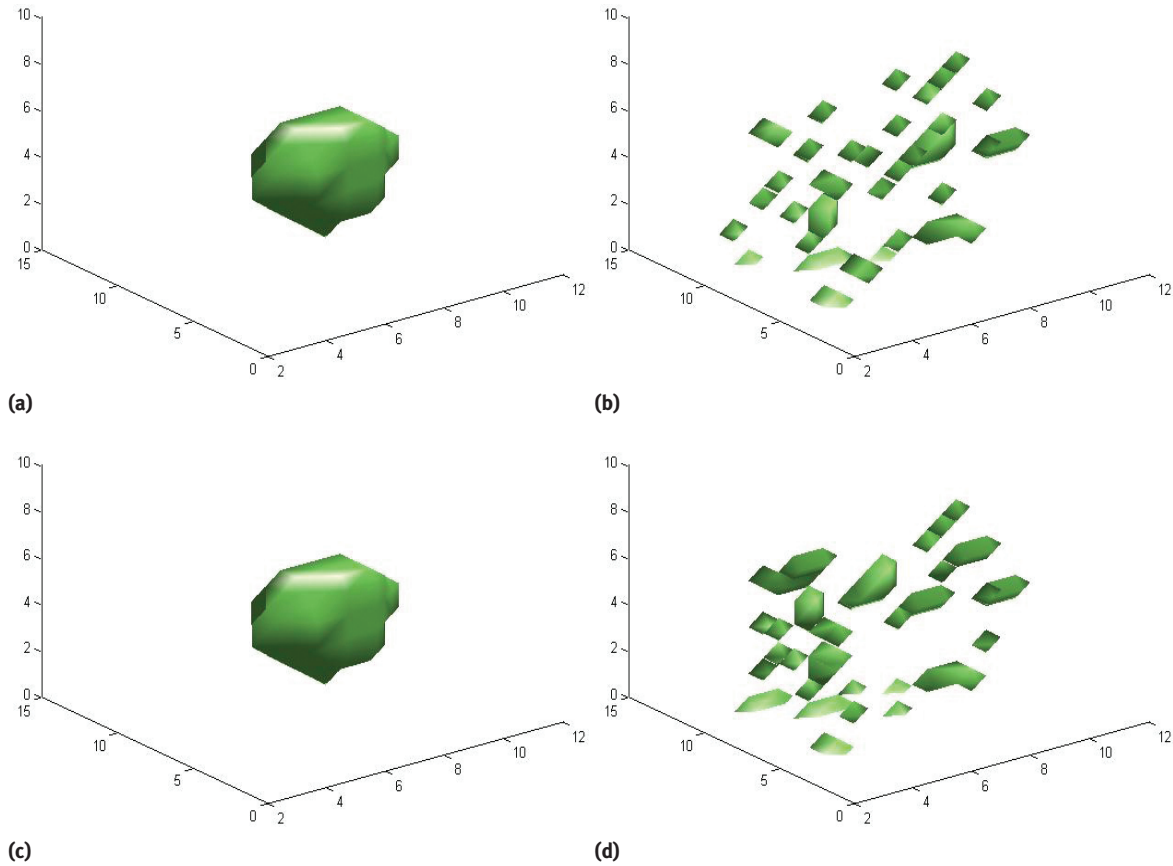


Figure 6: Robust: contour plot of the active region on IOFC. The left column is the estimated region in 3DIF and the right column is the estimated region with FSVD method. The top row is the result for NFL with normal factor loadings and on the bottom is the result for HRF with HRF incorporated in factor 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 $\mathbf{W} = \iiint \boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}^\top(\mathbf{x}) d\mathbf{x}$, 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 \mathbf{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 \mathbf{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 $Y = [Y_1, Y_2, \dots, Y_N]$ 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 eigen-decomposition 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 $\widehat{\xi}_\ell$

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.

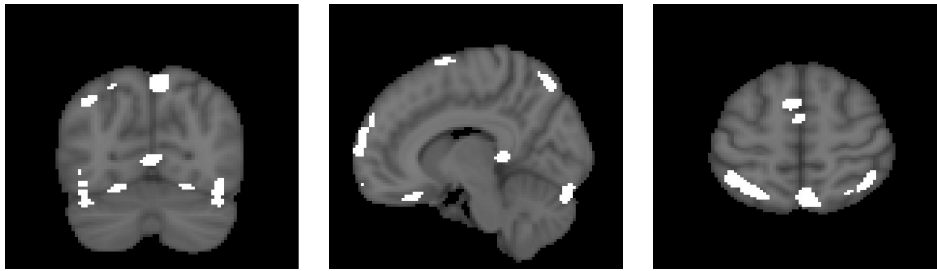
brain activity during the RPID experiment. Alternatively, the dominant component explains 96.50 % variation in FSVD, 70 % in FPCA and 98.67 % in SVD. Though numerically important, the first spatial factor has less psychological meaning and is irrelevant to any important risk related regions documented 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 IOFC ($\tilde{\xi}_5$) explains 2.73 % (the difference between 92.78 % for $L = 6$ and 90.05 % for $L = 4$), while FSVD ($\tilde{\xi}_5$) and SVD ($\tilde{\xi}_5$) both contribute 0.09 % and FPCA ($\tilde{\xi}_3$) 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 orbitofrontal cortex (IOFC) 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 orbitofrontal 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.

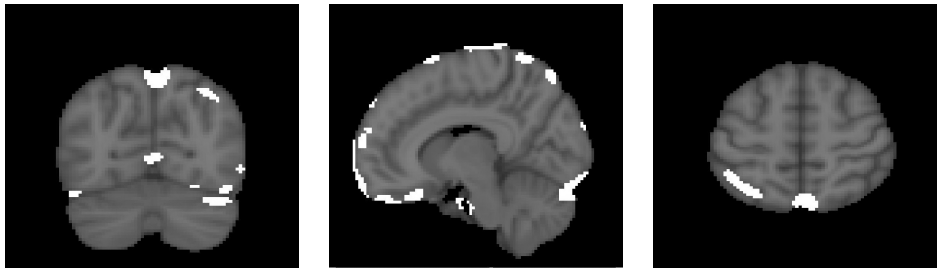
Figure 8 displays details of the detected regions by 3DIF. The relevant spatial factors are $\tilde{\xi}_\ell(x_1, x_2, x_3)$ with $\ell = 3, 4, 5, 12, 18, 19$. In particular, $\tilde{\xi}_3$ and $\tilde{\xi}_{12}$ are located in PC and attributed to risk related processes and selective attention (see [6, 41]); $\tilde{\xi}_4$ is related to the VLPFC region that stands for value processing. The regions mOFC and IOFC picked up by $\tilde{\xi}_5$ that are associated with evaluating and contrasting of different choice options [45]. The aINS region is captured by $\tilde{\xi}_{18}$ and related to risk and uncertainty [18], and the DLPFC area is highlighted by $\tilde{\xi}_{19}$. Figures 9–11 display the detected risk related brain regions by the alternative approaches. The identified regions of IOFC and VLPFC in Figures 9–11 are similar due to the nearby coordinates of the regions. The center coordinates of the identified IOFC is (61, 94, 31) and of the VLPFC is (30, 94, 36).

	PC	VLPFC	IOFC	aINS	DLPFC	mOFC	MPFC
3DIF	✓	✓	✓	✓	✓		
GLM	✓	✓		✓	✓		✓
PDSFM	✓					✓	
FSVD	✓	✓	✓				
FPCA	✓	✓	✓				
SVD	✓	✓	✓				

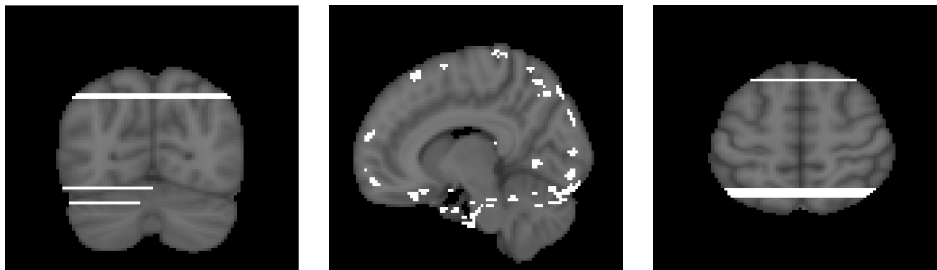
Table 5: Detected risk related brain regions for the same fMRI data of the RPID experiments in [33].



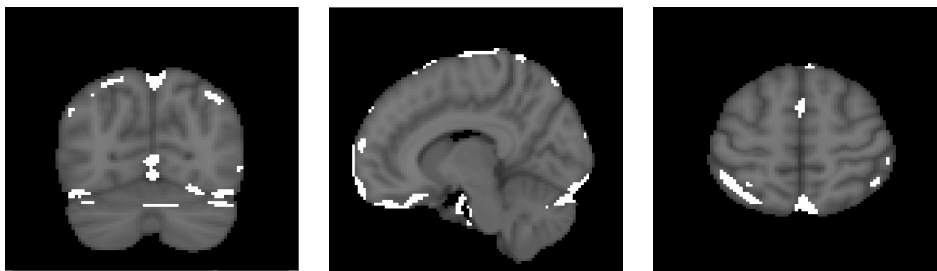
(a) 3DIF.



(b) FSVD.

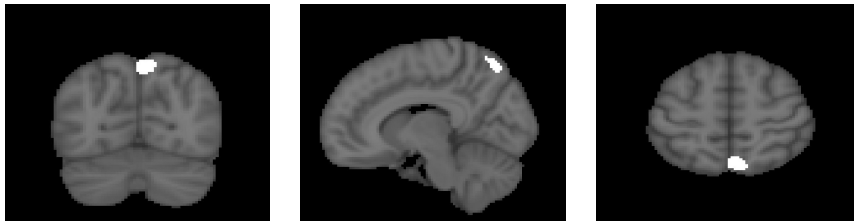


(c) FPCA.

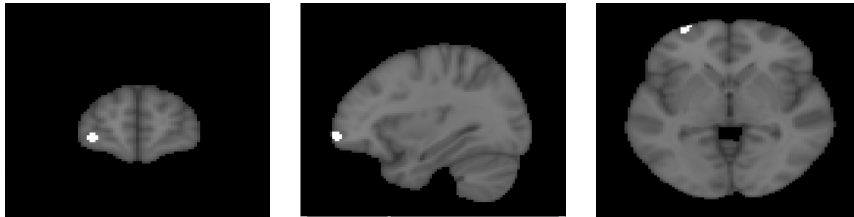


(d) SVD.

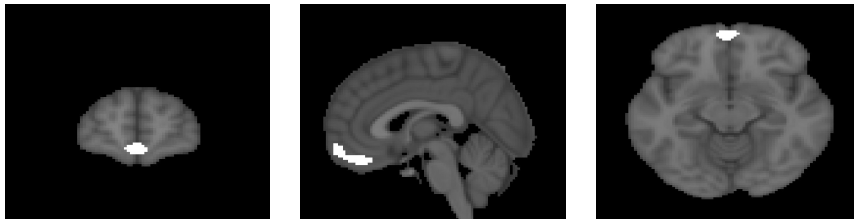
Figure 7: Detected risk-related brain regions by the first twenty eigenfunctions using (a) the 3DIF and alternative methods including (b) FSVD, (c) FPCA and (d) SVD.



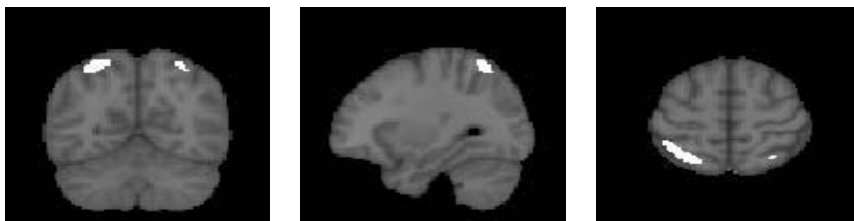
(a) Parietal Cortex.



(b) VLPFC.



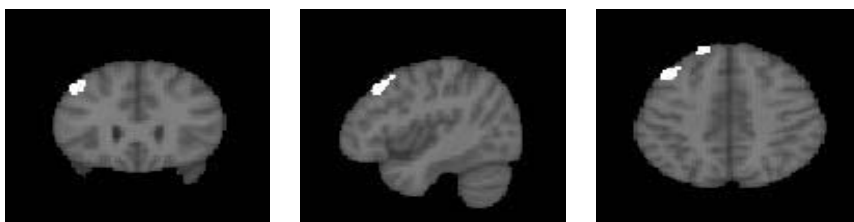
(c) IOFC.



(d) Parietal Cortex.

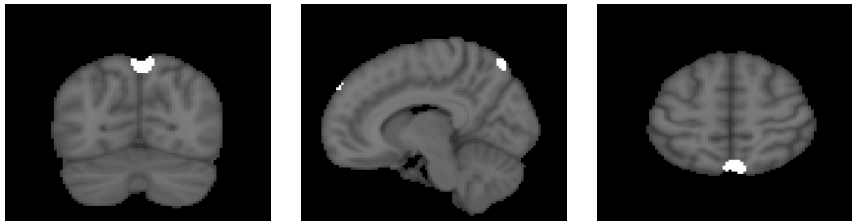


(e) aINS.

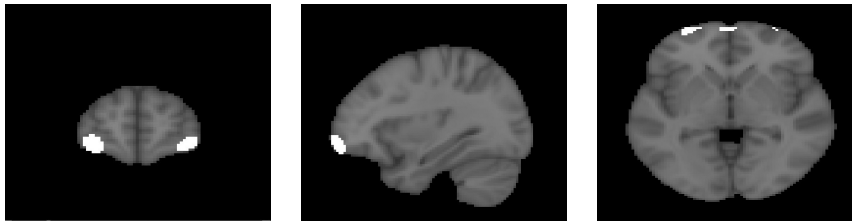


(f) DLPFC

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 IOFC; (d) $\widehat{\xi}_3$ in Parietal Cortex; (e) $\widehat{\xi}_{18}$ in aINS; (f) $\widehat{\xi}_{19}$ in DLPFC.

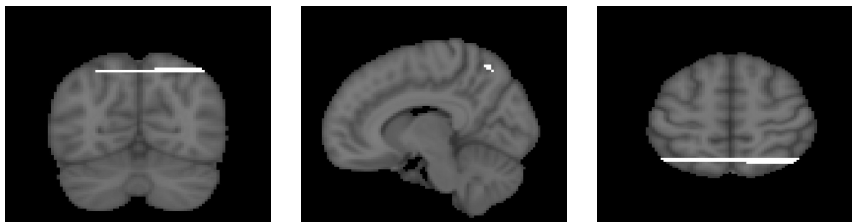


(a) Parietal Cortex.

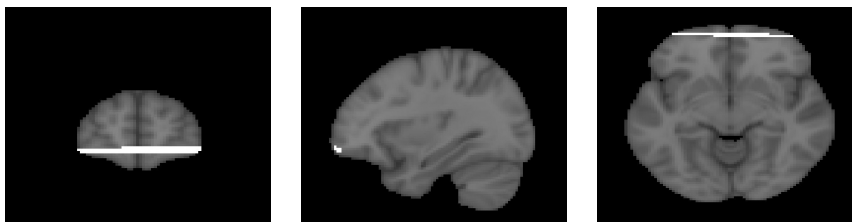


(b) VLPFC & IOFC.

Figure 9: FSVD: Selected identified risk related regions. (a) Estimated function $\hat{\xi}_{10}$ in Parietal Cortex; (b) $\hat{\xi}_5$ in VLPFC and IOFC.



(a) Parietal Cortex.



(b) VLPFC & IOFC.

Figure 10: FPCA: Selected identified risk related regions. (a) Estimated function $\hat{\xi}_2$ in Parietal Cortex; (b) $\hat{\xi}_3$ in VLPFC and IOFC.

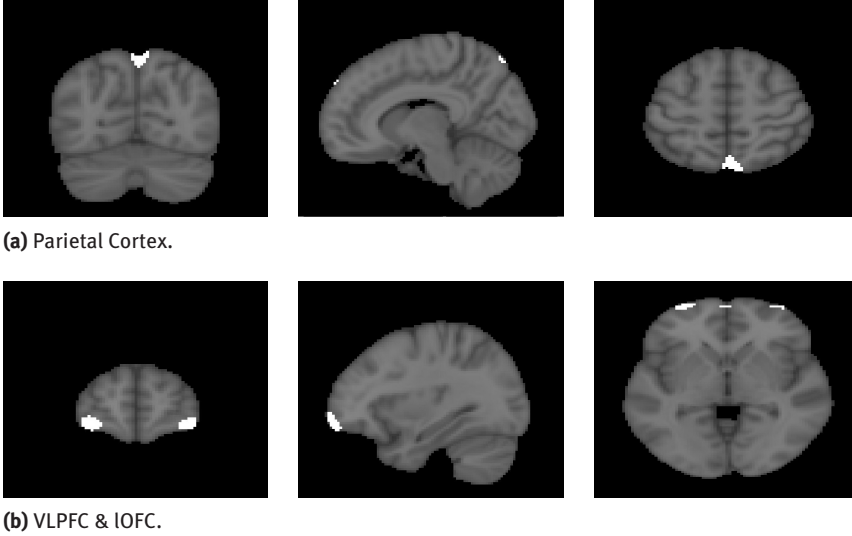


Figure 11: SVD: Selected identified risk related regions by SVD. (a) Estimated function $\hat{\xi}_{10}$ in Parietal Cortex; (b) $\hat{\xi}_5$ in VLPFC and IOFC.

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 $j = 1, \dots, 17$, $\ell = 1, \dots, 20$, and $t = 1, \dots, 1360$. 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

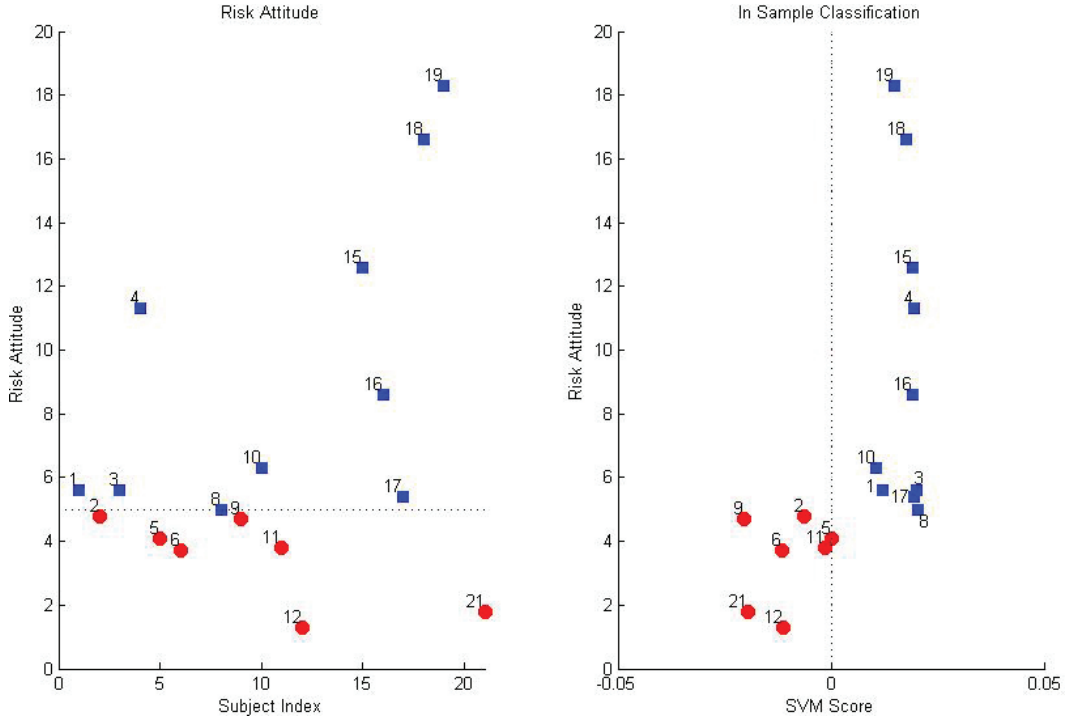


Figure 12: Risk attitudes and SVM scores of seventeen subjects. Subjects with risk attitude ≤ 5 are marked as red circles, otherwise as blue squares.

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 $Z_{\ell,t}^{(j)}$, $\ell = 3, 4, 5, 12, 18, 19$, 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_{ℓ,t_s}^j to capture the peak of the HRF. We compute the average to stand for the average reaction to stimulus s

$$\bar{\Delta Z}_{\ell,t_s}^{(j)} = \frac{1}{3} \sum_{\tau=1}^3 \hat{Z}_{\ell,t_s+\tau}^{(j)} - \hat{Z}_{\ell,t_s}^{(j)}$$

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

k	Overall				Strong				Weak			
	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 $k = 1, \dots, 6$. 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, IOFC, 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. Behrman, 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, Mahwah (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. Zippunikov, 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.



Forecasting limit order book liquidity supply–demand curves with functional autoregressive dynamics

Ying Chen, Wee Song Chua & Wolfgang Karl Härdle

To cite this article: Ying Chen, Wee Song Chua & Wolfgang Karl Härdle (2019) Forecasting limit order book liquidity supply–demand curves with functional autoregressive dynamics, *Quantitative Finance*, 19:9, 1473-1489, DOI: [10.1080/14697688.2019.1622290](https://doi.org/10.1080/14697688.2019.1622290)

To link to this article: <https://doi.org/10.1080/14697688.2019.1622290>



Published online: 09 Jul 2019.



Submit your article to this journal [↗](#)



Article views: 212



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 1 View citing articles [↗](#)

Forecasting limit order book liquidity supply–demand curves with functional autoregressive dynamics

YING CHEN^{†‡§}, WEE SONG CHUA^{*‡} and WOLFGANG KARL HÄRDLE^{¶||}

[†]Department of Mathematics, Faculty of Science, National University of Singapore, Block S17, Level 8, 2 Science Drive 2, Singapore 117543, Singapore

[‡]Department of Statistics & Applied Probability, Faculty of Science, National University of Singapore, Block S16, Level 7, 6 Science Drive 2, Singapore 117546, Singapore

[§]Risk Management Institute, National University of Singapore, 21 Heng Mui Keng Terrace, I³ Building #04-03, Singapore 119613, Singapore

[¶]Ladislav von Bortkiewicz Chair of Statistics, C.A.S.E. – Center for Applied Statistics & Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, Berlin 10099, Germany

^{||}Sim Kee Boon Institute for Financial Economics, Singapore Management University, 81 Victoria Street, Singapore 188065, Singapore

(Received 28 March 2018; accepted 6 May 2019; published online 9 July 2019)

We develop a dynamic model to simultaneously characterize the liquidity demand and supply in a limit order book. The joint dynamics are modeled in a unified Vector Functional AutoRegressive (VFAR) framework. We derive a closed-form maximum likelihood estimator under sieves and establish asymptotic consistency of the proposed method under mild conditions. We find the VFAR model presents strong interpretability and accurate out-of-sample forecasts. In application to limit order book records of 12 stocks in the NASDAQ, traded from 2 January 2015 to 6 March 2015, the VFAR model yields R^2 values as high as 98.5% for in-sample estimation and 98.2% in out-of-sample forecast experiments. It produces accurate 5-, 25- and 50-min forecasts, with RMSE as low as 0.09–0.58 and MAPE as low as 0.3–4.5%. The predictive power stably reduces trading cost in the order splitting strategies and achieves excess gains of 31 basis points on average.

Keywords: Liquidity demand and supply curves; Order splitting strategy; Vector functional autoregression; Liquidity forecasting; Time series

JEL Classification: C13, C32, C53

1. Introduction

Liquidation of large orders has attracted much attention from researchers and practitioners. Markets address the large order liquidation problem in one of three ways: call auctions, dealer markets and limit order books, see Foucault *et al.* (2005). Among them, Limit Order Book (LOB) has emerged as the main source for liquidity and exhibits a growing importance worldwide. LOB records investors' orders on both the bid and ask sides with price and volume constraints. With a limit order, investors can improve the execution price, either buying or selling, according to their choice, but the execution

is not immediate as a market order, nor guaranteed. Harris (1990) defines three components of liquidity as being: (1) tightness such as bid–ask spread at the best price level; (2) depth measured by quantities, e.g. eXchange Liquidity Measure (XLM); and (3) resilience referring to the recovery for deviations of spreads from their competitive level. As such, LOB contains the comprehensive information on market liquidity, not only a single-valued liquidity measurement at the best bid–ask price level, but also the queuing liquidities at deeper levels in the book. In this paper, we develop a dynamic model to simultaneously characterize the liquidity demand and supply in the LOB. The objectives are to understand the joint dynamics of liquidity at multiple levels and on both the

*Corresponding author. Email: a0054070@u.nus.edu

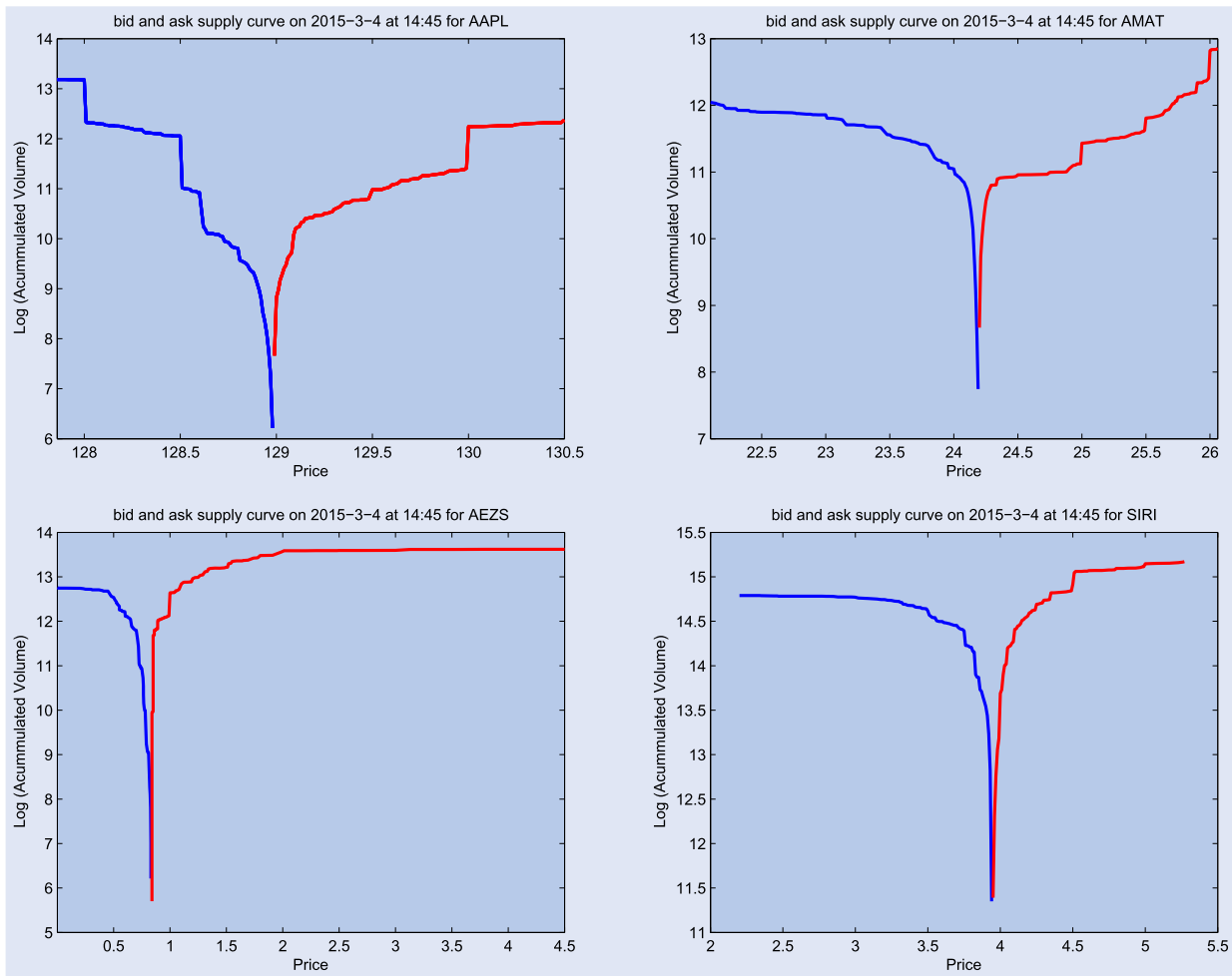


Figure 1. AAPL, AMAT, AEZS, and SIRI liquidity demand and supply curves at an arbitrarily selected time point. In our study, AAPL has the largest market value; AEZS has the smallest value and the smallest bid–ask spread on average; SIRI is the most active stock; and AMAT is the relatively less active stock among the rest.

bid and ask sides of the LOB, and explore the applicability of a dynamic model in the order splitting strategy.

Liquidity in the LOB can be well represented by a pair of demand and supply curves that are defined as the (log) accumulated volumes on the quoted prices. The demand curve corresponds to the bid side and the supply is associated with the ask side. As an illustration, figure 1 displays the liquidity curves of four stocks: Apple Inc. (AAPL); Applied Materials, Inc. (AMAT); Aeterna Zentaris Inc. (AEZS); and Sirius XM Holdings Inc. (SIRI), based on the snapshots of book on 4 March 2015 at 14:45:00. Each pair of demand and supply curves forms a V-shape that is monotonically decreasing on the bid side and monotonically increasing on the ask side. The gap at the center represents the bid–ask spread, i.e. the *market tightness* at the best price level. Moreover, the *market depth* is reflected by the gradients of the liquidity curves. More precisely, the steeper the curves are, the less price impact there is for large orders, and thus the more liquidity is ready to be supplied or consumed in the market. Liquidity is concentrated on relatively few quoted prices near the best bid and ask prices, while the tails are relatively flat. This flattening out of the tail, or the gentle gradient in the tails, implies low liquidity. A buy or sell of large volumes at the extreme prices will trigger a drastic change in the price and thus increase trading cost. The

dynamic dependence of the series on the liquidity demand and supply curves naturally inherits the *market resilience*, the third component of liquidity.

Liquidity is serially dependent, i.e. the current value of liquidity depends on its own past values. Though with limited information, the popular single-valued liquidity measures are found to be serially dependent; e.g. bid–ask spread (see Benston and Hagerman 1974, Stoll 1978, Fleming and Remolona 1999) and XLM (see Cooper *et al.* 1985, Gomber *et al.* 2015). These findings motivate as a first proxy the adoption of autoregressive models for liquidity in the LOB. Groß-Klußmann and Hautsch (2013) propose a long memory autoregressive conditional Poisson model for the quoted bid–ask spreads. Huberman and Halka (2001) evidence the serial dependence of bid–ask spread and market depth in an autoregressive framework. Härdle *et al.* (2015) propose a local adaptive multiplicative error model to forecast the high-frequency series of 1-min cumulative trading volumes of several NASDAQ blue chip stocks. Chordia *et al.* (2005) document in a vector autoregressive model the cross-sectional dependence among the liquidity measures of bid–ask spread, market depth and order flow, and other statistics of volatility and returns in the stock and bond markets, where the liquidity measures depend on both their own past values and the

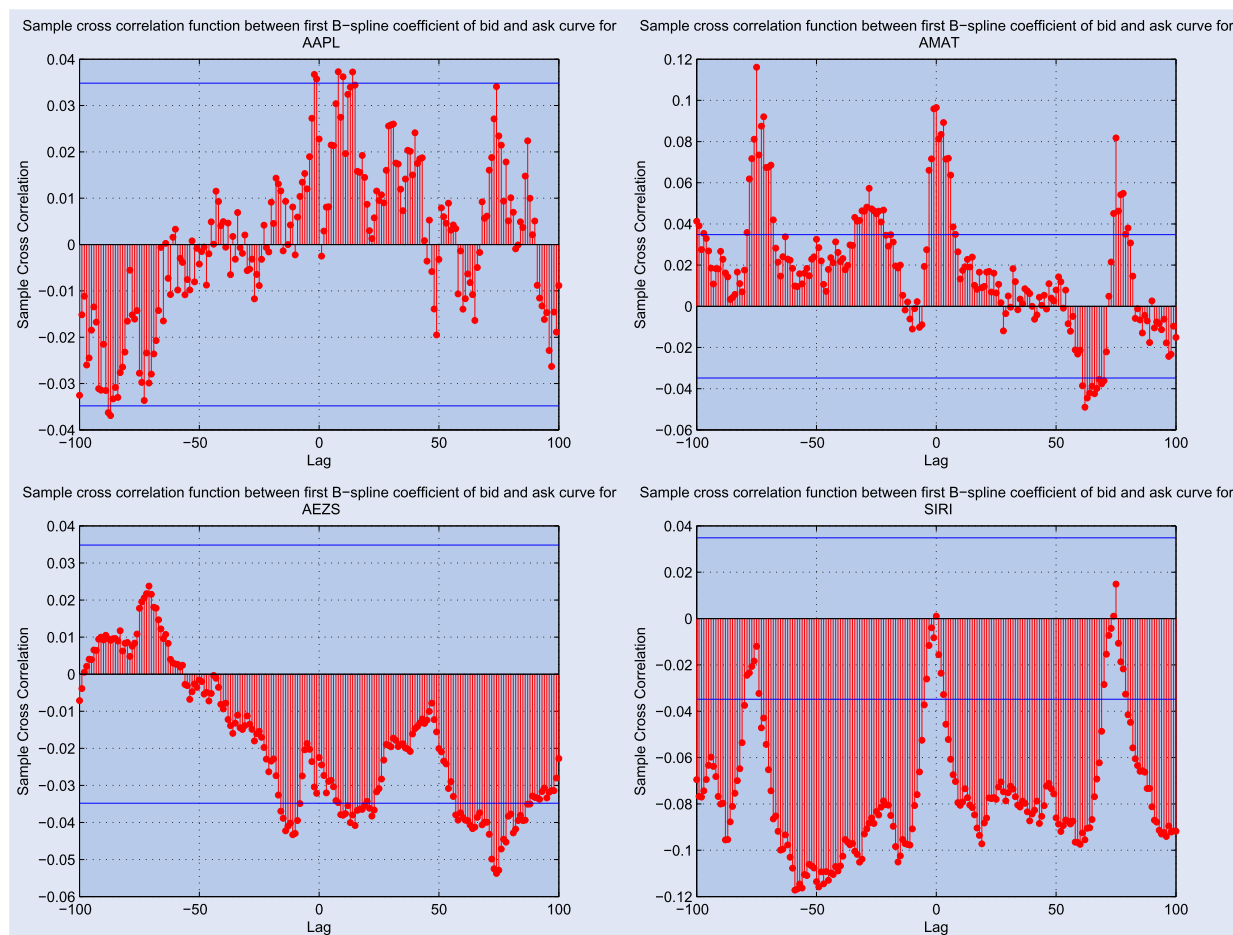


Figure 2. Sample cross-correlation function between first B -spline coefficient of the bid and ask curve for AAPL, AMAT, AEZS, and SIRI.

historical values of the other measures. Härdle *et al.* (2012) empirically analyze the seasonally-adjusted liquidity supply curves in the LOB using a dynamic semiparametric factor model, where the extracted factors of the curves are modeled in the Vector Error Correction (VEC) framework.

It is worth noting that there also exists serial cross-dependence in the bivariate series of liquidity demand and supply. The impact of public information on limit orders result in similar changes on both bid and ask sides, and can cause investors to switch from one side to the other. Thus it introduces lead–lag cross-dependence in both sides of liquidity. The joint serial cross-dependence suggests richer dynamics should be utilized in liquidity analysis. Building on this idea, we develop a Vector Functional AutoRegressive (VFAR) model to describe the joint dynamics of the bivariate series—liquidity demand and supply curves on the bid and ask sides of an electronic open LOB—simultaneously in a unified framework. While there is in general co-integration or a common trend in demand and supply when both variables are scalar time series, the model features the dependence of the series of curves, which is stable over time. Figure 2 displays the sample cross-correlations between the fundamental representatives of the demand and supply curves of the four stocks (the ‘first B -spline coefficients’ of the bid and ask curves). This first proxy shows that there are significant lead–lag cross-dependencies between the demand and supply curves, but the patterns deviate far from the persistence of a unit root.

In the VFAR model, we derive a closed-form maximum likelihood estimator under a sieve and establish the asymptotic consistency of the method. We investigate the finite sample performance of the proposed model along with the LOB records of 12 stocks traded in NASDAQ from 2 January 2015 to 6 March 2015, where the stocks are carefully selected to represent various types with different market capitalization and liquidity. We find the VFAR presents a strong predictability in liquidity, producing R^2 values as high as 98.5 % for in-sample estimation and 98.2 % in out-of-sample forecast experiments. Moreover, it yields accurate 5-, 25- and 50-min forecasts, with RMSE as low as 0.09–0.58 and MAPE as low as 0.3–4.5%. Finally, the predictive power stably reduces trading cost in the order splitting strategies and achieves excess gains of 31 basis points on average.

We would like to highlight the difference between our study and the existing ones in the literature. Above all, we develop a dynamic model to estimate and forecast liquidity demand and supply curves simultaneously and demonstrate the application in order splitting execution. Härdle *et al.* (2012) implement a dynamic semiparametric factor model for liquidity curves, but the extracted factors are handled separately on each side, although it is possible to capture the joint evolution of bid and ask sides. Secondly, we develop the FAR modeling for multiple functional time series, where the continuous curves are modeled in a convolutional VFAR that is stationary. Whereas in the functional time series literature,

Bosq (2000) has proposed the FAR model for univariate functional time series and developed Yule–Walker estimation (see also Besse *et al.* 2000, Guillas 2001, Antoniadis and Sapatinas 2003, Chaudhuri *et al.* 2016). Mourid and Bensmain (2006) propose a maximum likelihood estimation with Fourier expansions, as far as we know this is the first work to model multiple functional time series. We investigate its theoretical properties, and within a maximum likelihood estimator approach based on *B*-spline expansions, we provide more flexibility in fitting beyond the Fourier expansion in Chen and Li (2017). Although the implementation focuses on the bivariate liquidity demand and supply curves in our study, the developed model is general and can be used for analyzing multiple functional time series in other research areas.

This paper is structured as follows. In Section 2, we describe the LOB data. Section 3 introduces the VFAR modeling in detail including the estimation approach and the theoretical properties. Section 4 presents the modeling setup and in-sample estimation results, reports the out-of-sample forecast results and demonstrates the application to an order execution strategy. Section 5 provides concluding remarks. All of the theoretical proofs are contained in the Appendix.

2. Data

We consider the LOB records of 12 stocks from 2 January 2015 to 6 March 2015 (44 trading days). The LOB records contain the quoted prices and volumes up to 100 price levels on each side of ask and bid. All the quotes are timestamped with decimal precision up to nanoseconds ($= 10^{-9}$ s). In total, the (buy or sell) order book contains 400 values from the best ask price, best ask volume, best bid price, and best bid volume until the 100-th best ask (bid) price and corresponding volume. The data was obtained from LOBSTER through the Research Data Center of the Collaborative Research Center 649 (<https://sfb649.wiwi.hu-berlin.de/fedc/>). Note that the records of the 12 stocks in LOBSTER only contain the information in the National Association of Securities Dealers Automated Quotations (NASDAQ) stock market. NASDAQ is a continuous auction trading platform where the normal

continuous trading hours are between 9:30 am to 4:00 pm from Monday to Friday. During the normal trading, if an order cannot be executed immediately or completely, the remaining volumes are queued in the bid and ask sides according to a strict price-time priority order.

The stocks correspond to high variations in terms of market capitalization, liquidity tightness and depth. They are Apple Inc. (AAPL), Microsoft Corporation (MSFT), Intel Corporation (INTC), Cisco Systems, Inc. (CSCO), Sirius XM Holdings Inc. (SIRI), Applied Materials, Inc. (AMAT), Comcast Corporation (CMCSA), AEterna Zentaris Inc. (AEZS), eBay Inc. (EBAY), Micron Technology, Inc. (MU), Whole Foods Market, Inc. (WFM), and Starbucks Corporation (SBUX). The largest stock is AAPL with market value of USD737.41 billions, and the smallest is AEZS with market value of USD35.38 millions. When considering the 5-min queuing volume in the LOB, the most active stock is SIRI, with value of 3.73 millions on the bid side and 7.61 millions on the ask side. The least active stocks are CMCSA with value of 0.02 millions on the bid side and SBUX with value of 0.03 millions on the ask side. Moreover, the average value of the bid–ask spread varies from 0.0062 (AEZS) to 0.0213 (SBUX), see table 1.

In data pre-processing, we remove the first 15 min after opening and the last 5 min before closing to eliminate the market opening and closing effect. The accumulated bid and ask volumes are log-transformed when constructing liquidity curves to reduce the impact of extraordinarily large volumes. The liquidity curves are smoothed over the 100 price levels of LOB on each side using *B*-spline basis functions. Moreover, to remove the impact of microstructure noise, the sampling frequency is set to be 5 min for a good strike between bias and variance, see AitSahalia *et al.* (2005), Zhang *et al.* (2005), and Hårdle *et al.* (2018). As such, there are 75 pairs of bid and ask liquidity curves on each day for each stock. Over the whole sample period of 44 trading days, it amounts to 3300 pairs of bid and ask supply curves for each stock.

The liquidity curves exhibit significant serial dependence over time. As an illustration, figure 3 shows the sample cross-correlations between the log-accumulated volumes at best bid and ask prices for four representative stocks including AAPL with the largest market value, AEZS with the smallest value

Table 1. Summary statistics on liquidity measures for the 12 stocks traded in NASDAQ.

Ticker symbol	Mean spread (USD)	Bid vol		Ask vol	
		Min	Max	Min	Max
AAPL	0.0125	52,267	710,020	61,305	1,298,696
MSFT	0.0101	90,344	928,319	122,377	621,471
INTC	0.0102	158,900	557,251	146,959	1,142,641
CSCO	0.0101	134,790	1,316,058	266,455	4,458,672
SIRI	0.0101	1,266,528	3,725,304	3,002,680	7,605,467
AMAT	0.0102	78,944	334,794	180,749	787,983
CMCSA	0.0106	23,668	128,916	40,638	146,724
AEZS	0.0062	145,635	767,785	472,689	1,158,740
EBAY	0.0110	42,060	160,572	52,813	415,033
MU	0.0107	95,907	497,910	102,357	595,200
WFM	0.0153	34,538	114,386	41,019	159,488
SBUX	0.0213	27,467	151,022	34,914	166,932

Note: Sampling frequency is 5 min.

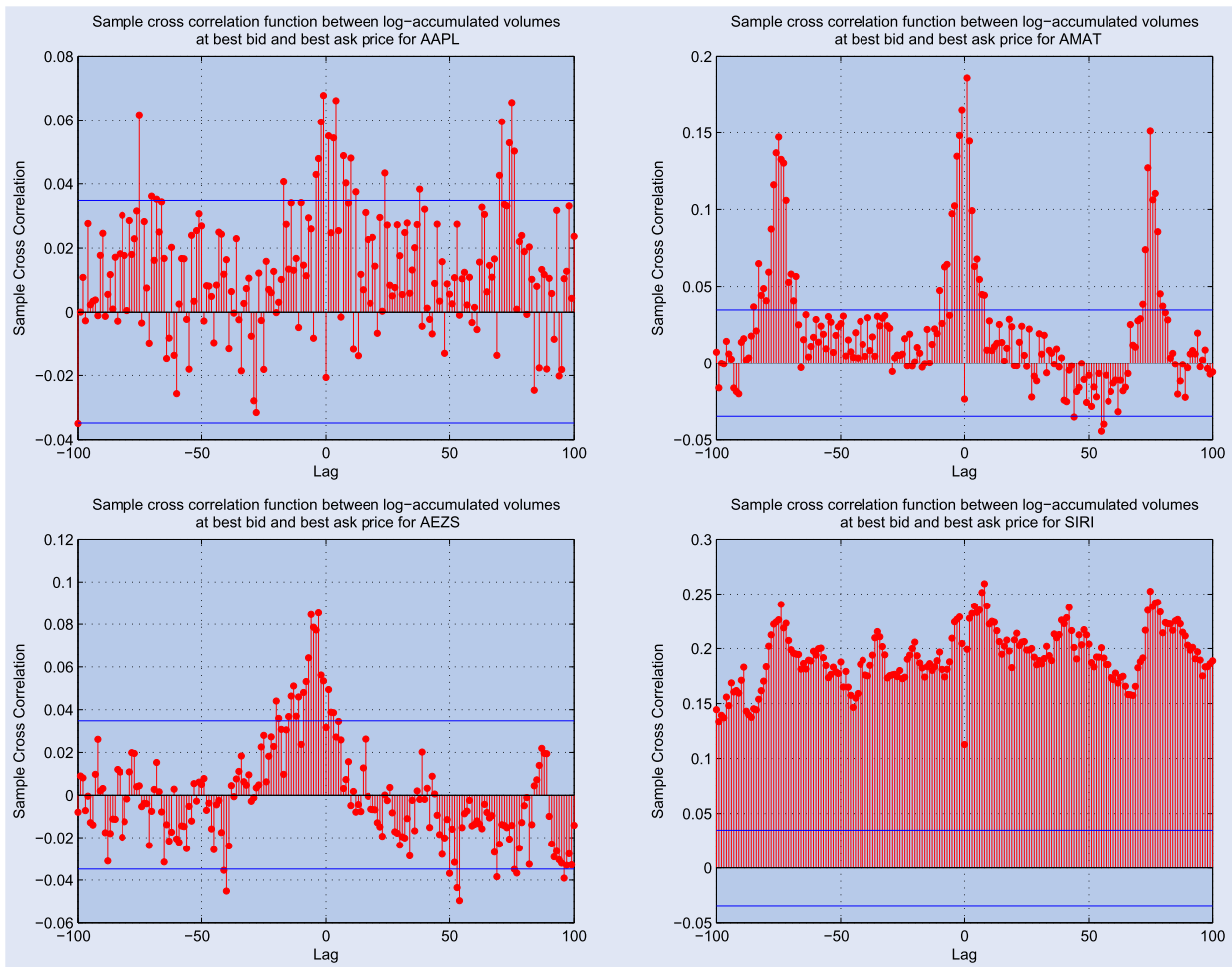


Figure 3. Sample cross-correlation function between log-accumulated volumes at best bid and ask price for AAPL, AMAT, AEZS, and SIRI.

and the smallest bid–ask spread on average, SIRI the most active stock, and AMAT the relatively mid active stock among the rest. While the simultaneous dependence between the bid and ask sides is insignificant or negatively correlated, there are significant positive serial correlations on the first lagged values of the opposite side and then decays for higher order. Similar features are observed in the other eight stocks, which are not displayed. The bid–ask cross-dependency motivates analyzing the liquidity demand and supply jointly.

In addition, the serial cross-dependence between the series of curves is reflected in figure 2, where the sample cross-correlations are computed based on the first B -spline coefficients of the bid and ask curves, which will be detailed later. It shows that there are significant lead–lag cross-dependence between the demand and supply curves, but the patterns deviate far from the persistence of unit root. We also perform the Johansen (1991) co-integration test to assess the null hypothesis of co-integration of bid and ask sides. The results support rejecting co-integration.

3. Vector functional autoregression

In this section, we present the Vector Functional AutoRegressive (VFAR) model that is to describe the joint serial cross-dependence of multiple series of continuous curves.

We show how to estimate the functional parameters of the bivariate liquidity demand and supply curves, with the help of B -spline expansions and sieve method. A closed-form estimator is presented and asymptotic consistency provided.

The liquidity curves are defined on the quoted price, denoted as τ , which is assumed to be exogenous. The quoted prices are observed on a dense tick grid, that is re-scaled separately for bid and ask sides into a continuous interval $[0, 1]$, with the minimum price as 0 and maximum price as 1. Denote by $X_t^{(a)}(\tau)$ and $X_t^{(b)}(\tau)$ the liquidity supply and demand curves on the ask (a) side and the bid (b) side at $t = 1, \dots, n$. The curves are realizations of a functional stochastic process in the space $C_{[0,1]}$ of real continuous functions on $[0, 1]$. In other words, each pair of the liquidity curves is one functional object. Over time, the curves form bivariate time series of n functional objects, each on the bid and ask sides. In our study, the curves are obtained by smoothing over the discrete log-accumulated volumes against the quoted prices at every time point.

To handle the joint dynamics of the two continuous liquidity curves, we propose a Vector Functional AutoRegressive (VFAR) model of order p :

$$\begin{bmatrix} X_t^{(a)} - \mu_a \\ X_t^{(b)} - \mu_b \end{bmatrix} = \sum_{k=1}^p \begin{bmatrix} \rho^{aa,k} & \rho^{ab,k} \\ \rho^{ba,k} & \rho^{bb,k} \end{bmatrix} \begin{bmatrix} X_{t-k}^{(a)} - \mu_a \\ X_{t-k}^{(b)} - \mu_b \end{bmatrix} + \begin{bmatrix} \varepsilon_t^{(a)} \\ \varepsilon_t^{(b)} \end{bmatrix}, \quad (1)$$

where the operators $\rho^{aa,k}$, $\rho^{ab,k}$, $\rho^{ba,k}$, and $\rho^{bb,k}$ measure the serial cross-dependence among the liquidity demand and supply curves on their k th lagged values. The operators are bounded linear operator from \mathcal{H} to \mathcal{H} , a real separable Hilbert space endowed with its Borel σ -algebra $B_{\mathcal{H}}$. The mean function is denoted as $(\mu_a(\tau), \mu_b(\tau))^{\top} \stackrel{\text{def}}{=} (\mathbb{E}[X_t^{(a)}(\tau)], \mathbb{E}[X_t^{(b)}(\tau)])^{\top}$. Under stationarity, both the serial cross-dependence and the mean are constant over time. The innovations $\{\varepsilon_t^{(a)}\}_{t=1}^n$ and $\{\varepsilon_t^{(b)}\}_{t=1}^n$ are strong \mathcal{H} -white noise, independently and identically distributed with zero mean and finite second moment, $0 < \mathbb{E}\|\varepsilon_1^{(a)}\|^2 = \dots = \mathbb{E}\|\varepsilon_n^{(a)}\|^2 < \infty$ and $0 < \mathbb{E}\|\varepsilon_1^{(b)}\|^2 = \dots = \mathbb{E}\|\varepsilon_n^{(b)}\|^2 < \infty$, where the norm $\|\cdot\|$ is induced from the inner product $\langle \cdot, \cdot \rangle$ of \mathcal{H} . The innovation processes $\varepsilon_t^{(a)}$ and $\varepsilon_t^{(b)}$ need not be cross-independent.

In the following, we derive the estimation for the VFAR model of order 1, which can be generalized for higher order. For the notational simplification, the superscript k is dropped. We consider the convolutional VFAR, where each operator ρ is represented by a convolution kernel Hilbert–Schmidt operator,

$$\begin{aligned} X_t^{(a)}(\tau) - \mu_a(\tau) &= \int_0^1 \kappa_{ab}(\tau - s) \{X_{t-1}^{(b)}(s) - \mu_b(s)\} ds \\ &\quad + \int_0^1 \kappa_{aa}(\tau - s) \{X_{t-1}^{(a)}(s) - \mu_a(s)\} ds + \varepsilon_t^{(a)}(\tau), \\ X_t^{(b)}(\tau) - \mu_b(\tau) &= \int_0^1 \kappa_{bb}(\tau - s) \{X_{t-1}^{(b)}(s) - \mu_b(s)\} ds \\ &\quad + \int_0^1 \kappa_{ba}(\tau - s) \{X_{t-1}^{(a)}(s) - \mu_a(s)\} ds + \varepsilon_t^{(b)}(\tau). \end{aligned} \quad (2)$$

The kernel function $\kappa_{xy} \in L^2([0, 1])$ and $\|\kappa_{xy}\|_2 < 1$ for $xy = aa, ab, ba$, and bb , where $\|\cdot\|_2$ denotes the L^2 norm in $C_{[0,1]}$. Note that a linear operator ρ on a Hilbert space \mathcal{H} with norm $\|\cdot\|$ and inner product $\langle \cdot, \cdot \rangle$ is Hilbert–Schmidt if $\rho(\cdot) = \sum_j \lambda_j \langle \cdot, e_j \rangle f_j$, where $\{e_j\}$ and $\{f_j\}$ are orthonormal bases of \mathcal{H} and $\{\lambda_j\}$ is a real sequence such that $\sum_j \lambda_j^2 < \infty$.

Expand the functional terms in (2) using the B -spline basis functions in $L^2([0, 1])$:

$$\begin{aligned} B_{j,m}(\tau) &= \frac{\tau - w_j}{w_{j+m-1} - w_j} B_{j,m-1}(\tau) + \frac{w_{j+m} - \tau}{w_{j+m} - w_{j+1}} B_{j+1,m-1}(\tau), \\ m &\geq 2, \end{aligned}$$

where m is the order, $w_1 \leq \dots \leq w_{J+m}$ denote the sequence of knots, and

$$B_{j,1}(\tau) = \begin{cases} 1 & \text{if } w_j \leq \tau < w_{j+1}, \\ 0 & \text{otherwise.} \end{cases}$$

Plug-in the B -spline expansions to the VFAR model (2), we obtain the relationship of the B -spline coefficients in the

framework of VFAR:

$$\begin{aligned} d_{t,h}^a &= p_h^a + d_h^a(\varepsilon_t^{(a)}) + \sum_{i=1}^{\infty} \\ &\quad \times \left\{ \sum_{j=1}^{\infty} \left(\frac{w_{j+m} - w_{j+1}}{w_{j+m} - w_j} - \frac{w_{j+m+1} - w_{j+2}}{w_{j+m+1} - w_{j+1}} \right) c_j^{aa} - c_h^{aa} \right\} \\ &\quad \times \frac{w_{i+m} - w_i}{m} d_{t-1,i}^a + \sum_{i=1}^{\infty} \\ &\quad \times \left\{ \sum_{j=1}^{\infty} \left(\frac{w_{j+m} - w_{j+1}}{w_{j+m} - w_j} - \frac{w_{j+m+1} - w_{j+2}}{w_{j+m+1} - w_{j+1}} \right) c_j^{ab} - c_h^{ab} \right\} \\ &\quad \times \frac{w_{i+m} - w_i}{m} d_{t-1,i}^b, \\ d_{t,h}^b &= p_h^b + d_h^b(\varepsilon_t^{(b)}) + \sum_{i=1}^{\infty} \\ &\quad \times \left\{ \sum_{j=1}^{\infty} \left(\frac{w_{j+m} - w_{j+1}}{w_{j+m} - w_j} - \frac{w_{j+m+1} - w_{j+2}}{w_{j+m+1} - w_{j+1}} \right) c_j^{bb} - c_h^{bb} \right\} \\ &\quad \times \frac{w_{i+m} - w_i}{m} d_{t-1,i}^b + \sum_{i=1}^{\infty} \\ &\quad \times \left\{ \sum_{j=1}^{\infty} \left(\frac{w_{j+m} - w_{j+1}}{w_{j+m} - w_j} - \frac{w_{j+m+1} - w_{j+2}}{w_{j+m+1} - w_{j+1}} \right) c_j^{ba} - c_h^{ba} \right\} \\ &\quad \times \frac{w_{i+m} - w_i}{m} d_{t-1,i}^a, \end{aligned} \quad (3)$$

where $d_{t,j}^a$ and $d_{t,j}^b$ are the B -spline coefficients for the observed functional data $X_t^{(a)}$ and $X_t^{(b)}$, respectively; p_h^a are the coefficients associated with the expansion of the mean on the ask side $\mu_a(\tau) - \int_0^1 \kappa_{ab}(\tau - s) \mu_b(s) ds - \int_0^1 \kappa_{aa}(\tau - s) \mu_a(s) ds$ and p_h^b are the coefficients on the bid side for $\mu_b(\tau) - \int_0^1 \kappa_{bb}(\tau - s) \mu_b(s) ds - \int_0^1 \kappa_{ba}(\tau - s) \mu_a(s) ds$; $d_j^a(\varepsilon_t^{(a)})$ and $d_j^b(\varepsilon_t^{(b)})$ are the B -spline coefficients for the unknown innovations $\varepsilon_t^{(a)}$ and $\varepsilon_t^{(b)}$, respectively; and c_j^{aa} , c_j^{ab} , c_j^{ba} , and c_j^{bb} are the B -spline coefficients for the unknown kernel functions κ_{aa} , κ_{ab} , κ_{ba} , and κ_{bb} , respectively. As such, the original problem of estimating the functional parameters can now be equivalently solved by the estimation of the B -spline coefficients in (3).

Given a finite sample of the functional objects, it is however impossible to estimate the infinite coefficients in (3) for $i, j = 1, \dots, \infty$. The estimation is conducted with the help of a sieve.

3.1. Maximum likelihood estimator under sieve

We introduce a sequence of subsets—named sieve for the parameter space Θ , which is denoted by $\{\Theta_{J_n}\}$ with $J_n \rightarrow +\infty$ as $n \rightarrow +\infty$, see e.g. Grenander (1981) on the theory of sieves. In other words, the dimension of the subset is allowed to increase with the sample size. We have $\Theta_{J_n} \subseteq \Theta_{J_{n+1}}$ and the union of subsets $\bigcup \Theta_{J_n}$ is dense in the parameter space.

The sieve is defined as follows:

$$\Theta_{J_n} = \left\{ \kappa_{xy} \in L^2 \mid \kappa_{xy}(\tau) = \sum_{l=1}^{J_n} c_l^{xy} B_{l,m}(\tau), \tau \in [0, 1], \right. \\ \left. \times \sum_{l=1}^{J_n} l^2 (c_l^{xy})^2 \leq \nu J_n \right\}, \quad (4)$$

where ν is some known positive constant such that the constraint on c_l^{xy} can be satisfied generally without sacrifice of the growth rate of J_n . We will show the estimation in the finite subsets of the parameter space.

Under the sieve with J_n , Equation (3) can be represented in a form as follows:

$$y_t = \nu + C y_{t-1} + u_t, \quad (5)$$

where $y_t = (d_{t,1}^a, \dots, d_{t,J_n}^a, d_{t,1}^b, \dots, d_{t,J_n}^b)^\top$, $\nu = (p_1^a, \dots, p_{J_n}^a, p_1^b, \dots, p_{J_n}^b)^\top$, $u_t = (d_1^a(\varepsilon_t^{(a)}), \dots, d_{J_n}^a(\varepsilon_t^{(a)}), d_1^b(\varepsilon_t^{(b)}), \dots, d_{J_n}^b(\varepsilon_t^{(b)}))^\top$, and $C = [{}_{R^{ba}R^{bb}}^{R^{aa}R^{ab}}]$ with R^{xy} being a $J_n \times J_n$ matrix with elements $r_{h,i}^{xy} = \{\sum_{j=1}^{J_n} ((w_{j+m} - w_{j+1})/(w_{j+m} - w_j) - (w_{j+m+1} - w_{j+2})/(w_{j+m+1} - w_{j+1})) c_j^{xy} - c_h^{xy}\} ((w_{i+m} - w_i)/m)$, for $xy = aa, ab, ba$, and bb .

We impose an assumption that the B -spline coefficients $d_j^a(\varepsilon_t^{(a)})$ and $d_j^b(\varepsilon_t^{(b)})$ are independently and identically Gaussian distributed with mean zero and constant variance $\sigma_{j,a}^2$ and $\sigma_{j,b}^2$, respectively. Following Geman and Hwang (1982), we define the likelihood function for (5) over the approximating subspace (4) of the original parameter space. The transition density is as follows:

$$g(X_t^{(a)}, X_t^{(b)}, X_{t-1}^{(a)}, X_{t-1}^{(b)}, \rho^{aa}, \rho^{ab}, \rho^{ba}, \rho^{bb}) \\ = \frac{1}{(2\pi)^{Kn/2}} |I_n \otimes \Sigma_u|^{-1/2} \\ \times \exp \left\{ -\frac{1}{2} (\mathbf{y} - (Z^\top \otimes I_K) \boldsymbol{\beta})^\top (I_n \otimes \Sigma_u^{-1}) (\mathbf{y} - (Z^\top \otimes I_K) \boldsymbol{\beta}) \right\},$$

where $\mathbf{y} = \text{vec}(y_1, \dots, y_n)$, $Z = [{}_{y_0 \dots y_{n-1}}^1 \dots 1]$, $\boldsymbol{\beta} = \text{vec}(\nu, C)$, $\mathbf{u} = \text{vec}(u_1, \dots, u_n)$, $K = 2J_n$, I_n is an $n \times n$ identity matrix, and vec is the column stacking operator.

The Maximum Likelihood Estimators (MLEs) are obtained with closed-form:

$$\widehat{\boldsymbol{\beta}} = \left\{ (ZZ^\top)^{-1} Z \otimes I_K \right\} \mathbf{y} \quad \text{or equivalently,} \\ \widehat{B} = (\widehat{\nu}, \widehat{C}) = YZ^\top (ZZ^\top)^{-1}, \quad (6) \\ \widehat{\Sigma}_u = \frac{1}{n} (Y - BZ)(Y - BZ)^\top,$$

where the first column of \widehat{B} in (6) contains the estimators of coefficients for the mean function $\nu = (p_1^a, \dots, p_{J_n}^a, p_1^b, \dots, p_{J_n}^b)^\top$. Let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_1, \theta_2, \dots, \theta_2)$, with $\theta_1 = (c_1^{aa}, \dots, c_{J_n}^{aa}, c_1^{ba}, \dots, c_{J_n}^{ba})^\top$ and $\theta_2 = (c_1^{ab}, \dots, c_{J_n}^{ab}, c_1^{bb}, \dots, c_{J_n}^{bb})^\top$,

such that $\boldsymbol{\theta}$ contains J_n columns of θ_1 and J_n columns of θ_2 . The estimator for c_j^{xy} for $xy = aa, ab, ba, bb$ is as follows:

$$\widehat{\boldsymbol{\theta}} = Q^{-1} YZ^\top (ZZ^\top)^{-1} (\mathbf{0}_{2J_n \times 1}, I_{2J_n})^\top W,$$

where $W = \text{diag}(m/(w_{1+m} - w_1), \dots, m/(w_{J_n+m} - w_{J_n}), m/(w_{1+m} - w_1), \dots, m/(w_{J_n+m} - w_{J_n}))$, $Q = [{}_{\mathbf{0}_{J_n \times J_n} Q^1}^{Q^1 \mathbf{0}_{J_n \times J_n}}]$, with Q^1 being a $J_n \times J_n$ matrix with elements in the j th diagonal equals $q_j - 1$ and the remaining elements in the j th column equals q_j , $q_j = (w_{j+m} - w_{j+1})/(w_{j+m} - w_j) - (w_{j+m+1} - w_{j+2})/(w_{j+m+1} - w_{j+1})$, and $\mathbf{0}$ is the zero matrix.

3.2. Asymptotic property

We establish the consistency property of the sieve estimators. Let $H(\boldsymbol{\rho}, \boldsymbol{\psi})$ denote the conditional entropy between a set of operators $\boldsymbol{\rho} = (\rho^{aa}, \rho^{ab}, \rho^{ba}, \rho^{bb})$ and a given set of operators $\boldsymbol{\psi}$:

$$H(\boldsymbol{\rho}, \boldsymbol{\psi}) = \mathbf{E}_\rho [\log g(X_t^{(a)}, X_t^{(b)}, X_{t-1}^{(a)}, X_{t-1}^{(b)}, \boldsymbol{\psi})].$$

THEOREM 3.1 Assume $\{\Theta_{J_n}\}$ is chosen such that conditions **Con1** and **Con2** in Appendix 3 are in force. Suppose that for each $\delta > 0$, we can find subsets $\Gamma_1, \Gamma_2, \dots, \Gamma_{l_n}$ of Θ_{J_n} , $J_n = 1, 2, \dots$ such that

- (i) $D_{J_n} \subseteq \bigcup_{k=1}^{l_n} \Gamma_k$, where $D_{J_n} = \{\boldsymbol{\rho} \in \Theta_{J_n} | H(\boldsymbol{\rho}_{0|\Theta_{J_n}}, \boldsymbol{\rho}) \leq H(\boldsymbol{\rho}_{0|\Theta_{J_n}}, \boldsymbol{\rho}_{J_n}) - \delta\}$ for every $\delta > 0$ and every J_n .
- (ii) $\sum_{n=1}^{+\infty} l_n (\varphi_{J_n})^n < +\infty$, where given l sets $\Gamma_1, \dots, \Gamma_l$ in Θ_{J_n} , $\varphi_{J_n} = \sup_k \inf_{t \geq 0} \mathbf{E}_{\boldsymbol{\rho}_{0|\Theta_{J_n}}} \exp\{t \log(g(X_t^{(a)}, X_t^{(b)}, X_{t-1}^{(a)}, X_{t-1}^{(b)}, \Gamma_k)/g(X_t^{(a)}, X_t^{(b)}, X_{t-1}^{(a)}, X_{t-1}^{(b)}, \boldsymbol{\rho}_{J_n}))\}$.

Then we have $\sup_{\boldsymbol{\rho}_n \in M_{J_n}^n} \|\widehat{\boldsymbol{\rho}}_n - \boldsymbol{\rho}_{0|\Theta_{J_n}}\|_{\mathcal{S}} \rightarrow 0$ a.s.

The norm $\|\cdot\|_{\mathcal{S}}$ is a Hilbert–Schmidt norm for the convolution kernel operator and its Hilbert–Schmidt norm is $\|\rho\|_{\mathcal{S}} = (\sum_j \lambda_j^2)^{1/2}$. The use of Hilbert–Schmidt norm comes from the fact that it forms a class of operators embedded in the whole space of Hilbert–Schmidt operators and for any convolution kernel operator ρ , the Hilbert–Schmidt norm of ρ is equal to the L^2 norm of its kernel function, in particular, $\|\rho\|_{\mathcal{S}} = \|\kappa\|_2$.

Note that in Theorem 3.1, $g(X_t^{(a)}, X_t^{(b)}, X_{t-1}^{(a)}, X_{t-1}^{(b)}, \Gamma_k) = \sup_{\boldsymbol{\psi} \in \Gamma_k} g(X_t^{(a)}, X_t^{(b)}, X_{t-1}^{(a)}, X_{t-1}^{(b)}, \boldsymbol{\psi})$. We define the set of all the MLEs on Θ_{J_n} given the sample size n as $M_{J_n}^n = \{\boldsymbol{\rho} \in \Theta_{J_n} | \ell(X_1^{(a)}, \dots, X_n^{(a)}, X_1^{(b)}, \dots, X_n^{(b)}; \boldsymbol{\rho}) = \sup_{\boldsymbol{\psi} \in \Theta_{J_n}} \ell(X_1^{(a)}, \dots, X_n^{(a)}, X_1^{(b)}, \dots, X_n^{(b)}; \boldsymbol{\psi})\}$. Let $\boldsymbol{\rho}_0$ denotes the true set of values for the set of parameters $(\rho_0^{aa}, \rho_0^{ab}, \rho_0^{ba}, \rho_0^{bb})$. We follow Mourid and Bensmain (2006) for the proof of Theorem 3.1 to show the convergence of the ML estimator to $\boldsymbol{\rho}_{0|\Theta_{J_n}}$, the projections of the true operators on sieve, see Appendix 3 for details. Together with the convergence of $\boldsymbol{\rho}_{0|\Theta_{J_n}}$ to the true set of operators $\boldsymbol{\rho}_0$ as the sieve dimension grows, we prove that the ML estimator converges to the true set of operators $\boldsymbol{\rho}_0$.

THEOREM 3.2 If $J_n = \mathcal{O}(n^{1/3-\eta})$ for $\eta > 0$, then $\|\widehat{\boldsymbol{\kappa}}_{J_n} - \boldsymbol{\kappa}_{0|\Theta_{J_n}}\|_2 \rightarrow 0$ a.s. when $n \rightarrow +\infty$. $\widehat{\boldsymbol{\kappa}}_{J_n} = (\widehat{\boldsymbol{\kappa}}_{aa, J_n}, \widehat{\boldsymbol{\kappa}}_{ab, J_n}, \widehat{\boldsymbol{\kappa}}_{ba, J_n}, \widehat{\boldsymbol{\kappa}}_{bb, J_n})$ is the set of sieve estimators on Θ_{J_n} and $\boldsymbol{\kappa}_{0|\Theta_{J_n}} = (\boldsymbol{\kappa}_{aa, 0|\Theta_{J_n}}, \boldsymbol{\kappa}_{ab, 0|\Theta_{J_n}}, \boldsymbol{\kappa}_{ba, 0|\Theta_{J_n}}, \boldsymbol{\kappa}_{bb, 0|\Theta_{J_n}})$ is the projection of the set of true kernel functions $\boldsymbol{\kappa}_0$ on Θ_{J_n} . $\|\widehat{\boldsymbol{\kappa}}_{J_n} - \boldsymbol{\kappa}_{0|\Theta_{J_n}}\|_2 \rightarrow$

Table 2. R^2 , RMSE, and MAPE for in-sample estimation of the 12 stocks.

Ticker symbol	VFAR			RW vs. VFAR		
	R^2 (%)	RMSE	MAPE (%)	R^2	RMSE	MAPE
AAPL	92.03	0.34	3.61	0.97	1.18	1.05
MSFT	95.19	0.18	0.95	0.98	1.16	1.07
INTC	94.79	0.19	0.92	0.98	1.15	1.07
CSCO	96.16	0.19	0.86	0.99	1.13	1.06
SIRI	98.29	0.09	0.29	1.00	1.09	1.00
AMAT	95.83	0.18	0.89	0.99	1.15	1.09
CMCSA	93.39	0.19	1.20	0.97	1.18	1.13
AEZS	98.48	0.42	2.18	0.98	1.45	1.05
EBAY	94.88	0.23	1.55	0.98	1.15	1.06
MU	95.14	0.26	1.17	0.98	1.16	1.08
WFM	95.52	0.20	1.57	0.98	1.16	1.01
SBUX	94.77	0.22	2.51	0.98	1.17	1.05

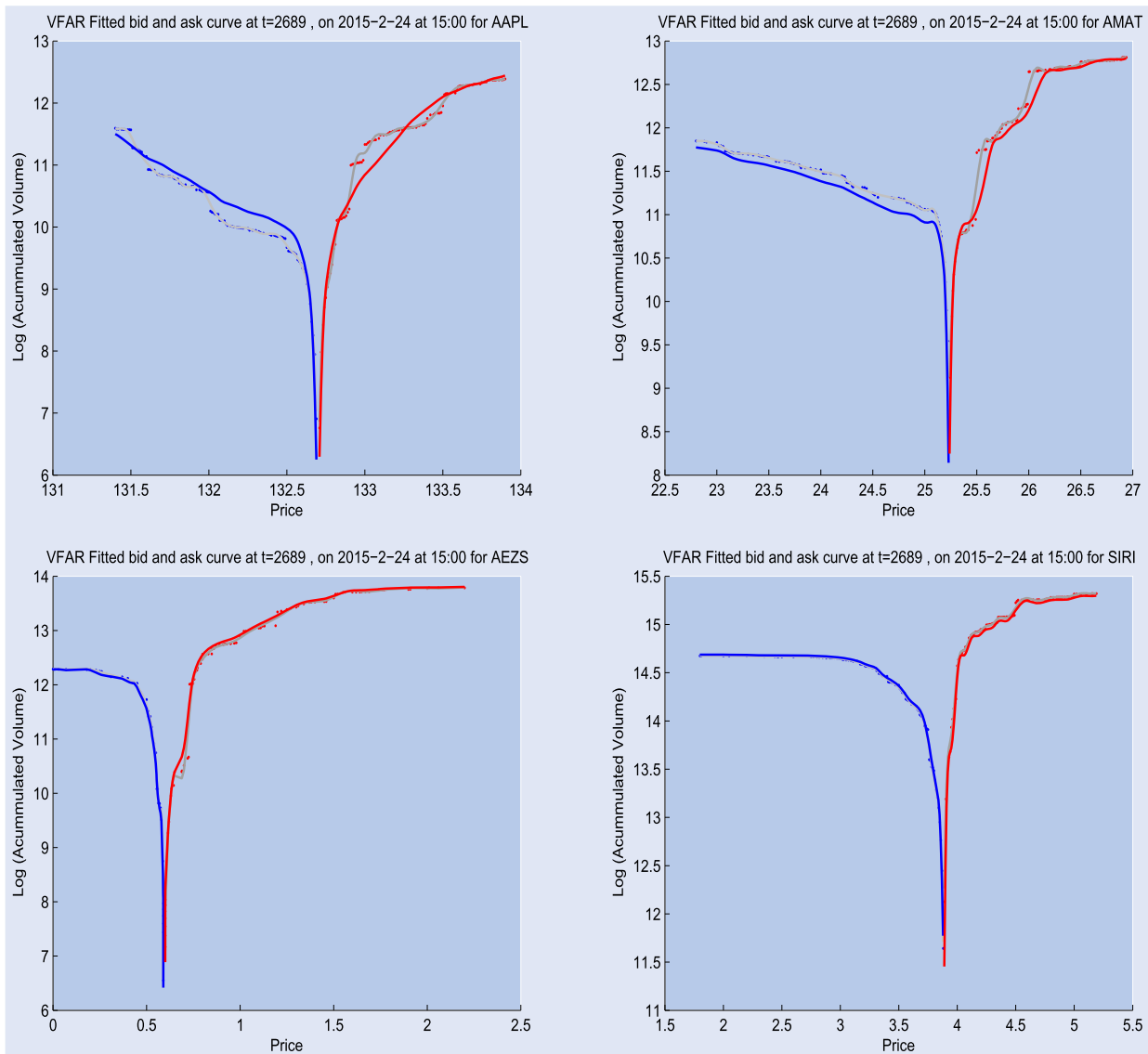


Figure 4. Estimated bid (and ask) supply curves vs. the actually observed.

0 a.s. means that each $\|\widehat{\kappa}_{xy,J_n} - \kappa_{xy,0|\Theta_{J_n}}\|_2 \rightarrow 0$ a.s. for $xy = aa, ab, ba, bb$.

By checking the conditions of Theorem 3.1, we can achieve the proof of Theorem 3.2. The proof is detailed in Appendix 4.

As $n, J_n \rightarrow \infty$, we have $\kappa_{0|\Theta_{J_n}} \rightarrow \kappa_0$ as $\kappa_{xy,0|\Theta_{J_n}}$ in $\kappa_{0|\Theta_{J_n}}$ is just the B -spline truncation of the corresponding true kernel $\kappa_{xy,0}$ in κ_0 on Θ_{J_n} . Finally we have the sieve estimator $\widehat{\kappa}_{J_n}$ converges to the true set of kernel functions κ_0 .

4. Modeling liquidity demand and supply curves

We apply the convolutional VFAR model to study the joint dynamics of the liquidity demand and supply curves in the LOB. We investigate the in-sample and out-of-sample predictability based on the records of the 12 stocks with high variations over 44 trading days from date 2 January 2015 to 6 March 2015. We evaluate the accuracy of prediction and also demonstrate the application of the VFAR forecast in order execution strategy.

On each day, the liquidity demand and supply curves are obtained by using the *B*-spline expansions on the log-accumulated volumes. Throughout the analysis, we assumed prices are known exogenous variables. The equally spaced price percentiles are used as nodes and $J_n = 20$ is chosen in the sieve. The value of J_n is selected for giving on average the highest explanatory power over all the 12 stocks in our analysis. There are in total 20 coefficients on the bid side and another 20 on the ask side.

One may suspect co-integration between the bid and ask sides, though there is no empirical evidence on the existence of co-integration in figure 2, we also consider the Random Walk (RW) model as an alternative, where the liquidity curves are predicted by the most recent curves at the previous time point. The selection of random walk is also motivated by the fact the it provides a general good predictability and is hard to beat under market efficiency.

Three measures are used to evaluate the prediction performance. They are Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) for accuracy, and R^2 for the explanatory power:

$$\begin{aligned}
 \text{RMSE} &= \sqrt{\frac{\sum_{xy=a,b} \sum_{t=1}^n \sum_{\tau} \left\{ X_t^{(xy)}(\tau) - \widehat{X}_t^{(xy)}(\tau) \right\}^2}{\sum_{t=1}^n N_t}}, \\
 \text{MAPE} &= \frac{\sum_{xy=a,b} \sum_{t=1}^n \sum_{\tau} \frac{\left| X_t^{(xy)}(\tau) - \widehat{X}_t^{(xy)}(\tau) \right|}{X_t^{(xy)}(\tau)}}{\sum_{t=1}^n N_t}, \\
 R^2 &= 1 - \frac{\sum_{xy=a,b} \sum_{t=1}^n \sum_{\tau} \left\{ X_t^{(xy)}(\tau) - \widehat{X}_t^{(xy)}(\tau) \right\}^2}{\sum_{xy=a,b} \sum_{t=1}^n \sum_{\tau} \left\{ X_t^{(xy)}(\tau) - \bar{X} \right\}^2}, \tag{7}
 \end{aligned}$$

where X is the actual value, \widehat{X} denotes the estimate or forecast, and N_t is the total number of the observed price quotes on both sides of LOB at time point t . For each stock, we calculate these measures for the estimated/forecasted liquidity curves using the VFAR model and the alternative RW model, respectively.

4.1. In-sample estimation

We conduct the in-sample estimation over the whole time period of the 44 days. Table 2 reports the R^2 , RMSE and MAPE of the estimated liquidity curves. It shows that VFAR provides high explanatory power for all the stocks, with R^2 ranging from 92 % (AAPL) to 98 % (AEZS), and superior prediction accuracy with RMSE smaller than 0.42 (AEZS)

and MAPE lower than 3.61 % (AAPL). We compare the performance of VFAR and the alternative RW model. On the right panel, the ratio of each measure is computed for the estimates based on the RW model against those on the VFAR model. The best relative performance is marked in bold-face. Without exception, the VFAR model is better than the alternative. In terms of R^2 , VFAR outperforms by up to 3 % (AAPL the largest stock and CMCSA the least active stock). As for estimation accuracy, the relative performance reaches to 13 % in MAPE (CMCSA) and at least 9 % (SIRI, the most

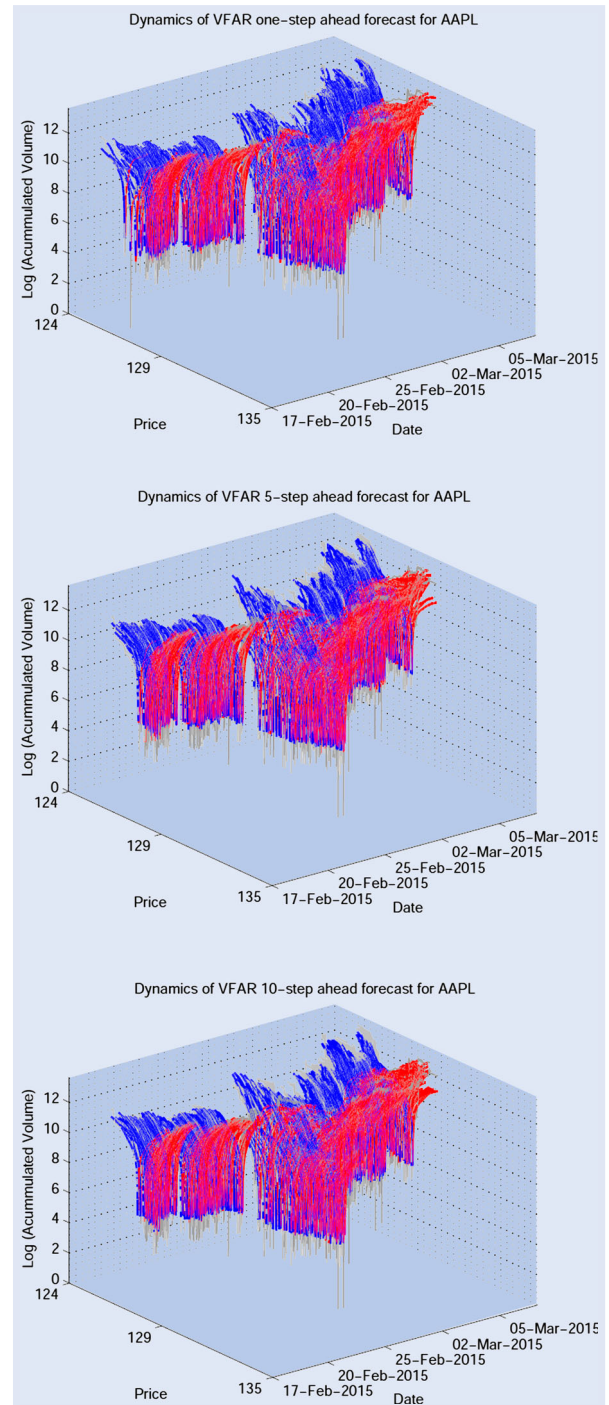


Figure 5. Dynamics of multi-step ahead forecast for AAPL. Top: 5-min ahead forecast; middle: 25-min ahead forecast; bottom: 50-min ahead forecast.

Table 3. R^2 , RMSE, MAPE for multi-step ahead VFAR forecast of the 12 stocks.

Ticker symbol	R^2 (%)			RMSE			MAPE (%)		
	1-step	5-steps	10-steps	1-step	5-steps	10-steps	1-step	5-steps	10-steps
AAPL	91.13	85.64	83.74	0.37	0.48	0.51	3.61	4.21	4.49
MSFT	95.38	91.56	89.65	0.18	0.24	0.27	0.93	1.42	1.63
INTC	94.02	89.44	86.95	0.19	0.26	0.28	0.95	1.46	1.74
CSCO	96.67	93.07	90.35	0.21	0.31	0.36	0.88	1.44	1.76
SIRI	98.14	96.23	95.31	0.09	0.13	0.14	0.30	0.52	0.62
AMAT	95.47	92.17	89.99	0.19	0.25	0.29	1.00	1.47	1.72
CMCSA	92.80	89.22	88.09	0.20	0.24	0.25	1.15	1.56	1.69
AEZS	98.23	97.71	97.43	0.48	0.55	0.58	2.22	2.85	3.14
EBAY	94.64	91.54	89.74	0.23	0.29	0.32	1.31	1.79	2.03
MU	95.37	92.35	90.60	0.22	0.28	0.31	1.18	1.70	1.99
WFM	95.18	92.24	91.15	0.20	0.26	0.27	1.25	1.76	1.94
SBUX	94.49	91.82	90.63	0.23	0.28	0.30	1.81	2.27	2.48

Table 4. Ratio of R^2 , RMSE, MAPE for multi-step ahead RW forecast to VFAR forecast of the 12 stocks.

Ticker symbol	R^2			RMSE			MAPE		
	1-step	5-steps	10-steps	1-step	5-steps	10-steps	1-step	5-steps	10-steps
AAPL	0.97	0.93	0.90	1.15	1.19	1.22	1.03	1.11	1.13
MSFT	0.99	0.98	0.97	1.10	1.12	1.14	0.97	0.98	1.02
INTC	0.98	0.96	0.94	1.12	1.16	1.19	1.02	1.05	1.07
CSCO	0.99	0.98	0.97	1.09	1.13	1.14	1.01	1.01	1.05
SIRI	1.00	0.99	0.99	1.04	1.10	1.09	0.90	0.90	0.90
AMAT	0.99	0.97	0.96	1.11	1.16	1.18	1.01	1.05	1.09
CMCSA	0.98	0.94	0.91	1.14	1.23	1.28	1.06	1.13	1.19
AEZS	0.98	0.98	0.98	1.36	1.40	1.39	1.05	1.12	1.12
EBAY	0.99	0.96	0.94	1.12	1.19	1.23	1.05	1.11	1.16
MU	0.98	0.96	0.95	1.15	1.20	1.22	1.08	1.12	1.15
WFM	0.99	0.96	0.94	1.13	1.23	1.29	1.05	1.12	1.20
SBUX	0.98	0.96	0.94	1.15	1.21	1.24	1.03	1.11	1.15

active stock) and up to 45 % (AEZS that has the smallest bid-ask spread on average) in RMSE. We find the superior performance of the VFAR is robust with respect to market capitalization, market tightness and depth.

Figure 4 visualizes the fitted liquidity demand and supply curves and the actual values at an arbitrarily selected date, 24 February 2015 at 3 pm, of the four representative stocks, AAPL, AMAT, AEZS and SIRI. The estimated curves reasonably trace the queuing orders displayed as discrete dots as well as the smoothed liquidity curves in gray color. The accuracy is quite stable, especially in the middle around the best quotes as well as the extreme in the tails.

4.2. Out-of-sample forecast

In this section, we analyze the model's forecasting performance in a realistic setup. In particular, a trader is assumed to observe the LOB at 5-min snapshots, with the information over the past 30 trading days. The trader can only submit orders every 5 min and thus asks for multi-step ahead out-of-sample forecasts for the liquidity curves, starting from the 31st trading day onwards. Among others, the trader is interested in 1-, 5- and 10-step ahead forecasts that correspond to 5-, 25- and 50-min ahead liquidity curves, respectively. As such, the first pair of the forecasted curves is for time $t = 2251$, based on the past 30 trading days of $30 \times 75 = 2250$ functional objects. Each time, he moves forward one period, i.e.

5 min and performs re-estimation and forecast until reaching the end of the sample at $t = 3300$.

Figure 5 gives graphical illustrations of the forecasted liquidity curves for AAPL with the VFAR model. The forecasts closely trace the realized liquidity curves. It is remarkable that the VFAR model is able to catch the dynamic movements of the liquidity curves over the period from 17 February to 06 March 2015 for different forecast horizon from 5- to 50-min.

Table 3 reports the RMSE, MAPE and predictive power of the liquidity curves forecast for the 12 stocks. Even if in the 'worst' case, the VFAR approach is able to achieve high R^2 ranging from 91.13 % (1-step AAPL) to 83.74 % (10-step AAPL), low RMSE of 0.48 (1-step AEZS) to 0.58 (10-step AEZS), and low MAPE of 3.61 % (1-step AAPL) to 4.49 % (10-step AAPL). The relative performance of the alternative RW model is summarized in table 4. Again, the VFAR model dominates the RW model across forecast horizons and forecast measures. Though the improvement in R^2 is weak, the advantage is obvious in terms of the forecast error reduction. In terms of RMSE, the VFAR model reaches about 4 % (1-step SIRI) in the worst case and 36 % (1-step AEZS) and 40 % (5-step AEZS) in the best case. On the other hand, the VFAR model does not always yield improvement in the MAPE comparison. However, the RW performs better than VFAR only in 5 out of 36 instances. In other cases, VFAR outperforms the RW by up to 20 %. The relative superior performance grows as the forecast horizon increases, indicating that the

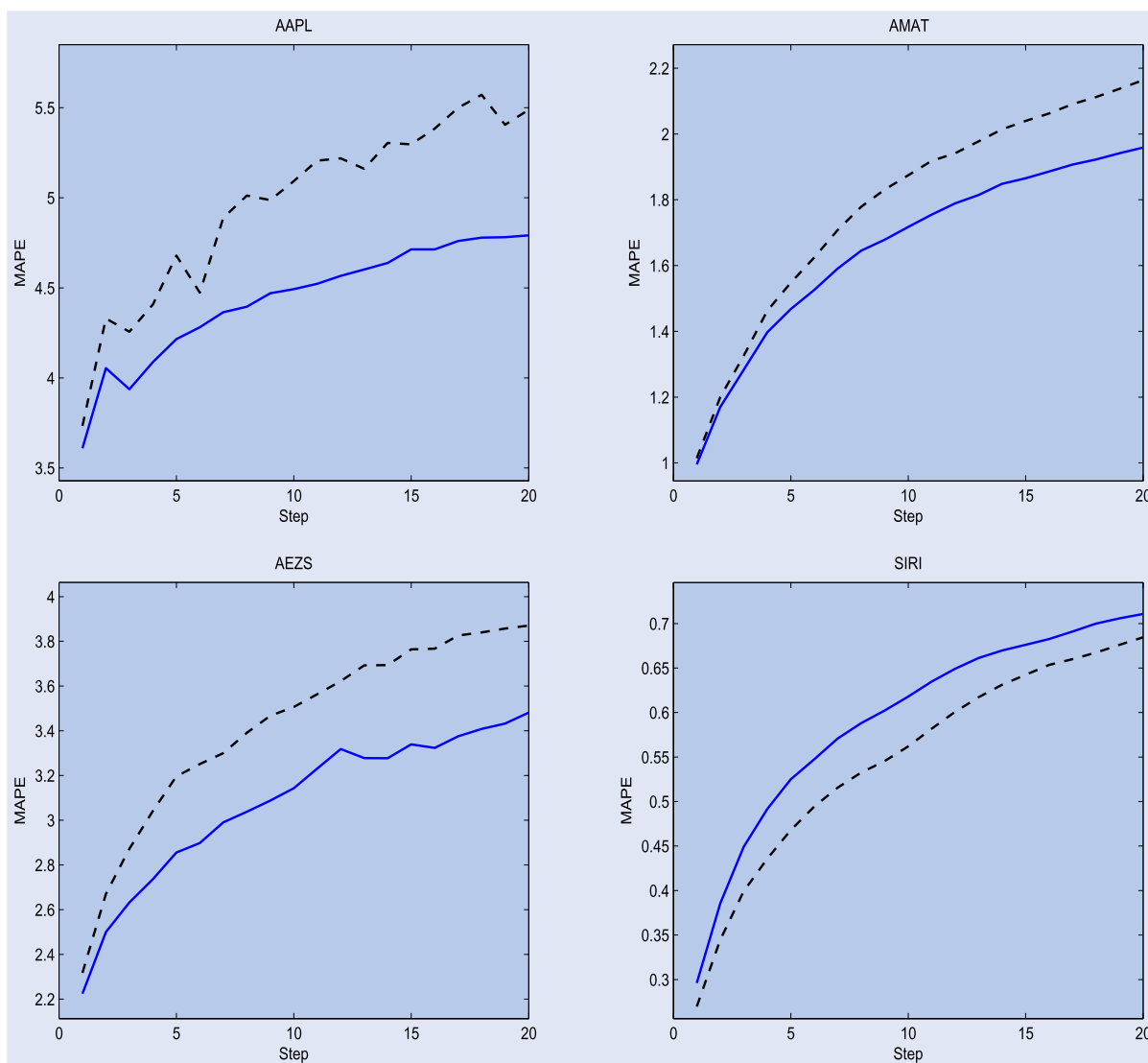


Figure 6. Mean absolute percentage errors (MAPEs) implied by the VFAR approach (blue) and the RW approach (black dashed) for different number of steps ahead forecasts, corresponding to 5–100 min.

utilization of cross-dependence in liquidity curves helps to improve out-of-sample prediction.

We also find that the strong predictability of the VFAR model is stable with respect to forecast horizons. Figure 6 displays the MAPEs for various multi-steps ahead forecasts ranging from 1 to 20 steps ahead for the four representative stocks AAPL, AMAT, AEZS, and SIRI. Except SIRI, the VFAR forecasts outperform the RW with lower MAPEs. As we forecast further into the future, the advantage of VFAR over RW increases for AAPL, AMAT and AEZS. As the most active stock considered in our data analysis, SIRI has weaker predictive power in the out-of-sample forecasting. It is the only asset for which MAPE is greater than the random walk specification alternative, although this difference shrinks as the steps forward increase. This weaker predictive power could be due to the fact that SIRI is the most active stock, and its dynamics is difficult to characterize, even by the VFAR.

To summarize, the proposed VFAR model is able to successfully predict the liquidity curves over various forecasting periods. These results can be applied to various financial and

economics applications, and we will show next an application to order execution strategy as an example.

4.3. Application to order execution strategy

In this section, we show how to utilize the forecasting results in the previous section to a practical application on order execution. Assume that an investor decides to buy (sell) ν number of shares over a course of a trading day, starting from 9:45 to 15:55. The volume ν to be traded is chosen to be 5 or 10 times the average pending volume at the best bid (ask) price, yielding the following buy (sell) quantities in the respective two cases of (a) high and (b) very high liquidity demand:

- (a) AAPL-4000 (4000); MSFT-26 000 (29 000); INTC-27 000 (31 000); CSCO-67 000 (53 000); SIRI-675 000 (654 000); AMAT-17 000 (18 000); CMCSA-7000 (8000); AEZS-12 000 (17 000); EBAY-6000 (6000); MU-8000 (8000); WFM-2000 (3000); SBUX-2000 (2000).

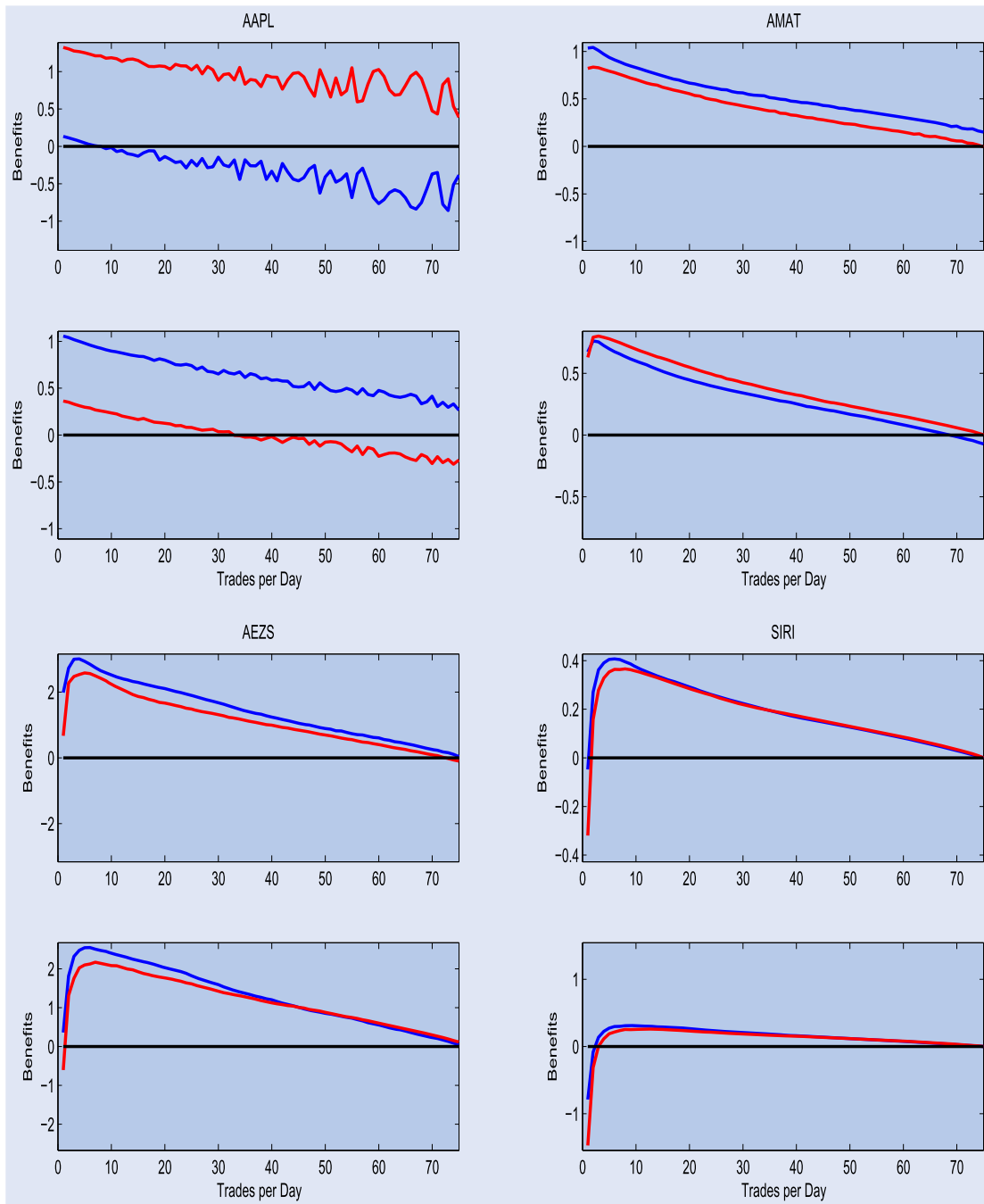


Figure 7. Average percentage gains by strategy (ii) in excess of the equal-splitting strategy (i) when buying (blue) and selling (red) shares based on m VFAR-predicted time points per day. For each stock, top: high liquidity demand corresponding to 5 times the average first level market depth; bottom: very high liquidity demand corresponding to 10 times the average first level market depth.

(b) AAPL-8000 (8000); MSFT-52 000 (58 000); INTC-54 000 (62 000); CSCO-134 000 (106 000); SIRI-1 350 000 (1 308 000); AMAT-34 000 (36 000); CMCSA-14 000 (16 000); AEZS-24 000 (34 000); EBAY-12 000 (12 000); MU-16 000 (16 000); WFM-4000 (6000); SBUX-4000 (4000).

Suppose that trading is only performed on a 5 min grid throughout the day corresponding to 75 possible trading time points, and investor can only make decision at 9:45 but not monitor the market anymore during the day. The forecasting horizon covers $h = 75$ periods on each trading day. The following are two execution strategies the investor can use:

- (i) Splitting the buy (sell) order of size v equally in 5-min frequency over the trading day, resulting into 75 trades of size $v/75$ each; see Almgren and Chriss (2000).
- (ii) Placing different orders at m (5-min interval) time points throughout the day where the VFAR predicted implied trading costs c of volume v are smallest. The volume is split over the m time points according to the relative proportion of expected trading costs, i.e. at time point i , $w_i \cdot v$ shares are traded, with $w_i = c_i / \sum_{j=1}^m c_j$ for $i = 1, \dots, m$; see Härdle et al. (2012).

Note that strategy (i) is a special case of strategy (ii) with $m = 75$ and the volume v is equally split. For strategy (ii),

if $m = 1$, it is the extreme case where the whole quantity is traded only once. The VFAR prediction of the trading costs involves predicting the whole bid and ask curve at each time point and compute the effective cost of trading using prevailing bid and ask quotes assumed to be known at the respective time points.

We implement these strategies to 14 forecasting days, from 17 February to 06 March 2015. Figure 7 shows the average percentage reduction in trading costs of strategy (ii) in excess of the equal-splitting strategy (i) for different values of $m \in [1, 75]$ for AAPL, AMAT, AEZS, and SIRI. Overall, we observe that strategic placement of orders according to VFAR predictions achieve excess gains of 31 basis points on average. Generally, the behavior of the curves is similar, increasing as m increases from 1, and converge to zero as m reaches the upper limit of 75. The pattern shows that making small number of large market orders is superior to an equal-splitting strategy, while for the extreme case where $m = 1$, there is lesser benefit or even loss since the transactions have to walk up the book too severely and cause large price impacts; and where $m = 75$, the relative benefit only results from the strategic non-equal weighting scheme. All in all, the VFAR model is successful in predicting times when the market is sufficiently deep to execute large orders.

5. Conclusion

Predictions of future liquidity supply and demand in the limit order book (LOB) help in analyzing optimal splitting strategies for large orders to reduce cost. To capture not only the volume around the best bid and ask price in the LOB, but also the pending volumes more deeply in the book, it becomes an ultra-high dimensional problem. Motivated by the significant cross-dependency of the bid and ask side of the market, we proposed a Vector Functional AutoRegressive (VFAR) model to estimate and forecast the liquidity supply and demand curves in the functional domain.

The model is applied to 12 stocks traded in the National Association of Securities Dealers Automated Quotations (NASDAQ) stock market. It is shown that the VFAR model gives R^2 values as high as 98.5 % for in-sample estimation. In out-of-sample forecast experiments, it produces accurate 5-, 25- and 50-min forecasts, with MAPE as low as 0.3–4.5%. The predictive power of the VFAR model can be further used to improve order execution strategies at lower trading cost.

Our results contribute to the finance domain in helping both practitioners and academics to better understand the dynamics of available liquidity of a LOB and aiding to construct a forward-looking trading strategy. In the area of financial econometrics, we extend the VAR framework to the functional domain, develop the method, and derive the theoretical results. For future work, one can consider adding exogenous variables to the VFAR model, extending the VARX framework in the functional domain, to develop the VFARX model. In application, one can jointly forecast the liquidity supply and demand of a basket of stocks, rather than just an individual stock that was presented in this paper, and include other

exogenous variables (in VFARX) such as oil prices, or electricity prices, that may affect the liquidity of certain stocks. Most important, the VFAR model is general, has high interpretability, and can be used for other multiple functional time series modeling and forecasting.

Acknowledgments

We would like to thank the editor and two anonymous referees for their constructive comments to help improve the quality of this paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The research of Ying Chen is supported by the Academic Research Funding R-155-000-178-114 and IDS Funding R-155-000-185-64 at the National University of Singapore. Support from IRTG 1792 ‘High Dimensional Non Stationary Time Series’, Humboldt-Universität zu Berlin, is gratefully acknowledged.

References

- Ait-Sahalia, Y., Mykland, P. and Zhang, L., How often to sample a continuous-time process in the presence of market microstructure noise. *Rev. Financ. Stud.*, 2005, **18**, 351–416.
- Almgren, R. and Chriss, N., Optimal execution of portfolio transactions. *J. Risk.*, 2000, **3**, 5–39.
- Antoniadis, A. and Sapatinas, T., Wavelet methods for continuous-time prediction using Hilbert-valued autoregressive processes. *J. Multivar. Anal.*, 2003, **87**, 135–158.
- Benston, G. and Hagerman, R., Determinants of bid-asked spreads in the over-the-counter market. *J. Financ. Econ.*, 1974, **1**, 353–364.
- Besse, P., Cardot, H. and Stephenson, D., Autoregressive forecasting of some functional climatic variations. *Scand. J. Stat.*, 2000, **27**, 673–687.
- Bosq, D., *Linear Processes in Function Spaces: Theory and Applications*, 2000 (Springer: New York).
- Chaudhuri, K., Kim, M. and Shin, Y., Forecasting distributions of inflation rates: A functional Autoregressive Approach. *J. R. Stat. Soc.: Ser. A (Stat. Soc.)*, 2016, **179**, 65–102.
- Chen, Y. and Li, B., An adaptive functional autoregressive forecast model to predict electricity price curves. *J. Bus. Econ. Stat.*, 2017, **35**(3), 371–388.
- Chordia, T., Sarkar, A. and Subrahmanyam, A., An empirical analysis of stock and bond market liquidity. *Rev. Financial Stud.*, 2005, **18**(1), 85–129.
- Cooper, K., Groth, J. and Avera, W., Liquidity, exchange listing, and common stock performance. *J. Econ. Bus.*, 1985, **37**, 19–33.
- Fleming, M. and Remolona, E., Price formation and liquidity in the U.S. treasury market: The response to public information. *J. Financ.*, 1999, **54**, 1901–1915.
- Foucault, T., Kadan, O. and Kandel, E., Limit order book as a market for liquidity. *Rev. Financ. Stud.*, 2005, **18**, 1171–1217.

Geman, S. and Hwang, C.R., Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Stat.*, 1982, **10**, 401–414.

Gomber, P., Schweickert, U. and Theissen, E., Liquidity dynamics in an electronic open limit order book: An event study approach. *Eur. Financ. Manage.*, 2015, **21**, 52–78.

Grenander, U., *Abstract Inference*, 1981 (Wiley: New York).

Groß-Klußmann, A. and Hautsch, N., Predicting bid ask spreads using long-memory autoregressive conditional Poisson models. *J. Forecast.*, 2013, **32**, 724–742.

Guillas, S., Rates of convergence of autocorrelation estimates for autoregressive Hilbertian processes. *Stat. Probab. Lett.*, 2001, **55**, 281–291.

Härdle, W., Hautsch, N. and Mihoci, A., Modelling and forecasting liquidity supply using semiparametric factor dynamics. *J. Empir. Financ.*, 2012, **19**, 610–625.

Härdle, W., Hautsch, N. and Mihoci, A., Local adaptive multiplicative error models for high-frequency forecasts. *J. Appl. Economet.*, 2015, **30**, 529–550.

Härdle, W. K., Chen, S., Liang, C. and Schienle, M., Time-varying limit order book networks. IRTG 1792 Discussion Paper 2018-016, IRTG 1792, Humboldt Universität zu Berlin, Germany, 2018.

Harris, L., Liquidity, trading rules and electronic trading systems. Technical report, NYU Salomon Center Series in Finance and Economics, 1990.

Huberman, G. and Halka, D., Systematic liquidity. *J. Financ. Res.*, 2001, **24**, 161–178.

Hwang, C., Gaussian measure of large balls in a Hilbert space. *Proc. Am. Math. Soc.*, 1980, **78**, 107–110.

Johansen, S., Estimating and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica*, 1991, **59**, 1551–1580.

Mourid, T. and Bensmain, N., Sieves estimator of the operator of a functional autoregressive process. *Stat. Probab. Lett.*, 2006, **76**, 93–108.

Stoll, H., The pricing of security dealer services: An empirical study of NASDAQ stocks. *J. Financ.*, 1978, **33**, 1153–1172.

Zhang, L., Mykland, P. and Ait-Sahalia, Y., A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *J. Am. Stat. Assoc.*, 2005, **100**, 1394–1411.

Appendix 1. Derivation of the B-spline coefficient relationship as shown in Section 3

$$\begin{aligned}
 X_t^{(a)}(\tau) &= \sum_{h=1}^{\infty} p_h^a B_{h,m}(\tau) \\
 &+ \int_0^1 \left\{ \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} c_j^{aa} d_{t-1,i}^a B_{j,m}(\tau-s) B_{i,m}(s) \right\} ds \\
 &+ \int_0^1 \left\{ \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} c_j^{ab} d_{t-1,i}^b B_{j,m}(\tau-s) B_{i,m}(s) \right\} ds \\
 &+ \sum_{j=1}^{\infty} d_j^a(\varepsilon_t^{(a)}) B_{j,m}(\tau) \\
 &= \sum_{h=1}^{\infty} p_h^a B_{h,m}(\tau) + \sum_{j=1}^{\infty} d_j^a(\varepsilon_t^{(a)}) B_{j,m}(\tau) \\
 &+ \sum_{h=1}^{\infty} \sum_{i=1}^{\infty}
 \end{aligned}$$

$$\begin{aligned}
 &\times \left\{ \sum_{j=1}^{\infty} \left(\frac{w_{j+m} - w_{j+1}}{w_{j+m} - w_j} - \frac{w_{j+m+1} - w_{j+2}}{w_{j+m+1} - w_{j+1}} \right) c_j^{aa} - c_h^{aa} \right\} \\
 &\times \frac{w_{i+m} - w_i}{m} d_{t-1,i}^a B_{h,m}(\tau) \\
 &+ \sum_{h=1}^{\infty} \sum_{i=1}^{\infty} \\
 &\times \left\{ \sum_{j=1}^{\infty} \left(\frac{w_{j+m} - w_{j+1}}{w_{j+m} - w_j} - \frac{w_{j+m+1} - w_{j+2}}{w_{j+m+1} - w_{j+1}} \right) c_j^{ab} - c_h^{ab} \right\} \\
 &\times \frac{w_{i+m} - w_i}{m} d_{t-1,i}^b B_{h,m}(\tau), \\
 X_t^{(b)}(\tau) &= \sum_{h=1}^{\infty} p_h^b B_{h,m}(\tau) \\
 &+ \int_0^1 \left\{ \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} c_j^{bb} d_{t-1,i}^b B_{j,m}(\tau-s) B_{i,m}(s) \right\} ds \\
 &+ \int_0^1 \left\{ \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} c_j^{ba} d_{t-1,i}^a B_{j,m}(\tau-s) B_{i,m}(s) \right\} ds \\
 &+ \sum_{j=1}^{\infty} d_j^b(\varepsilon_t^{(b)}) B_{j,m}(\tau) \\
 &= \sum_{h=1}^{\infty} p_h^b B_{h,m}(\tau) + \sum_{j=1}^{\infty} d_j^b(\varepsilon_t^{(b)}) B_{j,m}(\tau) \\
 &+ \sum_{h=1}^{\infty} \sum_{i=1}^{\infty} \\
 &\times \left\{ \sum_{j=1}^{\infty} \left(\frac{w_{j+m} - w_{j+1}}{w_{j+m} - w_j} - \frac{w_{j+m+1} - w_{j+2}}{w_{j+m+1} - w_{j+1}} \right) c_j^{bb} - c_h^{bb} \right\} \\
 &\times \frac{w_{i+m} - w_i}{m} d_{t-1,i}^b B_{h,m}(\tau) \\
 &+ \sum_{h=1}^{\infty} \sum_{i=1}^{\infty} \\
 &\times \left\{ \sum_{j=1}^{\infty} \left(\frac{w_{j+m} - w_{j+1}}{w_{j+m} - w_j} - \frac{w_{j+m+1} - w_{j+2}}{w_{j+m+1} - w_{j+1}} \right) c_j^{ba} - c_h^{ba} \right\} \\
 &\times \frac{w_{i+m} - w_i}{m} d_{t-1,i}^a B_{h,m}(\tau). \tag{A1}
 \end{aligned}$$

Rearranging the above equations gives the relationship of the B-spline coefficients in (3).

Next we show how the expansion was obtained in (A1). We only show for the first integral in (A1) as the expansion other integrals can be obtained similarly.

$$\begin{aligned}
 &\int_0^1 \left\{ \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} c_j^{aa} d_{t-1,i}^a B_{j,m}(\tau-s) B_{i,m}(s) \right\} ds \\
 &= \sum_{j=1}^{\infty} c_j^{aa} \int_0^1 B_{j,m}(\tau-s) \left\{ \sum_{i=1}^{\infty} d_{t-1,i}^a B_{i,m}(s) \right\} ds \\
 &= \sum_{j=1}^{\infty} c_j^{aa} \left\{ \frac{1}{m} \sum_{i=1}^{\infty} d_{t-1,i}^a (w_{i+m} - w_i) B_{j,m}(\tau-s) \right\} \Big|_{s=0}^{s=1}
 \end{aligned}$$

$$\begin{aligned}
& + \sum_{j=1}^{\infty} c_j^{aa} \int_0^1 \left[\frac{1}{m} \sum_{i=1}^{\infty} d_{t-1,i}^a (w_{i+m} - w_i) \right. \\
& \left. \left\{ \frac{m}{w_{j+m} - w_j} B_{j,m-1}(\tau - s) \right. \right. \\
& \quad \left. \left. - \frac{m}{w_{j+m+1} - w_{j+1}} B_{j+1,m-1}(\tau - s) \right\} \right] ds \\
& = - \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \frac{d_{t-1,i}^a (w_{i+m} - w_i)}{m} c_j^{aa} B_{j,m}(\tau) \\
& \quad + \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \frac{d_{t-1,i}^a (w_{i+m} - w_i)}{m} c_j^{aa} \int_{\tau-1}^{\tau} \left\{ \frac{m}{w_{j+m} - w_j} B_{j,m-1}(z) \right. \\
& \quad \left. - \frac{m}{w_{j+m+1} - w_{j+1}} B_{j+1,m-1}(z) \right\} dz \\
& = - \sum_{h=1}^{\infty} \sum_{i=1}^{\infty} \frac{d_{t-1,i}^a (w_{i+m} - w_i)}{m} c_h^{aa} B_{h,m}(\tau) \\
& \quad + \sum_{h=1}^{\infty} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \left(\frac{w_{j+m} - w_{j+1}}{w_{j+m} - w_j} - \frac{w_{j+m+1} - w_{j+2}}{w_{j+m+1} - w_{j+1}} \right) \\
& \quad \times c_j^{aa} \frac{w_{i+m} - w_i}{m} d_{t-1,i}^a B_{h,m}(\tau) \\
& = \sum_{h=1}^{\infty} \sum_{i=1}^{\infty} \left[\sum_{j=1}^{\infty} \left\{ \frac{w_{j+m} - w_{j+1}}{w_{j+m} - w_j} - \frac{w_{j+m+1} - w_{j+2}}{w_{j+m+1} - w_{j+1}} \right\} c_j^{aa} - c_h^{aa} \right] \\
& \quad \times \frac{w_{i+m} - w_i}{m} d_{t-1,i}^a B_{h,m}(\tau).
\end{aligned}$$

The second equality made use of integration by parts, with $(d/ds)B_{j,m}(\tau - s) = -(m/(w_{j+m} - w_j))B_{j,m-1}(\tau - s) - (m/(w_{j+m+1} - w_{j+1}))B_{j+1,m-1}(\tau - s)$ and $\int \sum_{i=1}^{\infty} d_{t-1,i}^a B_{i,m}(s) ds = (1/m) \sum_{i=1}^{\infty} d_{t-1,i}^a (w_{i+m} - w_i)$. In the third equality, we made the substitution of $z = \tau - s$. For the fourth equality, we made use of the formula: $\int_{-\infty}^{\tau} B_{j,m}(z) dz = ((w_{j+m+1} - w_{j+1})/(m + 1)) \sum_{h=1}^{\infty} B_{h,m+1}(\tau)$, and truncating the sum up till the J th term. We also swapped the notation j for the first summation with h in the fourth equality.

Appendix 2. Derivation of the ML estimator in (eqn6)

For $t = 1, \dots, n$, we write (5) compactly as the following:

$$Y = BZ + U. \quad (\text{A2})$$

By applying vec operator to (A2) yields

$$\begin{aligned}
\text{vec}(Y) &= \text{vec}(BZ) + \text{vec}(U) \\
&= (Z^{\top} \otimes I_K) \text{vec}(B) + \text{vec}(U)
\end{aligned}$$

or equivalently,

$$\mathbf{y} = (Z^{\top} \otimes I_K) \boldsymbol{\beta} + \mathbf{u},$$

where \otimes is the Kronecker product.

Assuming

$$\mathbf{u} = \text{vec}(U) = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} \sim \mathcal{N}(0, I_n \otimes \Sigma_u),$$

the probabilistic density of \mathbf{u} is

$$f_{\mathbf{u}}(\mathbf{u}) = \frac{1}{(2\pi)^{Kn/2}} |I_n \otimes \Sigma_u|^{-1/2} \exp \left\{ -\frac{1}{2} \mathbf{u}^{\top} (I_n \otimes \Sigma_u^{-1}) \mathbf{u} \right\}.$$

In addition,

$$\mathbf{u} = \begin{bmatrix} I_K & & & 0 \\ -C & \ddots & & \\ & \ddots & \ddots & \\ 0 & & -C & I_K \end{bmatrix} (\mathbf{y} - \mathbf{v}) + \begin{bmatrix} -C \\ 0 \\ \vdots \\ 0 \end{bmatrix} \mathbf{y}_0,$$

where $\mathbf{v} = (v, \dots, v)^{\top}$ is a $(Kn \times 1)$ vector. Consequently, $\partial \mathbf{u} / \partial \mathbf{y}^{\top}$ is a lower triangular matrix with unit diagonal which has unit determinant. Therefore using $\mathbf{u} = \mathbf{y} - (Z^{\top} \otimes I_K) \boldsymbol{\beta}$, the transition density is as follows:

$$\begin{aligned}
g(X_t^{(a)}, X_t^{(b)}, X_{t-1}^{(a)}, X_{t-1}^{(b)}; \rho^{aa}, \rho^{ab}, \rho^{ba}, \rho^{bb}) &= f_{\mathbf{y}}(\mathbf{y}) = \left| \frac{\partial \mathbf{u}}{\partial \mathbf{y}^{\top}} \right| f_{\mathbf{u}}(\mathbf{u}) \\
&= \frac{1}{(2\pi)^{Kn/2}} |I_n \otimes \Sigma_u|^{-1/2} \\
&\quad \exp \left\{ -\frac{1}{2} (\mathbf{y} - (Z^{\top} \otimes I_K) \boldsymbol{\beta})^{\top} (I_n \otimes \Sigma_u^{-1}) (\mathbf{y} - (Z^{\top} \otimes I_K) \boldsymbol{\beta}) \right\}.
\end{aligned}$$

The (approximated) log-likelihood function is

$$\begin{aligned}
\ell(X_1^{(a)}, \dots, X_n^{(a)}, X_1^{(b)}, \dots, X_n^{(b)}; \rho^{aa}, \rho^{ab}, \rho^{ba}, \rho^{bb}) &= \ell(\boldsymbol{\beta}, \Sigma_u) \\
&= -\frac{Kn}{2} \log 2\pi - \frac{n}{2} \log |\Sigma_u| - \frac{1}{2} (\mathbf{y} - (Z^{\top} \otimes I_K) \boldsymbol{\beta})^{\top} \\
&\quad \times (I_n \otimes \Sigma_u^{-1}) (\mathbf{y} - (Z^{\top} \otimes I_K) \boldsymbol{\beta}) \\
&= -\frac{Kn}{2} \log 2\pi - \frac{n}{2} \log |\Sigma_u| - \frac{1}{2} \sum_{t=1}^n (y_t - v - C y_{t-1})^{\top} \\
&\quad \times \Sigma_u^{-1} (y_t - v - C y_{t-1}) \\
&= -\frac{Kn}{2} \log 2\pi - \frac{n}{2} \log |\Sigma_u| - \frac{1}{2} \sum_{t=1}^n (y_t - C y_{t-1})^{\top} \\
&\quad \times \Sigma_u^{-1} (y_t - C y_{t-1}) + \mathbf{v}^{\top} \Sigma_u^{-1} \sum_{t=1}^n (y_t - C y_{t-1}) \\
&\quad - \frac{n}{2} \mathbf{v}^{\top} \Sigma_u^{-1} \mathbf{v} \\
&= -\frac{Kn}{2} \log 2\pi - \frac{n}{2} \log |\Sigma_u| - \frac{1}{2} \text{tr} \left((Y - BZ)^{\top} \Sigma_u^{-1} (Y - BZ) \right)
\end{aligned}$$

and the first order partial differentiations are as follows:

$$\begin{aligned}
\frac{\partial \ell}{\partial \boldsymbol{\beta}} &= (Z \otimes I_K) (I_n \otimes \Sigma_u^{-1}) (\mathbf{y} - (Z^{\top} \otimes I_K) \boldsymbol{\beta}) \\
&= (Z \otimes \Sigma_u^{-1}) \mathbf{y} - (ZZ^{\top} \otimes \Sigma_u^{-1}) \boldsymbol{\beta}, \quad (\text{A3})
\end{aligned}$$

$$\frac{\partial \ell}{\partial \Sigma_u} = -\frac{n}{2} \Sigma_u^{-1} + \frac{1}{2} \Sigma_u^{-1} (Y - BZ) (Y - BZ)^{\top} \Sigma_u^{-1}.$$

By equating the first order partial derivatives in (A3) to zero, we obtain the maximum likelihood estimators in (6).

Appendix 3. Proof of Theorem 3.1

The growth of J_n is determined by the following two conditions:

Con1: If there exists a sequence $\{\rho_{J_n}\}$ such that $\rho_{J_n} \in \Theta_{J_n} \forall n$ and $H(\rho_{0|\Theta_{J_n}}, \rho_{J_n}) \rightarrow H(\rho_{0|\Theta_{J_n}}, \rho_{0|\Theta_{J_n}})$, then $\|\rho_{J_n} - \rho_{0|\Theta_{J_n}}\|_{\mathcal{S}} \rightarrow 0$; meaning each $\|\rho_{J_n}^{xy} - \rho_{0|\Theta_{J_n}}^{xy}\|_{\mathcal{S}} \rightarrow 0$, for $xy = aa, ab, ba, bb$. Here $\rho_{0|\Theta_{J_n}}$ denotes the projection of the set of true operators ρ_0 on the sieve Θ_{J_n} .

Con2: There exists a sequence $\{\rho_{J_n}\}$ described in **Con1** such that $H(\rho_{0|\Theta_{J_n}}, \rho_{J_n}) \rightarrow H(\rho_{0|\Theta_{J_n}}, \rho_{0|\Theta_{J_n}})$.

Fix $\delta > 0$. Following Mourid and Bensmain (2006), we only need to show that

$$P(D_{J_n} \cap M_{J_n}^n \neq \emptyset) = 0, \quad (\text{A4})$$

because if (A4) holds, then with probability 1

$$\inf_{\rho \in M_{J_n}^n} H(\rho_{0|\Theta_{J_n}}, \rho) \geq H(\rho_{0|\Theta_{J_n}}, \rho_{J_n}) - \delta,$$

for all n sufficiently large. Since δ is arbitrary, and

$$H(\rho_{0|\Theta_{J_n}}, \rho_{J_n}) \rightarrow H(\rho_{0|\Theta_{J_n}}, \rho_{0|\Theta_{J_n}}),$$

by condition **Con2** we deduce

$$\liminf_{\rho \in M_{J_n}^n} H(\rho_{0|\Theta_{J_n}}, \rho) \geq H(\rho_{0|\Theta_{J_n}}, \rho_{0|\Theta_{J_n}}) \quad \text{a.s.}$$

Combining with

$$H(\rho_{0|\Theta_{J_n}}, \rho) \leq H(\rho_{0|\Theta_{J_n}}, \rho_{0|\Theta_{J_n}}),$$

we have

$$\lim_{n \rightarrow +\infty} \sup_{\rho \in M_{J_n}^n} |H(\rho_{0|\Theta_{J_n}}, \rho) - H(\rho_{0|\Theta_{J_n}}, \rho_{0|\Theta_{J_n}})| = 0 \quad \text{a.s.} \quad (\text{A5})$$

Fix $\varepsilon > 0$, and for each n choose $\psi_n \in M_{J_n}^n$ such that

$$\frac{d(\rho_{0|\Theta_{J_n}}, \psi_n)}{1 + d(\rho_{0|\Theta_{J_n}}, \psi_n)} > \sup_{\rho \in M_{J_n}^n} \frac{d(\rho_{0|\Theta_{J_n}}, \rho)}{1 + d(\rho_{0|\Theta_{J_n}}, \rho)} - \varepsilon.$$

Condition **Con1** combined with (A5) imply that

$$d(\rho_{0|\Theta_{J_n}}, \psi_n) \rightarrow 0 \quad \text{a.s.}$$

Hence,

$$\limsup_{\rho \in M_{J_n}^n} \frac{d(\rho_{0|\Theta_{J_n}}, \rho)}{1 + d(\rho_{0|\Theta_{J_n}}, \rho)} \leq \varepsilon.$$

Since ε is arbitrary, we deduce that $M_{J_n}^n \rightarrow \rho_{0|\Theta_{J_n}}$, which is the desired result. Therefore, it suffices to prove (A4).

For now, n and J_n are fixed. Then

$$\begin{aligned} & (D_{J_n} \cap M_{J_n}^n \neq \emptyset) \\ & \subseteq \left\{ \sup_{\rho \in D_{J_n}} \ell(X_1^{(a)}, \dots, X_n^{(a)}, X_1^{(b)}, \dots, X_n^{(b)}; \rho) \right. \\ & \geq \ell(X_1^{(a)}, \dots, X_n^{(a)}, X_1^{(b)}, \dots, X_n^{(b)}; \rho_{J_n}) \left. \right\} \\ & \subseteq \bigcup_{k=1}^{l_{J_n}} \left\{ \sup_{\rho \in \Gamma^k} \prod_{i=1}^n g(X_i^{(a)}, X_i^{(b)}, X_{i-1}^{(a)}, X_{i-1}^{(b)}, \rho) \right. \\ & \leq \prod_{i=1}^n g(X_i^{(a)}, X_i^{(b)}, X_{i-1}^{(a)}, X_{i-1}^{(b)}, \rho_{J_n}) \left. \right\} \\ & \subseteq \bigcup_{k=1}^{l_{J_n}} \left\{ \prod_{i=1}^n g(X_i^{(a)}, X_i^{(b)}, X_{i-1}^{(a)}, X_{i-1}^{(b)}, \Gamma_k) \right. \\ & \leq \prod_{i=1}^n g(X_i^{(a)}, X_i^{(b)}, X_{i-1}^{(a)}, X_{i-1}^{(b)}, \rho_{J_n}) \left. \right\}. \end{aligned}$$

Next we bound the probability of this latter set and called it π .

$$\pi \leq \sum_{k=1}^{l_{J_n}} P \left[\prod_{i=1}^n g(X_i^{(a)}, X_i^{(b)}, X_{i-1}^{(a)}, X_{i-1}^{(b)}, \Gamma_k) \right]$$

$$\begin{aligned} & \leq \prod_{i=1}^n g(X_i^{(a)}, X_i^{(b)}, X_{i-1}^{(a)}, X_{i-1}^{(b)}, \rho_{J_n}) \\ & = \sum_{k=1}^{l_{J_n}} P \left[\exp \sum_{i=1}^n \left\{ t_k \log \frac{g(X_i^{(a)}, X_i^{(b)}, X_{i-1}^{(a)}, X_{i-1}^{(b)}, \Gamma_k)}{g(X_i^{(a)}, X_i^{(b)}, X_{i-1}^{(a)}, X_{i-1}^{(b)}, \rho_{J_n})} \right\} \geq 1 \right] \\ & \leq \sum_{k=1}^{l_{J_n}} \mathbb{E}_{\rho_{0|\Theta_{J_n}}} \left[\exp \left\{ t_k \log \frac{g(X_t^{(a)}, X_t^{(b)}, X_{t-1}^{(a)}, X_{t-1}^{(b)}, \Gamma_k)}{g(X_t^{(a)}, X_t^{(b)}, X_{t-1}^{(a)}, X_{t-1}^{(b)}, \rho_{J_n})} \right\} \right]^n \end{aligned}$$

for any nonnegative arbitrary t_1, \dots, t_k and conditionally to $X_{i-1}^{(a)}$ and $X_{i-1}^{(b)}$, the laws of the real r.v. $g(X_i^{(a)}, X_i^{(b)}, X_{i-1}^{(a)}, X_{i-1}^{(b)}, \Gamma_k)$ and $g(X_i^{(a)}, X_i^{(b)}, X_{i-1}^{(a)}, X_{i-1}^{(b)}, \rho_{J_n})$ are images of g by the translations of the laws ε_i which are i.i.d. Hence, we get

$$\pi \leq l_{J_n} (\varphi_{J_n})^n.$$

Finally, result (A4) is deduced by condition (ii) of Theorem 3.1 and by the Borel–Cantelli lemma.

Appendix 4. Proof of consistency result in Theorem 3.2

Without loss of generality, we assume that p_j^a and p_j^b are all zeros. For non-zero cases, the same consistency results can be obtained. We check the condition **Con1**. We replace J_n by J in the remaining of this section for notational simplicity, and let all summation be from 1 to J . Using the definition of the entropy, we have

$$\begin{aligned} & H(\rho_{0|\Theta_J}, \rho_{0|\Theta_J}) - H(\rho_{0|\Theta_J}, \rho_{\Theta_J}) \\ & = H(\kappa_{0|\Theta_J}, \kappa_{0|\Theta_J}) - H(\kappa_{0|\Theta_J}, \kappa_{\Theta_J}) \\ & = -\frac{1}{2} \log |\Sigma_u| + \frac{1}{2} \log |\Sigma_{u,J}| + \mathbb{E} \left(-\frac{1}{2} x^\top \Sigma_u^{-1} x + \frac{1}{2} x_J^\top \Sigma_{u,J}^{-1} x_J \right), \end{aligned}$$

where

$$x = \begin{bmatrix} d_{t,1}^a - \sum_i \left(\sum_j q_j c_j^{aa} - c_1^{aa} \right) \frac{w_{i+m} - w_i}{m} d_{t-1,i}^a \\ - \sum_i \left(\sum_j q_j c_j^{ab} - c_1^{ab} \right) \frac{w_{i+m} - w_i}{m} d_{t-1,i}^b \\ \vdots \\ d_{t,J}^a - \sum_i \left(\sum_j q_j c_j^{aa} - c_J^{aa} \right) \frac{w_{i+m} - w_i}{m} d_{t-1,i}^a \\ - \sum_i \left(\sum_j q_j c_j^{ab} - c_J^{ab} \right) \frac{w_{i+m} - w_i}{m} d_{t-1,i}^b \\ \vdots \\ d_{t,1}^b - \sum_i \left(\sum_j q_j c_j^{ba} - c_1^{ba} \right) \frac{w_{i+m} - w_i}{m} d_{t-1,i}^a \\ - \sum_i \left(\sum_j q_j c_j^{bb} - c_1^{bb} \right) \frac{w_{i+m} - w_i}{m} d_{t-1,i}^b \\ \vdots \\ d_{t,J}^b - \sum_i \left(\sum_j q_j c_j^{ba} - c_J^{ba} \right) \frac{w_{i+m} - w_i}{m} d_{t-1,i}^a \\ - \sum_i \left(\sum_j q_j c_j^{bb} - c_J^{bb} \right) \frac{w_{i+m} - w_i}{m} d_{t-1,i}^b \end{bmatrix},$$

$$x_J = \begin{bmatrix} d_{t,1}^a - \sum_i \left(\sum_j q_j c_{j,J}^{aa} - c_{1,J}^{aa} \right) \frac{w_{i+m} - w_i}{m} d_{t-1,i}^a \\ - \sum_i \left(\sum_j q_j c_{j,J}^{ab} - c_{1,J}^{ab} \right) \frac{w_{i+m} - w_i}{m} d_{t-1,i}^b \\ \vdots \\ d_{t,J}^a - \sum_i \left(\sum_j q_j c_{j,J}^{aa} - c_{J,J}^{aa} \right) \frac{w_{i+m} - w_i}{m} d_{t-1,i}^a \\ - \sum_i \left(\sum_j q_j c_{j,J}^{ab} - c_{J,J}^{ab} \right) \frac{w_{i+m} - w_i}{m} d_{t-1,i}^b \\ \vdots \\ d_{t,1}^b - \sum_i \left(\sum_j q_j c_{j,J}^{ba} - c_{1,J}^{ba} \right) \frac{w_{i+m} - w_i}{m} d_{t-1,i}^a \\ - \sum_i \left(\sum_j q_j c_{j,J}^{bb} - c_{1,J}^{bb} \right) \frac{w_{i+m} - w_i}{m} d_{t-1,i}^b \\ \vdots \\ d_{t,J}^b - \sum_i \left(\sum_j q_j c_{j,J}^{ba} - c_{J,J}^{ba} \right) \frac{w_{i+m} - w_i}{m} d_{t-1,i}^a \\ - \sum_i \left(\sum_j q_j c_{j,J}^{bb} - c_{J,J}^{bb} \right) \frac{w_{i+m} - w_i}{m} d_{t-1,i}^b \end{bmatrix}.$$

Here $\Sigma_u, c_j^{aa}, c_j^{ab}, c_j^{ba}$, and c_j^{bb} denote the covariance matrix and B -spline coefficients for the kernel $\kappa_{0|\Theta_J}$; and $\Sigma_{u,J}, c_{j,J}^{aa}, c_{j,J}^{ab}, c_{j,J}^{ba}$, and $c_{j,J}^{bb}$ denote the covariance matrix and B -spline coefficients for the kernel κ_J . κ_J is the set of kernel functions for ρ_J with $\rho_J \in \Theta_J$; and $\kappa_{0|\Theta_J}$ is the projection of the set of true kernel functions κ_0 on Θ_J . Assuming $\Sigma_u = \Sigma_{u,J}$, we have

$$\begin{aligned} & H(\rho_{0|\Theta_J}, \rho_{0|\Theta_J}) - H(\rho_{0|\Theta_J}, \rho_{\Theta_J}) \\ &= \mathbb{E} \left(-\frac{1}{2} x^\top \Sigma_u^{-1} x + \frac{1}{2} x_J^\top \Sigma_u^{-1} x_J \right) \\ &= \frac{1}{2} \sum_{r,s} (\Sigma_u^{-1})_{r,s} \mathbb{E} \left\{ (x_J)_r (x_J)_s - (x)_r (x)_s \right\}, \end{aligned}$$

where $(\Sigma_u^{-1})_{r,s}$ is the r th row, s th column of Σ_u^{-1} , $(x)_r$ is the r th element of x_J , and $(x)_r$ is the r th element of x .

Since the only difference between $(x_J)_r (x_J)_s$ and $(x)_r (x)_s$ are the different B -spline coefficients, we can group the individual terms of the expansion of $(x_J)_r (x_J)_s$ and the expansion $(x)_r (x)_s$ together. After canceling out the common terms not containing the B -spline coefficients, each of the grouped terms will contain a product of some common terms and the subtraction between the B -spline coefficients (of the same index) of the two kernels or the subtraction between the product of B -spline coefficients of one kernel and that of the other kernel (of the same combination of indices). Hence, if $H(\kappa_{0|\Theta_J}, \kappa_{\Theta_J}) \rightarrow H(\kappa_{0|\Theta_J}, \kappa_{0|\Theta_J})$ as $n, J \rightarrow \infty$, we have $c_{j,J}^{aa} \rightarrow c_j^{aa}$, $c_{j,J}^{ab} \rightarrow c_j^{ab}$, $c_{j,J}^{ba} \rightarrow c_j^{ba}$, $c_{j,J}^{bb} \rightarrow c_j^{bb}$ and consequently $\rho_J \rightarrow \rho_{0|\Theta_J}$.

For the condition **Con2** and (i) of Theorem 3.1, we follow similar arguments as in Mourid and Bensmain (2006). To verify Theorem 3.1 (ii), we define

$$\varphi(t) = \mathbb{E}_{\kappa_{0|\Theta_J}} \left\{ \exp \left(t \log \frac{g(X_t^{(a)}, X_t^{(b)}, X_{t-1}^{(a)}, X_{t-1}^{(b)}, \Gamma_k)}{g(X_t^{(a)}, X_t^{(b)}, X_{t-1}^{(a)}, X_{t-1}^{(b)}, \kappa_J)} \right) \right\},$$

where $g(X_t^{(a)}, X_t^{(b)}, X_{t-1}^{(a)}, X_{t-1}^{(b)}, \Gamma_k) = \sup_{\psi \in \Gamma_k} g(X_t^{(a)}, X_t^{(b)}, X_{t-1}^{(a)}, X_{t-1}^{(b)}, \psi)$. Furthermore, we have $\varphi(0) = 1$ and $\varphi' = \mathbb{E}_{\kappa_{0|\Theta_J}} \log(g(X_t^{(a)}, X_t^{(b)}, X_{t-1}^{(a)}, X_{t-1}^{(b)}, \Gamma_k)/g(X_t^{(a)}, X_t^{(b)}, X_{t-1}^{(a)}, X_{t-1}^{(b)}, \kappa_J))$. For a fixed $\kappa \in \Gamma_k$, we have

$$\begin{aligned} A &= \mathbb{E}_{\kappa_{0|\Theta_J}} \log g(X_t^{(a)}, X_t^{(b)}, X_{t-1}^{(a)}, X_{t-1}^{(b)}, \Gamma_k) \\ &\quad - \mathbb{E} \log g(X_t^{(a)}, X_t^{(b)}, X_{t-1}^{(a)}, X_{t-1}^{(b)}, \kappa) \\ &= \mathbb{E}_{\kappa_{0|\Theta_J}} \sup_{\psi \in \Gamma_k} \left\{ \log g(X_t^{(a)}, X_t^{(b)}, X_{t-1}^{(a)}, X_{t-1}^{(b)}, \psi) \right. \\ &\quad \left. - \log g(X_t^{(a)}, X_t^{(b)}, X_{t-1}^{(a)}, X_{t-1}^{(b)}, \kappa) \right\} \\ &= \mathbb{E}_{\kappa_{0|\Theta_J}} \sup_{\psi \in \Gamma_k} \left\{ -\frac{1}{2} \log |\Sigma_{u,\psi}| + \frac{1}{2} \log |\Sigma_{u,\kappa}| \right. \\ &\quad \left. - \frac{1}{2} x_\psi^\top \Sigma_{u,\psi}^{-1} x_\psi + \frac{1}{2} x_\kappa^\top \Sigma_{u,\kappa}^{-1} x_\kappa \right\}, \end{aligned}$$

where x_ψ and x_κ have the same form as x_J , with J replaced by ψ and κ , respectively. $\Sigma_{u,\psi}, c_{j,\psi}^{aa}, c_{j,\psi}^{ab}, c_{j,\psi}^{ba}$, and $c_{j,\psi}^{bb}$ denote the covariance matrix and B -spline coefficients for the kernel ψ , while $\Sigma_{u,\kappa}, c_{j,\kappa}^{aa}, c_{j,\kappa}^{ab}, c_{j,\kappa}^{ba}$, and $c_{j,\kappa}^{bb}$ denote that for the kernel κ .

Assuming $\Sigma_{u,\psi} = \Sigma_{u,\kappa} = \Sigma_u$, we have

$$A = \mathbb{E}_{\kappa_{0|\Theta_J}} \sup_{\psi \in \Gamma_k} \left\{ \frac{1}{2} \sum_{r,s} (\Sigma_u^{-1})_{r,s} \left((x_\psi)_r (x_\psi)_s - (x_\kappa)_r (x_\kappa)_s \right) \right\},$$

where $(\Sigma_u^{-1})_{r,s}$ is the r th row, s th column of Σ_u^{-1} , $(x_\psi)_r$ is the r th element of x_ψ , and $(x_\kappa)_r$ is the r th element of x_κ .

We follow the similar conditions and arguments in Mourid and Bensmain (2006) and obtain $A \leq C_1/J^{\eta/2}$, where C_1 is a constant. In addition, for $\delta > 0$,

$$\varphi'(0) = H(\kappa_{0|\Theta_J}, \kappa) - H(\kappa_{0|\Theta_J}, \kappa_J) + A \leq C_2 J^{-\eta/2} - \delta.$$

Using Taylor expansion and the results from Hwang (1980) such that $\varphi''(t) \leq C_3 J^2$, we have $\varphi(1/J^2) \leq 1 - \delta/C_4 J^2$, where C_2, C_3 , and C_4 are constants. Since $\varphi_J = \sup_k \inf_{t \geq 0} \varphi(t)$, we can deduce that for sufficiently large J , we have

$$l_J(\varphi_J)^n \leq C J^{CJ^{1+\eta}} \left(1 - \frac{\delta}{CJ^2} \right)^n,$$

which is summable if $J = \mathcal{O}(n^{1/3-\delta})$ for $\delta > 0$ (see Hwang 1980). Note that C is a constant. Finally, we can apply Theorem 3.1 to obtain the result that the ML estimator $\hat{\kappa}$ obtained on Θ_{J_n} converges to the projected true set of kernel functions $\kappa_{0|\Theta_J}$. As $n, J_n \rightarrow \infty$, $\kappa_{0|\Theta_J} \rightarrow \kappa_0$ because each $\kappa_{xy,0|\Theta_J}$ in $\kappa_{0|\Theta_J}$ is just the B -spline truncation of the corresponding true kernel $\kappa_{xy,0}$ in κ_0 on Θ_{J_n} .

Network Quantile Autoregression

Xuening Zhu¹, Weining Wang^{2,4}, Hangsheng Wang³, and Wolfgang Karl Härdle^{2,5}

¹ *School of Data Science, Fudan University, Shanghai, China;*

² *Humboldt-Universität zu Berlin, C.A.S.E.-Center of Applied Statistics and Economics, Unter den Linden 6, 10099 Berlin, Germany;*


³ *Guanghua School of Management, Peking University, Beijing, China;*

⁴ *Department of Economics, City University London;*

⁵ *Research fellow in Sim Kee Boon Institute for Financial Economics, Singapore Management University, 90 Stamford Road, 6th Level, School of Economics, Singapore 178903.*

Abstract

The complex tail dependency structure in a dynamic network with a large number of nodes is an important object to study. Here we propose a network quantile autoregression model (NQAR), which characterizes the dynamic quantile behavior. Our NQAR model consists of a system of equations, of which we relate a response to its connected nodes and node specific characteristics in a quantile autoregression process. We show the estimation of the NQAR model and the asymptotic properties with assumptions on the network structure. For this propose we develop a network Bahadur representation that gives us direct insight into the parameter asymptotics. Moreover, innovative tail-event driven impulse functions are defined. Finally, we demonstrate the usage of our model by investigating the financial contagions in the Chinese stock market accounting for shared ownership of companies. We find higher network dependency when the market is exposed to a higher volatility level.

 www.quantlet.de

KEY WORDS: Social Network; Quantile Regression; Autoregression; Systemic Risk; Financial Contagion; Shared Ownership.

*Xuening Zhu¹ is supported by National Natural Science Foundation of China (NSFC, 11131002, 11271031, 11525101, 11690015, 71532001), NIDA, NIH grants P50 DA039838, and a NSF grant DMS 1512422. Weining Wang^{2,4} and Wolfgang Karl Härdle^{2,5} are supported by German Research Foundation (DFG) on “quantile method for complex financial system” and IRTG 1792. Hangsheng Wang³ is supported by National Natural Science Foundation of China (NSFC, 11131002, 11271031, 11525101, 71532001), the Business Intelligence Research Center at Peking University, the Center for Statistical Science at Peking University.

1. INTRODUCTION

Quantile regression is an increasingly popular tool for modern econometric analysis. Instead of studying the conditional mean function of the response variable, quantile regression is concerned with estimating the conditional quantile function. It has been applied to a wide range of econometric applications, such as labor economics (Koenker and Hallock, 2001; Fitzenberger et al., 2013) and financial risk management (Gaglianone et al., 2011; Härdle et al., 2016). Particularly, the linear quantile model has been studied by a seminal work by Koenker and Bassett (1978), and the asymptotic theory has been developed by Portnoy (1991, 1997). See Koenker (2005) for a comprehensive summary of the methods and applications.

Following the development of quantile methods in the existing literature, the quantile regression in time series is of particular interest. An early stream of researches such as Hasan and Koenker (1997) deal with linear quantile autoregression models, which focus on independent identically distributed (*iid*) innovations in a relatively restrictive location shift model. In another approach, Engle and Manganelli (2004) propose a set of autoregressive forms (the CaViaR model) for value-at-risk (VaR), which models directly the dynamics of the conditional quantiles. The framework is easy to apply but is quite difficult to directly infer the underlying process. As an alternative, Koenker and Xiao (2006) consider a quantile autoregressive method to model the conditional quantile function, which allows to study the asymptotic properties of the underlying process and does not assume an *iid* underlying process. This provides us with an interesting framework to understand the risk propagation within a complex financial system. We consider therefore a parametric approach involving a system of dynamic quantile autoregression equations. Thus our methodology characterizes a dynamic tail dependency graph, which facilitates tail event driven forecasting and impulse analysis in a complex financial system. This is in particular important in a financial network and complements to the literature on systemic risk; see for example Billio et al. (2012), Diebold and Yilmaz (2014).

The rapid development of modern data technology has allowed us to have access to large amount of data with possible network structures. On one hand, this poses serious challenges to the analysis of dynamic tail behavior especially within a network composed of a large number of nodes. On the other hand, the data structure bears the opportunity for naturally embedded network information. We take this opportunity of employing network structures, and propose a tail event driven network quantile autoregression model (NQAR), which allows us to make inference based on the underlying processes and to

estimate the network dependencies.

In the existing literature, great efforts have been taken to incorporate network information into an econometric framework. For instance, Sewell and Chen (2015) incorporate network information to study the dynamic social behavior of students in a Dutch class by a latent space model. Community detection and extraction methods are studied by Zhao et al. (2011), Amini et al. (2013), and Sewell et al. (2017) using a block structure. Comparably, our proposed framework is related to the recent autoregressive models in large-scale social networks. Based on the cited literature, it is assumed that the behavioural patterns of network users are related to their connected friends (Zhang and Chen, 2013; Zhu et al., 2017). Estimation and computation issues for this situation are intensively discussed (Zhou et al., 2017; Zhu et al., 2018).

A recent paper by Zhu et al. (2017) on network autoregression provide a modelling framework at the mean level. In this work, we extend this to a network quantile autoregression (NQAR) model in order to study conditional quantiles in complex financial networks. Consider a network of firms, connected by their shareholder relationships. Specifically, a nodal connection between two firms can be established if they share major common shareholders. In this context, it makes sense to assume that the conditional quantile function of the response variable (e.g., volatility of stock returns for the firms) is related to underlying exogenous factors. These may include nodal specific variables (e.g., firm specific variables), the lagged response of the same node (e.g., volatility of the same stock in the previous time point), and the lagged responses of other connected nodes. To estimate the parameter, a minimum contrast method is introduced, which is applied to a large-scale network. The corresponding asymptotic properties are established, where the conditions on network structures are given and discussed. Moreover, the stationarity of the NQAR model is investigated, and an impulse analysis under the NQAR model is discussed. Empirically, we discover strong asymmetric network effects of shocks at different quantile levels of stock volatilities in the Chinese financial market. Namely, the network dependence among the absolute returns becomes stronger at the tail level, while at normal times it is not significant.

Finally, our paper is closely related to the recent emerging literature on modeling tail dependence in a complex financial system. Examples include the quantile LASSO framework discussed by Hautsch et al. (2014) and Härdle et al. (2016), where the network relationship is estimated among the financial institutions by imposing an L_1 -penalty. Their estimation framework considers a more flexible network formation at the cost of s-

lower convergence rate, as it is in nature a nonparametric estimation. Furthermore, there is also a literature on the tail dependence in multi-dimensional dynamic settings. For example, Cappiello et al. (2014) develop an econometric measure for the comovement of quantiles. In addition, White et al. (2015) provide a very innovative vector autoregressive model for the dynamics of quantiles. Chavleishvili and Manganelli (2016) propose a framework related to the CaViaR process to identify the structural quantile shocks. Comparably, our approach is different mainly in the following three aspects. First, the proposed NQAR model embeds the observed network structure, which provides a parametric estimation framework. Second, it admits trackable quantile dynamics, which facilitates to conduct stationary and impulse response analysis. Third, the model allows for modeling a large number of nodes (a high dimensional setup), and controlling for the observed nodal heterogeneity.

Lastly, we summarize our contribution as follows. Firstly, we provide a novel network quantile model that characterizes the dynamic quantile behavior, which incorporates valuable network information from data. Secondly, we give new definitions of tail-event driven impulse functions under this innovative modeling framework. Thirdly, the asymptotic theories are derived for both the underlying process and estimated parameters. The model stationarity is discussed with insights on its relationship with the given network information. Moreover, detailed conditions on the network structures are derived to ensure the consistency and asymptotic normality of the estimator.

The rest of the paper is organized as follows. Section 2 introduces the network quantile autoregression model and its stationarity properties. Section 3 proposes a novel impulse analysis framework for the network quantile autoregression model. The parameter estimation method is given in Section 4, where the asymptotic properties are presented. An empirical analysis for stocks in Chinese financial markets are conducted in Section 5. Lastly, a conclusion is discussed in Section 6. Extensive numerical studies and all technical details are delegated to the supplementary material.

2. NETWORK QUANTILE AUTOREGRESSION

2.1. Model and Notations

Let U_{it} ($1 \leq i \leq N$, $1 \leq t \leq T$) be *iid* random variables following the standard uniform distribution on the set of $[0, 1]$. Assume that a q -dimensional random nodal covariate vector $Z_i = (Z_{i1}, \dots, Z_{iq})^\top \in \mathbb{R}^q$ is collected for each node. To record the

network relationship, we define $A = (a_{ij}) \in \mathbb{R}^{N \times N}$ as the adjacency matrix, where $a_{ij} = 1$ if there is an edge from i to j , otherwise $a_{ij} = 0$.

Following the standard conventions, the nodes are assumed to be not self-related (i.e., $a_{ii} = 0$). Motivated by the univariate autoregression quantile model (Koenker and Xiao, 2006), we consider the network quantile autoregression model as

$$Y_{it} = \beta_0(U_{it}) + \sum_{l=1}^q Z_{il}\gamma_l(U_{it}) + \beta_1(U_{it})n_i^{-1} \sum_{j=1}^N a_{ij}Y_{j(t-1)} + \beta_2(U_{it})Y_{i(t-1)}, \quad (2.1)$$

where β_j s ($0 \leq j \leq 2$) and γ_l s ($1 \leq l \leq q$) are unknown coefficient functions from $(0, 1)$ to \mathbb{R}^1 , and $n_i = \sum_{j \neq i} a_{ij}$ is the out-degree for the i th node.

Importantly, the NQAR model (2.1) induces a convenient form of the conditional quantile function of Y_{it} . Denote $Q_Y(\tau|X)$ as the τ th conditional quantile of Y seen as a function of X . By assuming that the right side of (2.1) is monotonically increasing in U_{it} , we can write the conditional quantile function of Y_{it} given (Z_i, \mathbb{Y}_{t-1}) as:

$$Q_{Y_{it}}(\tau|Z_i, \mathbb{Y}_{t-1}) = \beta_0(\tau) + \sum_{l=1}^q Z_{il}\gamma_l(\tau) + \beta_1(\tau)n_i^{-1} \sum_{j=1}^N a_{ij}Y_{j(t-1)} + \beta_2(\tau)Y_{i(t-1)}. \quad (2.2)$$

In (2.2), $\beta_0(\tau) + \sum_{l=1}^q Z_{il}\gamma_l(\tau)$ reflects the nodal impact invariant over t , where $\beta_0(\tau)$ is referred to as the *baseline function*. The covariates Z_{il} refer to node-specific variables, like, size, leverage ratio, which are invariant in time. It is assumed that the nodal covariates Z_i s are independent of the U_{it} s. Next, the second term $n_i^{-1} \sum_{j=1}^N a_{ij}Y_{j(t-1)}$ characterizes the network impact from the connected nodes (e.g., firms with common shareholders) (Zhu et al., 2017). The corresponding coefficient function $\beta_1(\tau)$ is then referred to as the *network function*. Lastly, $Y_{i(t-1)}$ captures the impact from the response of the same node in the previous time point. Accordingly, the coefficient function $\beta_2(\tau)$ is then referred to as the *momentum function*. The model (2.2) is related to the autoregression models in spatial econometrics literature, e.g. Lee (2004); Lee and Yu (2009). Although they share the similarity in the construction of the adjacency matrix A , the modelling interests are different. Specifically, the spatial models mainly characterize the simultaneous spatial effect across spatial locations, while our approach mainly focus on modelling the dynamic patterns of the responses. To better understand the NQAR model (2.1), we have the following remarks.

Remark 1. (Monotonicity) Monotonicity is a frequently discussed issue for the quantile

autoregression model. A specific example for the monotonicity of (2.1) to hold is that $\gamma_l(\cdot)$ s and $\beta_l(\cdot)$ s are all monotone increasing functions, and Z_{it} s, Y_{it} s are positive random variables. In other cases, certain data transformation techniques can be conducted to ensure this assumption; see Koenker and Xiao (2006) and Fan and Fan (2010) for more detailed discussions.

Remark 2. (Comparison with NAR Model) One certainly notes (2.2) to be an extension of the network vector autoregression (NAR) model proposed by Zhu et al. (2017). First of all, the NQAR model has a varying coefficient structure which requires to significantly different tools for deriving theoretical properties. Emphatically, the conditional quantile function in the NAR case is

$$Q_{Y_{it}}(\tau|Z_i, \mathbb{Y}_{t-1}) = \beta_0 + \sum_{l=1}^q Z_{il}\gamma_l + \beta_1 n_i^{-1} \sum_{j=1}^N a_{ij} Y_{j(t-1)} + \beta_2 Y_{i(t-1)} + c_\tau,$$

where c_τ is the quantile of the error distribution, and all the parameters $\beta_1, \beta_2, \gamma_l$ are not related to τ . In contrast, the NQAR model allows coefficient functions to vary over τ . This makes not only the location of the conditional density of Y_{it} be determined by τ , but also the shape of $Q_{Y_{it}}(\tau|Z_i, \mathbb{Y}_{t-1})$ be τ -related. In practice, this model formulation is of particular interest for financial risk management. Specifically, we discuss the following two scenarios in which NQAR is more powerful than the mean case (i.e., NAR model).

SCENARIO 1. (TAIL BEHAVIOR) The NQAR model captures asymmetric dependency between the responses at different quantile levels, especially at tail levels. For instance, to measure the conditional VaR of a firm, one can adopt the stock volatility as Y_{it} s for the i th firm and at $\tau = 0.95$. In this case, an asymmetric pattern indicates whether the financial institutions tend to have closer connections in the upper tail (e.g. when the market exhibits high turbulences) than other levels.

SCENARIO 2. (ROBUST ESTIMATION) General vector autoregression models are usually sensitive to outliers, which leads to a serious distortion of the estimation (Abello et al., 2013; Li et al., 2015). Consequently, compared to NAR, NQAR is more robust to outliers since it is established on the quantile framework. Specifically, the robust median estimation can be readily obtained by setting $\tau = 0.5$.

Remark 3. (Heteroskedasticity) Heteroskedasticity is a pervasive phenomenon in complex financial systems. The QNAR model could include a vector autoregression model with heteroskedasticity as a special case. We take for example the classical location shift

form as mentioned by Koenker and Hallock (2001) as

$$Y_{it} = b_0 + b_1 \bar{Y}_{i(t-1)} + b_2 Y_{i(t-1)} + \sigma_i \varepsilon_{it}, \quad (2.3)$$

where b_0, b_1, b_2 are constants, $\bar{Y}_{i(t-1)} = n_i^{-1} \sum_j a_{ij} Y_{j(t-1)}$, $\sigma_i = b_0 + b_1 \bar{Y}_{i(t-1)} + b_2 Y_{i(t-1)}$, and ε_{it} s are *iid* random variables with distribution function $F(\cdot)$. One could note that Model (2.3) involves the heteroscedasticity in the innovation term. It can be rewritten into,

$$Y_{it} = b_0(1 + \varepsilon_{it}) + b_1(1 + \varepsilon_{it})\bar{Y}_{i(t-1)} + b_2(1 + \varepsilon_{it})Y_{i(t-1)},$$

Specifically, it is a special case of the NQAR model (2.1) by specifying $\beta_0(U_{it}) = b_0(1 + \varepsilon_{it})$, $\beta_1(U_{it}) = b_1(1 + \varepsilon_{it})$, and $\beta_2(U_{it}) = b_2(1 + \varepsilon_{it})$, where $\varepsilon_{it} = F^{-1}(U_{it})$.

Remark 4. (Adjacency Matrix) It should be noted that the adjacency matrix A in (2.1) is allowed to take flexible forms according to specific application scenarios. For example, in social network analysis, the adjacency matrix is defined by the natural following-follower relationship (Chen et al., 2013; Zhou et al., 2017; Zhu et al., 2017). Specifically, $a_{ij} = 1$ if the i th node follows the j th node in the network; otherwise $a_{ij} = 0$. In economic and financial applications, one can take several strategies. For instance, Acemoglu et al. (2012) provide a characterization of intersectorial input-output linkage embedded in a network relationship. Alternatively, the network structures of financial institutions are usually constructed according to their financial fundamentals. Specifically, the industrial background, financial statement, shareholder information are commonly employed for network construction (Zou et al., 2017; Antón et al., 2018; Chen et al., 2018). In spatial econometrics, the adjacency matrix can be related to spatial distances between locations (or even economic distance such as a measure of trade flows, e.g., Novy (2013)), where the weight (i.e., a_{ij}) are usually assumed to be monotone decreasing with distance increasing (Cressie and Wikle, 2015; Lee, 2004). In addition, one could take a further flexible approach to model the adjacency matrix A at the first step according to different statistical models. Particularly, the random graph model (Hoff et al., 2012; Herz, 2015) and statistical tests (Granger et al., 2000) can be applied.

2.2. Vector Formulation of NQAR

Next, we organize the NQAR model in (2.1) into vector forms to facilitate further discussions. Define $\mathbb{Y}_t = (Y_{1t}, \dots, Y_{Nt})^\top \in \mathbb{R}^N$. Let $\mathbf{B}_{0t} = (\beta_0(U_{it}) + \sum_l Z_{il} \gamma_l(U_{it}), 1 \leq i \leq N)^\top \in \mathbb{R}^N$, $\mathbf{B}_{1t} = \text{diag}\{\beta_1(U_{it}), 1 \leq i \leq N\} \in \mathbb{R}^{N \times N}$, $\mathbf{B}_{2t} = \text{diag}\{\beta_2(U_{it}), 1 \leq i \leq$

$N\}^\top \in \mathbb{R}^{N \times N}$. One can easily verify that $\Gamma = \mathbf{E}(\mathbf{B}_{0t}) = c_0 \mathbf{1}_N \in \mathbb{R}^N$, where $c_0 = b_0 + c_Z$, $b_0 = \int_0^1 \beta_0(u) du$ and $c_Z = \mathbf{E}(Z_1)^\top r$ with $r = (\int_0^1 \gamma_l(u) du, 1 \leq l \leq q)^\top \in \mathbb{R}^q$. Then the NQAR model (2.1) can be re-written in a vector form as

$$\mathbb{Y}_t = \Gamma + G_t \mathbb{Y}_{t-1} + V_t, \quad (2.4)$$

where $G_t = \mathbf{B}_{1t}W + \mathbf{B}_{2t} \in \mathbb{R}^{N \times N}$, $W = (w_{ij}) = (n_i^{-1} a_{ij}) \in \mathbb{R}^{N \times N}$ is the row-normalized adjacency matrix, and $V_t = \mathbf{B}_{0t} - \Gamma \in \mathbb{R}^N$ is *iid* over t with mean $\mathbf{0}$ and covariance $\Sigma_V = \sigma_V^2 I_N \in \mathbb{R}^{N \times N}$ with $\sigma_V^2 = \sigma_{b_0}^2 + \mathbf{E}\{\gamma^\top(U_{1t})\Sigma_z\gamma(U_{1t})\}$, $\sigma_{b_0}^2 = \int_0^1 \beta_0^2(u) du - \{\int_0^1 \beta_0(u) du\}^2$, and $\Sigma_z = \text{Cov}(Z_1) \in \mathbb{R}^{q \times q}$.

Note that (2.4) is written in the form of a vector autoregression model (Lütkepohl, 2005) with a stochastic coefficient matrix G_t depending on t . It is not hard to see that G_t is linear in the adjacency matrix W . This form borrows the strength of network structure information (i.e., W), and greatly reduces the dimensionality of estimated parameters. For convenience, we discuss the model stationarity based on the vector form (2.4).

2.3. Covariance Stationarity

Given the NQAR model (2.4), it is then of great interest to check the stationarity of \mathbb{Y}_t . A process $\{\mathbb{Y}_t\}_{-\infty}^{+\infty}$ is covariance stationary if (a) $\mathbf{E}(\mathbb{Y}_t) = \mu_Y$ for a constant vector $\mu_Y \in \mathbb{R}^N$; (b) $\text{Cov}(\mathbb{Y}_t, \mathbb{Y}_{t-h}) = \mathbf{E}\{(\mathbb{Y}_t - \mu_Y)(\mathbb{Y}_{t-h} - \mu_Y)^\top\} = \Sigma(h)$ with $\Sigma(h) \in \mathbb{R}^{N \times N}$ only related to h . For convenience, let $b_1 = \mathbf{E}\{\beta_1(U_{it})\}$, $b_2 = \mathbf{E}\{\beta_2(U_{it})\}$, $\tilde{b}_1 = \{\mathbf{E}\beta_1^2(U_{it})\}^{1/2}$, $\tilde{b}_2 = \{\mathbf{E}\beta_2^2(U_{it})\}^{1/2}$, $G = \mathbf{E}(G_t) = b_1W + b_2I$, and $G^* = \mathbf{E}(G_t \otimes G_t)$. Then we have the following theorem.

THEOREM 1. *Assume $\tilde{b}_1 + \tilde{b}_2 < 1$ and $\mathbf{E}|V_{it}| < C$ for some positive constant C . Then the following conclusions hold.*

(a) *There exists a unique covariance stationary solution to the NQAR model (2.4) with finite first moment as*

$$\mathbb{Y}_t = \sum_{l=0}^{\infty} \Pi_l \Gamma + \sum_{l=0}^{\infty} \Pi_l V_{t-l}, \quad (2.5)$$

where $\Pi_l = \prod_{k=1}^l G_{t-k+1}$ for $l \geq 1$ and $\Pi_0 = I$.

(b) *Denote $\Sigma_Y = \Sigma(0) = \text{Cov}(\mathbb{Y}_t)$. The stationary mean is $\mu_Y = c_1^{-1} c_0 \mathbf{1}_N$ and*

$$\text{vec}(\Sigma_Y) = (M_1 - c_1^{-2} c_0^2) \mathbf{1}_{N^2} + 2c_1^{-1} c_0 (I - G^*)^{-1} \text{vec}(\Sigma_{bv}) + (I - G^*)^{-1} \text{vec}(\Sigma_V), \quad (2.6)$$

where $c_1 = (1 - b_1 - b_2)^{-1}$, $M_1 = c_1^{-1} c_0^2 (1 + b_1 + b_2) (I - G^*)^{-1}$, $\Sigma_{bv} = \sigma_{bv} I_N$, and

$\sigma_{bv} = \mathbb{E}[\{\beta_1(U_{it}) + \beta_2(U_{it})\}V_{it}]$. Moreover, we have $\Sigma(h) = G^h \Sigma_Y$ for any integer $h > 0$, and $\Sigma(h) = \Sigma_Y (G^\top)^{-h}$ for $h < 0$.

The proof of Theorem 1 is given in the supplementary material (Appendix A.1). It provides the unique covariance stationary solution of (2.5).

Remark 5. It is straightforward to verify $\tilde{b}_1 = (b_1^2 + \sigma_{b_1}^2)^{1/2}$, where $\sigma_{b_1}^2 = \mathbf{Var}\{\beta_1(U_{it})\}$. Similarly one can define $\sigma_{b_2}^2 = \mathbf{Var}\{\beta_2(U_{it})\}$ and $\tilde{b}_2 = (b_2^2 + \sigma_{b_2}^2)^{1/2}$. Therefore the stationarity assumption in Theorem 1 essentially sets constraints on the expectation and variance of the network and momentum functions (i.e., $\beta_1(\cdot)$ and $\beta_2(\cdot)$). It is noteworthy that the assumption does not require $|\beta_1(\tau)| + |\beta_2(\tau)|$ to be strictly less than one for all $\tau \in (0, 1)$. Instead, it imposes an upper bound in L_2 -norm, which allows for some ‘‘explosive’’ cases at a specific quantile (i.e., $|\beta_1(\tau)| + |\beta_2(\tau)| > 1$ for some τ). Particularly, if the network and momentum functions are constants, i.e., $\beta_1(\tau) = b'_1$ and $\beta_2(\tau) = b'_2$ (for some constants b'_1 and b'_2), the stationarity assumption reduces to $|b'_1| + |b'_2| < 1$, which corroborates to the stationary condition in the mean case (Zhu et al., 2017).

Remark 6. Let us look at the stationary mean μ_Y and covariance Σ_Y . First, note that $\mu_Y = c_1^{-1} c_0 \mathbf{1}_N$, thus the stationary mean is the same for all the nodes, and unrelated to the network structure. By contrast, the analytical form for the covariance Σ_Y is more complicated. To better understand how Σ_Y is affected by the network structure, we approximate Σ_Y in the case that β_1 is small (Chen et al., 2013; Zhou et al., 2017; Zhu et al., 2017), namely, $\tilde{b}_1 = o(1)$. For convenience, define $\tilde{b}_{12} = \mathbb{E}\{\beta_1(U_{it})\beta_2(U_{it})\}$, $\tilde{b}_{01} = \mathbb{E}\{\beta_1(U_{it})V_{it}\}$, and $\tilde{b}_{02} = \mathbb{E}\{\beta_2(U_{it})V_{it}\}$. Employing the Taylor’s expansion, Σ_Y can be approximated by

$$\mathbf{Var}(Y_{it}) \approx c_{b1} c_0^2 + \frac{1}{1 - \tilde{b}_2^2} \left[2(1 - b_2)^{-2} \{ (1 - b_2) \sigma_{bv} + b_1 \tilde{b}_{02} \} c_0 + \sigma_V^2 \right], \quad (2.7)$$

$$\mathbf{Cov}(Y_{i_1 t}, Y_{i_2 t}) \approx c_{b2} c_0^2 + \frac{1}{(1 - \tilde{b}_2^2)^2} \left\{ 2(1 - b_2)^{-1} \tilde{b}_{02} c_0 + \sigma_V^2 \right\} \left\{ b_1 b_2 (w_{i_1 i_2} + w_{i_2 i_1}) \right\}, \quad (2.8)$$

where $c_{b1} = [(1 - \tilde{b}_2^2)^{-1} \{ 1 - b_2^2 + 2b_1 + 2(1 - \tilde{b}_2^2)^{-1} (1 - b_2^2) \tilde{b}_{12} \} - (1 - b_2)^{-1} (1 - b_2 + 2b_1)] (1 - b_2)^{-2}$ and $c_{b2} = (1 - b_2)^{-2} (1 - b_2^2)^{-2} (1 - b_2^2 + 2b_1 + 2b_1 b_2) - (1 - b_2)^{-3} (1 - b_2 + 2b_1)$. Detailed verifications of (2.7) and (2.8) are given in the supplementary material (Appendix A.2). It is worth noting that the variance of Y_{it} is mainly determined by the momentum function $\beta_2(\cdot)$ and the baseline function $\beta_0(\cdot)$. Particularly, a larger \tilde{b}_2 will result in higher variance of Y_{it} . Next, the covariance between $Y_{i_1 t}$ and $Y_{i_2 t}$ is not only related to $\beta_2(\cdot)$, but is also related to the network function $\beta_1(\cdot)$. Nodes have a higher correlation with each other if b_1

is larger. Moreover, note that $w_{i_1 i_2} + w_{i_2 i_1} = n_{i_1}^{-1} a_{i_1 i_2} + n_{i_2}^{-1} a_{i_2 i_1}$. Therefore, the correlation between node i_1 and i_2 is higher if (a) they connect to each other in the network (i.e., $a_{i_1 i_2} = a_{i_2 i_1} = 1$) and (b) they both have small out-degrees (i.e., small n_{i_1} and n_{i_2}).

2.4. Asymptotic Stationary Distribution

Given the established covariance stationarity, it is then natural to derive the asymptotic stationary distribution. We focus on the long run average of \mathbb{Y}_t , namely, $\bar{\mathbb{Y}}_T = T^{-1} \sum_{t=1}^T \mathbb{Y}_t$. It reflects the average performance of \mathbb{Y}_t over the whole time period T , and its asymptotic properties are going to be investigated as $T \rightarrow \infty$. In this regard, two types of asymptotics exist. The first type is fixed N asymptotics, and the second one is $N \rightarrow \infty$ asymptotics. In the following theorem, we first give the result of fixed N asymptotics.

THEOREM 2. *Assume $c_\beta < 1$ and $\mathbb{E}(|V_{it}|^4) < M$, where $c_\beta = \|\beta_1\|_4 + \|\beta_2\|_4$ with $\|\beta_j\|_4 = \mathbb{E}\{\beta_j(U_{it})^4\}^{1/4}$ ($j = 1, 2$), and M is a finite positive constant. Then the average of \mathbb{Y}_t converges in law to a normal distribution,*

$$\sqrt{T}(\bar{\mathbb{Y}}_T - \mu_Y \mathbf{1}) \xrightarrow{\mathcal{L}} \mathbb{N}(0, \Sigma_Y^*) \quad \text{as } T \rightarrow \infty, \quad (2.9)$$

where $\Sigma_Y^* = G(I - G)^{-1} \Sigma_Y + \Sigma_Y (I - G^\top)^{-1}$.

The proof of Theorem 2 is deferred to the supplementary material (Appendix A.3). Via (2.9), the asymptotic normality of $\sqrt{T}(\bar{\mathbb{Y}}_T - \mu_Y \mathbf{1})$ is provided. One could see that the corresponding asymptotic covariance is equal to the long run covariance $\Sigma_Y^* = \sum_{h=-\infty}^{\infty} \Sigma(h) = (\sigma_{ij}^*) \in \mathbb{R}^{N \times N}$.

Note that the Theorem 2 is established for a fixed N . However, one might consider to extend the result directly to the case $N \rightarrow \infty$. On general grounds, this can be difficult, since the convergence in distribution in high dimensions is not well defined. As one possible solution, we discuss the problem under the framework of Gaussian approximation theory, which is formulated by Zhang and Cheng (2014) and Zhang and Wu (2015) for time series analysis. Before we introduce this $N \rightarrow \infty$ asymptotics, we first give definition of a convenient distance between two high dimensional vectors. Specifically, the Kolmogorov distance is employed and defined as follows.

DEFINITION 1. *Let $X = (X_1, \dots, X_N)^\top \in \mathbb{R}^N$, $Y = (Y_1, \dots, Y_N)^\top \in \mathbb{R}^N$ be N -*

dimensional random vectors. The Kolmogorov distance between X and Y is defined as

$$\rho(X, Y) = \sup_{t \in \mathbb{R}} \left| P(\|X\|_\infty \leq t) - P(\|Y\|_\infty \leq t) \right|,$$

where $\|X\|_\infty$ denotes $\max_{1 \leq i \leq N} |X_i|$ for any arbitrary vector X .

The Kolmogorov distance can be seen as a distance between two distribution functions. Using it, we are able to quantify the distance between $\sqrt{T}(\bar{Y}_T - \mu_Y \mathbf{1})$ and a given Gaussian random vector. Specifically, define $\tilde{Y}_t = Y_t - \mu_Y$, then we have $\text{Cov}(\tilde{Y}_t, \tilde{Y}_{t-h}) = \Sigma(h)$. Accordingly, let $Z \in \mathbb{R}^N$ be an N -dimensional Gaussian random vector with covariance equal to the long run covariance of \tilde{Y}_t as Σ_Y^* (defined in Theorem 2). For a finite sample, the long run covariance is usually approximated by $\Sigma_Y^{*(T)} = \sum_{h=-(T-1)}^{T-1} T^{-1}(T - |h|)\Sigma(h)$. We then have the following.

THEOREM 3. *Assume the same conditions as in Theorem 2. Further assume $\lambda_{\min}(\Sigma_Y^{*(T)}) \geq \tau$ for a positive constant τ . In addition, $N = \mathcal{O}\{\exp(T^\delta)\}$ for $0 \leq \delta < 1/11$. Then as $T \rightarrow \infty$, we have*

$$\rho\left(T^{-1/2} \sum_{t=1}^T \tilde{Y}_t, Z\right) \rightarrow 0. \quad (2.10)$$

The result (2.10) can be seen as an analogue of the central limit theorem in a high dimensional version. It should be noted that to guarantee the \sqrt{T} convergence rate of $T^{-1} \sum_{t=1}^T \tilde{Y}_t$, the network size N is required to expand in a rate not faster than $\exp(T^\delta)$; see Appendix A.4 for more proof details.

3. PARAMETER ESTIMATION

In this section, we provide an estimation method to the NQAR model (2.1). The asymptotic properties are also established. Let $\theta(\tau) = [\beta_0(\tau), \gamma^\top(\tau), \beta_1(\tau), \beta_2(\tau)]^\top \in \mathbb{R}^{q+3}$ be the parameter vector. In addition, define $X_{it} = (1, Z_i^\top, n_i^{-1} \sum_{j=1}^N a_{ij} Y_{jt}, Y_{it})^\top \in \mathbb{R}^{q+3}$. Then $\theta(\tau)$ is estimated by

$$\hat{\theta}(\tau) = \arg \min_{\theta} \sum_{i=1}^N \sum_{t=1}^T \rho_\tau \left\{ Y_{it} - X_{i(t-1)}^\top \theta(\tau) \right\}, \quad (3.1)$$

where $\rho_\tau(u) = u\{\tau - \mathbf{1}(u < 0)\}$ is the contrast (check) function for quantile regression. Note that the estimation problem given by (3.1) is equivalent to estimating the quantile regression problem, where Y_{it} is the response variable and X_{it} is the explanatory variable. Consequently, the standard algorithms to estimate the quantile regression model

(e.g., simplex methods or interior point methods) can be employed. With regards to the computational perspective, we refer to the Chapter 6 of Koenker (2005) for more details.

Let the conditional density function of Y_{it} given \mathcal{F}_{t-1} be $f_{i(t-1)}(\cdot)$. To facilitate the study of the asymptotic properties, define $\widehat{\Omega}_0 = (NT)^{-1} \sum_{i=1}^N \sum_{t=0}^{T-1} X_{it} X_{it}^\top$ and $\widehat{\Omega}_1(\tau) = (NT)^{-1} \sum_{i=1}^N \sum_{t=0}^{T-1} f_{it}\{X_{it}^\top \theta(\tau)\} X_{it} X_{it}^\top$ for $\tau \in (0, 1)$. Computationally, $f_{it}(X_{it}^\top \theta(\tau))$ is approximated by $\hat{f}_{it}(X_{it}^\top \hat{\theta}(\tau)) = \{X_{it}^\top (\hat{\theta}(\tau_l) - \hat{\theta}(\tau_{l-1}))\}^{-1} (\tau_l - \tau_{l-1})$ for $\tau \in [\tau_{l-1}, \tau_l]$, where $\{\tau_l\}$ is a chosen grid. Next, to prove the asymptotic properties of the estimated parameters, the following assumptions are required.

(C1) (MOMENT ASSUMPTION) Assume $c_\beta < 1$, where c_β is defined in Theorem 2. Further, assume that Z_{is} are *iid* random vectors, with mean 0, covariance $\Sigma_z \in \mathbb{R}^{q \times q}$ and finite fourth order moment. The same assumption is needed for V_{it} , $1 \leq i \leq N$ and $0 \leq t \leq T$. Lastly, assume that $\{Z_i\}$ and $\{U_{it}\}$ to be mutually independent.

(C2) (NETWORK STRUCTURE)

(C2.1) (CONNECTIVITY) Treat W as the transition probability matrix of a Markov chain, whose state space is defined as the set of all the nodes $\{1, \dots, N\}$. Assume the Markov chain to be irreducible and aperiodic. In addition, define $\pi = (\pi_1, \dots, \pi_N)^\top \in \mathbb{R}^N$ to be the stationary distribution vector of the Markov chain (i.e., $\pi_i \geq 0$, $\sum_i \pi_i = 1$, and $W^\top \pi = \pi$). It is assumed that $\sum_{i=1}^N \pi_i^2 \rightarrow 0$ as $N \rightarrow \infty$.

(C2.2) (UNIFORMITY) Assume $|\lambda_1(W^*)| = \mathcal{O}(\log N)$, where W^* is the symmetric matrix as $W^* = W + W^\top$.

(C2.3) (CONVERGENCE) Assume that the following limits exist and finite: $\kappa_1 = \lim_{N \rightarrow \infty} N^{-1} \text{tr}(\Sigma_Y)$, $\kappa_2 = \lim_{N \rightarrow \infty} N^{-1} \text{tr}(W \Sigma_Y)$, $\kappa_3 = \lim_{N \rightarrow \infty} N^{-1} \text{tr}(W \Sigma_Y W^\top)$, $\kappa_4 = \lim_{N \rightarrow \infty} N^{-1} \text{tr}\{(I - G)^{-1}\}$, and $\kappa_5 = \lim_{N \rightarrow \infty} N^{-1} \text{tr}\{W(I - G)^{-1}\}$.

(C3) (EIGENVALUE-BOUND) Let $\widehat{\Omega}_1(\tau) \rightarrow_p \Omega_1(\tau)$ as $\min\{N, T\} \rightarrow \infty$ for any $\tau \in (0, 1)$, where $\Omega_1(\tau) \in \mathbb{R}^{N \times N}$ is a positive definite matrix. Moreover, there exists positive constants $0 < c_1 < c_2 < \infty$ such that $c_1 \leq \lambda_{\min}(\Omega_1(\tau)) \leq \lambda_{\max}(\Omega_1(\tau)) \leq c_2$ for any $\tau \in (0, 1)$.

(C4) (MONOTONICITY) It is assumed that $X_{it}^\top \theta(\tau)$ ($1 \leq i \leq N, 1 \leq t \leq T$) is a monotone increasing function with respect to $\tau \in (0, 1)$.

To gain insights into the conditions, we comment as follows. Condition (C1) is standard conditions on the noise term V_{it} s, nodal covariates Z_i s and $\beta(U_{it})$ s for the parameter consistency results. This condition can be relaxed to allow for the weak dependence or mixing case over time as long as the asymptotic normality still holds. Condition (C2) is set for the network structure. Specifically, condition (C2.1) ensures the connectivity on the network structure. It implies that all nodes in the network could connect each other within a finite number of steps. This condition can be supported by the empirical phenomenon named as “six degrees of separation” (Newman et al., 2011). Condition (C2.2) assures that the network has certain uniformity properties, i.e. the divergence rate of $\lambda_1(W^*)$ should be of the same rate or slower than $\log(N)$. Consider a fully connected network for example, it can be verified in such a case $\lambda_1(W^*)$ is of order $\mathcal{O}(1)$, which satisfies the condition perfectly. However, this assumption might be violated if huge heterogeneity occurs among nodes (e.g., a “super star” shaped network). Condition (C2.3) states the convergence assumption, which are the values related to network structure. To better understand the condition, we take κ_1 for illustration proposes. It can be written as $N^{-1} \text{tr}(\Sigma_Y) = N^{-1} \sum_i \Sigma_{Y,ii}$, where $\Sigma_{Y,ii}$ is the variance of node i . Consequently, this assumption is satisfied if the average variations of all the nodes in the network converge to a finite constant. Subsequently, condition (C3) assures that the law of large number assumption holds for $\widehat{\Omega}_1(\tau)$. Moreover, the condition guarantees that eigenvalues of the asymptotic covariance matrix in Theorem 4 are bounded from above and below for any $\tau \in B$. Lastly the monotonicity assumption is imposed by condition (C4) to ensure the validity of the quantile regression. Given the conditions, we provide a theorem named as *Network Bahadur Representation*, which leads to the consistency of the parameter estimation.

THEOREM 4 (Network Bahadur Representation). *Under conditions (C1)–(C4), the following representation holds uniformly over $\tau \in B$ (i.e., B is a compact set in $(0, 1)$),*

$$\widehat{\theta}(\tau) - \theta(\tau) = (NT)^{-1} \Omega_1(\tau)^{-1} \sum_{i=1}^N \sum_{t=1}^T X_{it} \psi_\tau(V_{it\tau}) + r_{NT}(\tau), \quad (3.2)$$

where $\psi_\tau(u) = \tau - I(u < 0)$, $V_{it\tau} = Y_{it} - X_{i(t-1)}^\top \theta(\tau)$, and The remainder term satisfies $\sup_{\tau \in B} |r_{NT}(\tau)| = \mathcal{O}_p\{(NT)^{-1/2}\}$. Therefore, we have $\widehat{\theta}(\tau) \xrightarrow{p} \theta(\tau)$ uniformly for $\tau \in B$ as $\min\{N, T\} \rightarrow \infty$.

The proof of Theorem 4 is given in the supplementary material (Appendix B.2). With the consistency of the parameters, we may now present the asymptotic distribution of the

estimated parameters.

THEOREM 5. *Under conditions (C1)–(C4), we have*

$$\sqrt{NT}\Sigma_\theta^{-1/2}(\tau)\{\widehat{\theta}(\tau) - \theta(\tau)\} \xrightarrow{\mathcal{L}} B_{q+3}(\tau)$$

as $\min\{N, T\} \rightarrow \infty$, where $\Sigma_\theta(\tau) = \Omega_1^{-1}(\tau)\Omega_0\Omega_1^{-1}(\tau)$ with

$$\Omega_0 = \begin{pmatrix} 1 & \mathbf{0}^\top & c_b & c_b \\ \mathbf{0} & \Sigma_z & \kappa_5 \Sigma_z r & \kappa_4 \Sigma_z r \\ c_b & \kappa_5 r^\top \Sigma_z & \kappa_3 + c_b^2 & \kappa_2 + c_b^2 \\ c_b & \kappa_4 r^\top \Sigma_z & \kappa_2 + c_b^2 & \kappa_1 + c_b^2 \end{pmatrix} \quad (3.3)$$

$c_b = c_1^{-1}c_0$, and $B_{q+3}(\tau)$ is a $(q+3)$ -dimensional Brownian bridge.

The proof of Theorem 5 is given in the supplementary material (Appendix B.3). In Theorem 4 and Theorem 5, both N and T are required to diverge to infinity to obtain a \sqrt{NT} -consistent result. Although the NQAR model requires that the adjacency matrix should be correctly specified, it is found that the consistency result still holds if the misspecification amount is under control. See Appendix B.4 in the supplementary materials for a discussion. It is noteworthy that since the nodal covariates Z_i are invariant to time, the minimum requirement is $N \rightarrow \infty$ to obtain the consistent estimation of $\gamma(\tau)$.

To better understand the convergence result given in Theorem 5, consider the case that for any fixed τ , $B_{q+3}(\tau)$ reduces to $N(0, \tau(1-\tau)I_{q+3})$. Specifically, we have the following Corollary on the asymptotic result for fixed τ .

COROLLARY 1. *Under conditions (C1)–(C4), for any fixed $\tau \in B$ we have the result $\sqrt{NT}\{\widehat{\theta}(\tau) - \theta(\tau)\} \xrightarrow{\mathcal{L}} N(0, \tau(1-\tau)\Sigma_\theta(\tau))$ as $\min\{N, T\} \rightarrow \infty$, where $B \subset (0, 1)$ is a compact set.*

Corollary 1 is a direct implication of Theorem 5. Indeed, by Corollary 1, the asymptotic normality can be obtained at any arbitrary fixed τ . This enables us to conduct pointwise (for any fixed τ) inference on the estimated parameters.

Remark 7. Given the estimated QNAR model, one is interested in measuring goodness-of-fit of the model. A possible solution is based on the approach of Koenker and Machado (1999) who looks at:

$$R^1(\tau) = 1 - \widehat{Q}(\tau)/\widetilde{Q}(\tau),$$

where $\widehat{Q}(\tau) = \sum_{i=1}^N \sum_{t=1}^T \rho_\tau\{Y_{it} - X_{i(t-1)}^\top \widehat{\theta}(\tau)\}$ and $\widetilde{Q}(\tau) = \min_{\beta_0} \sum_{i=1}^N \sum_{t=1}^T \rho_\tau(Y_{it} - \beta_0)$. Similarly, a goodness-of-fit measure can be defined for model prediction, which could also be used for model comparison.

To corroborate the theoretical results, we conduct a number of simulation studies. The details can be found in the Appendix C.1 and C.2 in the supplementary material. In the next section, we discuss an important aspect of the NQAR model: impulse analysis.

4. IMPULSE ANALYSIS

Given the NQAR structure (2.1) it is vital to check marginal influence via an impulse analysis: how does a node in the network react to an exogenous shock imposed on the other nodes at different quantile levels? Particularly, consider a stimulus $\Delta = (\delta_1, \dots, \delta_N)^\top \in \mathbb{R}^N$ imposed on V_t , i.e. shock to $V_t + \Delta$. Note here we do not consider the structural shock analysis to facilitate a simple discussion.

Then, the response for the i th node at time point t (i.e., Y_{it}) will grow to $Y_{it} + \delta_i$. Following the NQAR model (2.4), the response at time point $(t + l)$, $l \geq 1$ (i.e., \mathbb{Y}_{t+l}) is increased by

$$\text{IE}_t(l) = \prod_{k=0}^{l-1} G_{t+l-k} \Delta, \quad (4.1)$$

where $\text{IE}_t(l)$ refers to the *impulse effect* from time t to $t + l$. For instance, if $\Delta = (1, 0, \dots, 0)^\top$, then the $\text{IE}_t(l)$ is the first column of $\prod_{k=0}^{l-1} G_{t+l-k}$. Note in the a standard impulse analysis of VAR model, the autoregression matrix G_{t+l-k} is a constant matrix. Take the NAR model proposed by Zhu et al. (2017) for example, i.e., $\mathbb{Y}_t = \beta_0 + \beta_1 W \mathbb{Y}_{t-1} + \beta_2 \mathbb{Y}_{t-1} + Z_t^\top \gamma + \mathcal{E}_t$, where all the coefficients take constant forms. Immediately one could obtain the autoregression matrix $G = \beta_1 W + \beta_2 I$. However, in the NQAR model, the autoregression matrix G_t is a stochastic matrix related to $\{U_{it} : 1 \leq i \leq N\}$. As a result, $\text{IE}_t(l)$ cannot be directly evaluated as it is a random process. Therefore, we propose various impulse effects at any tail level in a tractable way.

4.1. Measurements of Impulse Effect

Before we go into the details, we discuss a straightforward way to measure the impulse effect, which can be referred to as average impulse effect (AIE). Naturally, the AIE is directly defined as the expectation of $\text{IE}_t(l)$ as $\mathbf{E}(\text{IE}_t(l)) = G^l \Delta = (b_1 W + b_2 I_N)^l \Delta$. Specifically, the AIE is only related to the average network (b_1) and momentum effect

(b_2). It is noteworthy that the AIE is no longer related to t but only depends on the time lag l . It can be further derived $|\mathbf{1}^\top \mathbf{E}(\text{IE}_t(l))| \leq N(\tilde{b}_1 + \tilde{b}_2)^l C_\Delta$, where $C_\Delta = \max_i |\delta_i|$. Therefore, it can be concluded that the AIE will reduce to 0 as $l \rightarrow \infty$, if the stationary condition in Theorem 1 is satisfied. Although the AIE is easy to understand, it could only measure the average level of impulse effect. As an extension, we propose the following three quantile specific measurements to measure the impulse effect.

TYPE I. (INTERVAL IMPULSE EFFECT) It can be noted that the AIE can characterize the impulse effect on average. However, it cannot capture asymmetric effects between different tail levels. To this end, we define the interval impulse effect (IIE) from t to $t+l$ within the interval $[\tau_1, \tau_2]$, ($0 < \tau_1 < \tau_2 < 1$) as

$$\begin{aligned} \text{IIE}_{\tau_1\tau_2}(l) &= \mathbf{E} \left\{ \prod_{k=0}^{l-1} G_{t+l-k} \Delta \mid U_{im} \in [\tau_1, \tau_2], 1 \leq i \leq N, t+1 \leq m \leq t+l \right\} \\ &= (c_{\beta_1, \tau_1\tau_2} W + c_{\beta_2, \tau_1\tau_2} I_N)^l \Delta, \end{aligned}$$

where $c_{\beta_1, \tau_1\tau_2} = \int_{\tau_1}^{\tau_2} \beta_1(u) du$ and $c_{\beta_2, \tau_1\tau_2} = \int_{\tau_1}^{\tau_2} \beta_2(u) du$. As one can see, the size of IIE is determined by the integration of $\beta_1(u)$ and $\beta_2(u)$ within any selected $[\tau_1, \tau_2]$. For example, to measure the effects in the upper tail, at the middle level and in the lower tail, one can split $(0, 1)$ equally into three intervals (i.e., $(0, 1/3)$, $[1/3, 2/3)$, $[2/3, 1)$) and compare the IIEs for different intervals respectively.

TYPE II. (IMPULSE EFFECT INTENSITY) IIE can distinguish effects at different quantiles. However, due to the unknown function forms of β_1 and β_2 , the integration can still be hard to compute. On the other hand, note that the IIE can be defined in any interval in $(0, 1)$. Motivated by this, we consider a sufficiently small interval $[\tau, \tau + \delta]$, and define the impulse effect intensity (IEI) at τ as

$$\begin{aligned} \text{IEI}_\tau(l) &= \lim_{\delta \rightarrow 0} \delta^{-l} \mathbf{E} \left(\prod_{k=0}^{l-1} G_{t+l-k} \Delta \mid U_{im} \in [\tau, \tau + \delta], 1 \leq i \leq N, t+1 \leq m \leq t+l \right) \\ &= \{ \beta_1(\tau) W + \beta_2(\tau) I_N \}^l \Delta, \end{aligned}$$

where $\beta_1(u)$ and $\beta_2(u)$ are assumed to be continuous at τ . By definition, $\text{IEI}_{l,\tau}$ can reflect the impulse impact at the τ th quantile, and is easy to compute as long as the estimates of $\beta_1(\tau)$ and $\beta_2(\tau)$ are obtained.

TYPE III. (PSEUDO QUANTILE IMPULSE RESPONSE FUNCTION) Similar to White et al. (2015), we can define the pseudo impulse response function. Recall that we impose

a stimulus on \mathbb{Y}_t and turn it to $\mathbb{Y}_t + \Delta$, and the l -step ahead the impulse effect $\text{IE}_t(l)$ is given by (4.1). We are interested in checking the change to the conditional quantile $Q_{Y_{i(t+l)}}(\tau|\mathbb{Y}_{t+l-1}, Z_i)$, which is the pseudo quantile impulse response function according to White et al. (2015). From (2.2), we have

$$Q_{Y_{i(t+l)}}(\tau|Z_i, \mathbb{Y}_{t+l-1}) = \beta_0(\tau) + \sum_{l=1}^q Z_{il}\gamma_l(\tau) + \beta_1(\tau) \sum_{j=1}^N w_{ij}Y_{j(t+l-1)} + \beta_2(\tau)Y_{i(t+l-1)}.$$

Therefore, the pseudo quantile impulse response function is given by

$$Q_{\tilde{Y}_{i(t+l)}}(\tau|Z_i, \tilde{\mathbb{Y}}_{t+l-1}) - Q_{Y_{i(t+l)}}(\tau|Z_i, \mathbb{Y}_{t+l-1}) = \beta_1(\tau) \sum_{j=1}^N w_{ij}\text{IE}_{t,j}(l-1) + \beta_2(\tau)\text{IE}_{t,i}(l-1),$$

where $\text{IE}_{t,j}(l-1)$ is the j th element of $\text{IE}_t(l-1)$. Due to the randomness of $\text{IE}_t(l-1)$, the pseudo quantile impulse response function can be measured by the above two methods.

Given the different types of impulse effect measurement, a cross-sectional impulse analysis can be conducted. Assume that one unit stimulus is imposed on the i th node, a cross-sectional impulse analysis aims to analyzing its impact on the other nodes. For instance, the impulse analysis can be critical in a network of financial institutions. It delivers an important message on the systemic risk spillover of an institution. Take the IEI as an example and assume $\Delta = (\delta_i)^\top$ with only $\delta_i = 1$ and $\delta_{i'} = 0$ (for all $i' \neq i$). The IEI from node i to j can be defined by the j th element of $\text{IEI}_\tau(l)$, which is then denoted as $\text{IEI}_\tau^{(i,j)}(l)$. Equivalently, $\text{IEI}_\tau^{(i,j)}(l)$ is equal to the (j, i) th element of the matrix $\{\beta_1(\tau)W + \beta_2(\tau)I_N\}^l$. If $\text{IEI}_\tau^{(i,j)}(l)$ is sufficiently large and decays slowly as $l \rightarrow \infty$, the j th node (e.g. risk factor) can be seriously affected by the shock on the i th node for a long time.

4.2. Influential Node Analysis

The impulse effect measures the impact from t to $t+l$ given a stimulus $\Delta = (\delta_i)^\top$. Specifically, consider that one unit stimulus is imposed on the node i with $\delta_i = 1$ and $\delta_j = 0$ for all $j \neq i$. As a result, the impulse impact of that unit stimulus could be measured with respect to a particular node i . Such amount of influence could reflect the influential power of the node, which leads to a quantification of influential nodes. Empirically, the influential nodes in a complex financial system should be paid particular attention with financial regulation.

To facilitate the analysis, first define the total network average impulse effect (TNAIE) as the summation of the cumulated AIE as $\text{TNAIE}(\Delta) = \sum_{l=0}^{\infty} \mathbf{1}^\top \mathbf{E}(\text{IE}_t(l)) = \sum_{l=0}^{\infty} \mathbf{1}^\top G^l \Delta = \mathbf{1}^\top (I - G)^{-1} \Delta$. Note that the definition is given by using the impulse measurement AIE, but it would be similar to use the other impulse measures (IIE and IEI). Further write $\text{TNAIE}(\Delta)$ as $\text{TNAIE}(\Delta) = \sum_{i=1}^N \nu_i \delta_i$, where ν_i is the i th element of the vector $\nu = (I - G^\top)^{-1} \mathbf{1}$. Then we have $\text{TNAIE}(\Delta) = \nu_i$ if we set $\delta_i = 1$ and $\delta_j = 0$ if $j \neq i$. It measures the effects of one unit perturbation from the node i on the whole network. We thus define ν_i as the *influential power* of node i .

However, in practice, ν_i can be hard to compute, as the calculation of ν involves the inverse of a high dimensional matrix $(I - G^\top) \in \mathbb{R}^{N \times N}$. Following the idea of Remark 4 of Theorem 1, we can approximate ν_i by the first order Taylor's expansion, which would lead to

$$\nu_i \approx 1/(1 - b_2) + (1 - b_2)^{-2} b_1 \sum_j n_j^{-1} a_{ji}.$$

Suppose $b_1 > 0$, then the influential power of node i is mainly determined by the quantity $\sum_j n_j^{-1} a_{ji}$, which is referred to as the *weighted degree* of the node i . Generally speaking, the weighted degree is an approximation to the influential power of the nodes. Therefore one may rank the nodes' influences based on the weighted degrees. Computationally, the calculation of this weighted degree does not involve complex computation of inverse of a high dimensional matrix $(I - G^\top)$, as well as specific values of b_1 and b_2 , but only the network structure information. As a result, nodes with larger weighted degrees tend to be followed (connected) by a large amount of nodes (i.e., $\sum_j a_{ji}$). Moreover, at the same time, the connected nodes should have less out-degrees (i.e., small n_j s).

5. FINANCIAL CONTAGION AND SHARED OWNERSHIP

In this section, we study financial risk contagion mechanisms arising from the common shared ownership information. Specifically, we focus on the Chinese Stock Market in 2013. The dataset consists of $N = 2,442$ stocks in the Chinese A share market, which are traded in the Shanghai Stock Exchange and the Shenzhen Stock Exchange. Here $N = 2,442$ is the size of cross section. For each stock, the weekly price is recorded for $T = 52$ weeks. The Y_{it} is the log-transformed weekly absolute return volatility, where the absolute return volatility is calculated as the absolute stock return for $t = 1, \dots, T$. The average volatility of all stocks at $t = 1, \dots, T$ is calculated and visualized in the left panel of Figure 1. A relatively higher volatility level can be captured in May and July. To describe the cross

sectional information, we average for each stock the volatilities over time. That leads to a median volatility level of 0.014. In addition, we calculate the cross sectional correlations for all the stocks, which leads to $N(N - 1)$ correlations. The histogram of cross sectional correlations is given in the right panel of Figure 1, where the mean correlation level is 0.105. This implies on average the stocks tend to be positively correlated.

To construct the network structure, the top ten shareholders' information for each stock are collected, which are referred to as *major* shareholders of the stock. For the i th and j th stock, $a_{ij} = 1$ if they share at least one major common shareholder, otherwise $a_{ij} = 0$. The shareholder network reflects an important information of inter-corporate dependence. Particularly, this is an important research problem of financial risk management. Corsi et al. (2016) argue about the common shareholder effect from the perspective of the diversification cost. They discovered that a reduction of diversification cost will lead to increasing level of diversification and thus increases the degree of overlap (common shareholder). This explains that why financial institutions, who have common shareholders are more likely to be highly correlated. We visualize the network structure among the top 100 stocks ranked by market values in Figure 2. The resulting network density is 3.9%. In addition, we would like to comment that one could construct the adjacency matrix by other approaches (e.g., according to their industrial background) and the proposed method could still be applied.

Besides this shared ownership information, firm specific variables are also taken into consideration. Motivated by Fama and French (2015), consider the following $K = 6$ covariates to represent stocks' fundamentals: SIZE (measured by the logarithm of market value), BM (book to market ratio), PR (increased profit ratio compared to the last year), AR (increased asset ratio compared to the last year), LEV (leverage ratio), and Cash (cash flow of the firm). Eventually all covariates are normalized within the interval $[0, 1]$.

We then proceed with the network analysis using the NQAR model. The results of our NQAR model yields Table 1. Both the estimates and the p -values are reported at quantiles $\tau = 0.05, 0.30, 0.50, 0.70,$ and 0.95 (from left to right). One could discover that the stocks have stronger network effect and momentum effect in the upper tail (i.e., $\tau = 0.95$) than the other quantiles. This indicates that stocks tend to have higher dependence through the network when the market is exposed to a higher volatility level. While on normal occasions (e.g., $\tau = 0.5$), the network effect tends to be insignificant. Besides, the size (i.e., CAP), the book to market ratio (BM), and the leverage ratios (LEV) are shown to have negative correlations with the conditional quantile level of the volatility at $\tau = 0.95$

and $\tau = 0.5$. To assess the model fitting level, we apply a model diagnosis procedure on residuals by using the QACF (quantile ACF) measure proposed by Li et al. (2015). The details and discussions are given in Appendix C.3 in the supplementary material. There is no strong evidence showing dynamic and cross-sectional dependence remained in the residuals.

Lastly, we include the NAR model (Zhu et al., 2017) for a comparison. The corresponding estimation results are given in Table 2. It can be seen that the network effect (i.e., $\hat{\beta}_1$) is no longer significant with p -value much larger than 10%. That implies the NAR model might not be suitable to detect the asymmetric effects in this volatility autoregression problem. Next, we compare the NQAR model with the LASSO method in multivariate quantile regression QLASSO (The tuning parameter is chosen to maximize $R^1(\tau)$ of (5.1)) see Hautsch et al. (2014). Specifically, we use the time periods with $t = 1, \dots, m$ for model training and the remaining for prediction. Following Koenker and Machado (1999), we define the prediction goodness-of-fit measure as

$$R^1(\tau) = 1 - \widehat{Q}(\tau)/\widetilde{Q}(\tau), \quad (5.1)$$

where $\widehat{Q}(\tau) = \sum_{i=1}^N \sum_{t=m+1}^T \rho_\tau\{Y_{it} - X_{i(t-1)}^\top \widehat{\theta}(\tau)\}$ and $\widetilde{Q}(\tau) = \min_{\beta_0} \sum_{i=1}^N \sum_{t=m+1}^T \rho_\tau(Y_{it} - \beta_0)$. Since for the LASSO method it is required that the number of nodes should be less than the total time periods for feasible estimation, we then randomly select $n = 40$ nodes each time, and keep $m = 46$ for training and the last 5 weeks for prediction. In addition, we set $\tau = 0.95$ to compare the prediction accuracy for the tail event. The experiment is repeated for $R = 500$ times to obtain reliable results. The goodness-of-fit measures are reported in Figure 5. It is evident that the NQAR model has better prediction power than the QLASSO model.

The impulse analysis as discussed in Section 4 is applied for stocks of five well-known banks in China. They include Bank of China (BOC), China Merchants Bank (CMB), Industrial and Commercial Bank of China (ICBC), Ping An Bank (PAB), and Shanghai Pudong Development Bank (SPDB). Specifically, $\text{IEI}_\tau(l)$ ($\tau = 0.05, 0.5, 0.95$) in Section 4.1 is computed within 5 banks at time lags $l = 1, \dots, 10$, which are plotted in Figure 3. The (i, j) th panel in Figure 3 denotes the impulse impact of the j th bank on the i th bank (i.e., $\text{IEI}_\tau^{(j,i)}(l)$). Significant asymmetric effects across different quantiles can be observed, where larger IEI can be detected in the upper tail ($\tau = 0.95$). Note that the estimated network effect $\beta_1(\tau)$ is very small at $\tau = 0.05$ and $\tau = 0.5$. Therefore, the resulting impulses $\text{IEI}_\tau^{(j,i)}(l)$ is much smaller than the higher quantile level $\tau = 0.95$,

which results in an almost flat impulse line in Figure 3. Moreover, it is observed that the mutual impulse impacts between BOC, CMB, and ICBC are much stronger than between the other two banks. Next, to evaluate the nodal influence, the influential node analysis is conducted. By Section 3.2, the influential power can be calculated by $\hat{\nu} = \{(1 - \hat{b}_2)I_N - \hat{b}_1 W^\top\}^{-1} \mathbf{1}$, where $\hat{b}_1 = 10^{-1} \sum_{m=0}^9 \hat{\beta}_1(0.05 + 0.1m)$ and $\hat{b}_2 = 10^{-1} \sum_{m=0}^9 \hat{\beta}_2(0.05 + 0.1m)$ are computed as the numerical approximations for b_1 and b_2 . Subsequently the influential power is plotted against the weighted degrees on the right panel of Figure 4, where a strong linear pattern is detected. In addition, the histogram of the weighted degrees is given on the left panel of Figure 4, where a skewed distribution pattern can be noticed. This indicates that a small portion of nodes possesses a large amount of influential power.

6. CONCLUSION

In this article, we propose a network quantile autoregression framework, which models network dynamics in a complex tail event-driven system. The stationarity of the model is discussed by considering the underlying stochastic process. To estimate the NQAR parameters, a minimum contrast estimation is employed. The asymptotic properties of the resulting estimators are investigated, which are closely related to the given network information. We illustrate the performance of the NQAR model via simulation studies and an application in the Chinese stock markets. In particular, significant asymmetric dependency at different levels of quantiles can be detected. Specifically, a stronger network effect can be found between stocks when higher volatility level is exposed to the market. This further confirms the usefulness of our proposed methodology.

We discuss here several potential future research topics. First, we comment on possible extensions of the model forms. First, it could be noted that the responses of the NQAR model are continuous. As a natural extension, a quantitative framework can be established for discrete response variables. Second, it should be noted that the NQAR model focus on lagged dependence. Thus an alternative modelling framework is to consider the contemporaneous correlation of nodes at the same time period, which would lead to the spatial quantile autoregression model. Next, suggested by the empirical evidence, we find that the dynamic patterns could be different during different time periods. See Appendix C.4 in the supplementary materials for discussions. Particularly, one tends to see stronger asymmetric network effects when the market exhibits higher turbulence. Therefore, more flexible model forms can be designed to model this phenomenon. Lastly, heteroskedasticity is a pervasive phenomenon in financial data and should be taken into

consideration. Although it has been shown in Remark 3 that the NQAR model could allow for certain types of heteroskedasticity, however, it is still worthwhile to discuss more general settings. In addition, with respect to the model form, one could consider allowing node specific network and momentum functions to reflect more heterogeneity in the network.

Next, the specification of the adjacency matrix should be investigated. It is found that the estimation would be biased when the network relationship is seriously mis-specified. As a complement to the NQAR model, one could consider the approach of Hautsch et al. (2014) to estimate the network relationship among the nodes. In addition to that, bias correction methods should be developed accordingly. Another flexible approach is to make the adjacency matrix to be time varying and related to exogenous covariates. Under that specification, new estimation methods should be discussed.

Thirdly, note that the NQAR model (2.1) requires continuous observations for each subject i . This may not be applicable in some scenarios. For example, the daily stock price will be missing when the stock market is closed on weekends. The model should be further adjusted to allow for missing values in such scenarios.

Lastly, the impulse analysis should be further investigated to allow for possibly structural shocks. Consider the NAR model for example, i.e., $\mathbb{Y}_t = \beta_0 + \beta_1 W \mathbb{Y}_{t-1} + \beta_2 \mathbb{Y}_{t-1} + Z_i^\top \gamma + \mathcal{E}_t$, where the coefficients take constant forms and $\text{Cov}(\mathcal{E}_t) = \Sigma_\varepsilon$. By conducting a Cholesky decomposition on Σ_ε and assuming an empirical causal chain of the nodes on the identification of the structural model, we could have $\Sigma_\varepsilon = LL^\top$. An equivalent structural VAR form is $L^{-1}\mathbb{Y}_t = \beta_0 L^{-1} + L^{-1}G\mathbb{Y}_{t-1} + L^{-1}Z_i^\top \gamma + e_t$, where $G = \beta_1 W + \beta_2 I$ and $e_t = L^{-1}\mathcal{E}_t$. Further it can be transformed to $\mathbb{Y}_t = (I - L^{-1})\mathbb{Y}_t + \beta_0 L^{-1} + L^{-1}G\mathbb{Y}_{t-1} + L^{-1}Z_i^\top \gamma + e_t$. Given this form, one could be able to analyze the instantaneous effect by making an impulse on \mathbb{Y}_t . We refer to Lütkepohl (2005) for more discussions. However, in the case of NQAR model, it is not so straightforward. That is because that the NQAR model has a non-linear model structure. Furthermore, it is assumed the underlying generating process U_{it} is independent with each other. A possible extension is to allow dependency across nodes for U_{it} . That will facilitate the discussion of the structural shock for the NQAR model. However, in such a case, the dependency structure will exist simultaneously with $\{\beta_k(U_{it}), 1 \leq i \leq N\}$ for $k = 0, 1, 2$. The analytical form of the impulse function would thus not be explicit. Since the extension might be non-trivial, we would like to leave it for future research topics.

References

- Abello, J., Pardalos, P. M., and Resende, M. G. (2013). Handbook of massive data sets, volume 4. Springer.
- Acemoglu, D., Carvalho, V. M., Ozdaglar, A., and Tahbaz-Salehi, A. (2012). The network origins of aggregate fluctuations. Econometrica, 80(5):1977–2016.
- Amini, A. A., Chen, A., Bickel, P. J., Levina, E., et al. (2013). Pseudo-likelihood methods for community detection in large sparse networks. The Annals of Statistics, 41(4):2097–2122.
- Antón, M., Ederer, F., Giné, M., and Schmalz, M. C. (2018). Common ownership, competition, and top management incentives. Working Paper.
- Billio, M., Getmansky, M., Lo, A. W., and Pelizzon, L. (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. Journal of Financial Economics, 104(3):535–559.
- Cappiello, L., Gérard, B., Kadareja, A., and Manganelli, S. (2014). Measuring comovements by regression quantiles. Journal of Financial Econometrics, 12(4):645–678.
- Chavleishvili, S. and Manganelli, S. (2016). Quantile impulse response functions. Working Paper.
- Chen, C. Y.-H., Härdle, W. K., and Okhrin, Y. (2018). Tail event driven networks of sifis. SFB discussion paper 2017004, submitted to Journal of Econometrics.
- Chen, X., Chen, Y., and Xiao, P. (2013). The impact of sampling and network topology on the estimation of social intercorrelations. Journal of Marketing Research, 50(1):95–110.
- Corsi, F., Marmi, S., and Lillo, F. (2016). When micro prudence increases macro risk: The destabilizing effects of financial innovation, leverage, and diversification. Operations Research, 64(5):1073–1088.
- Cressie, N. and Wikle, C. K. (2015). Statistics for spatio-temporal data. John Wiley & Sons.
- Diebold, F. X. and Yilmaz, K. (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. Journal of Econometrics, 182(1):119–134.

- Engle, R. F. and Manganelli, S. (2004). Caviar: Conditional autoregressive value at risk by regression quantiles. Journal of Business and Economic Statistics, 22:367–381.
- Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. Journal of Financial Economics, 116(1):1–22.
- Fan, J. and Fan, Y. (2010). Issues on quantile autoregression. Manuscript.
- Fitzenberger, B., Koenker, R., and Machado, J. A. (2013). Economic applications of quantile regression. Springer Science & Business Media.
- Gaglianone, W. P., Lima, L. R., Linton, O., and Smith, D. R. (2011). Evaluating value-at-risk models via quantile regression. Journal of Business & Economic Statistics, 29(1):150–160.
- Granger, C. W., Huang, B.-N., and Yang, C.-W. (2000). A bivariate causality between stock prices and exchange rates: evidence from recent asian flu? The Quarterly Review of Economics and Finance, 40(3):337–354.
- Härdle, W. K., Wang, W., and Yu, L. (2016). TENET: Tail-event driven network risk. Journal of Econometrics, 192(2):499–513.
- Hasan, M. N. and Koenker, R. W. (1997). Robust rank tests of the unit root hypothesis. Econometrica: Journal of the Econometric Society, 65(1):133–161.
- Hautsch, N., Schaumburg, J., and Schienle, M. (2014). Financial network systemic risk contributions. Review of Finance, 19(2):685–738.
- Herz, A. (2015). Relational constitution of social support in migrants’ transnational personal communities. Social Networks, 40:64–74.
- Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2012). Latent space approaches to social network analysis. Journal of the American Statistical Association, 97(460):1090–1098.
- Koenker, R. (2005). Quantile regression. Number 38. Cambridge university press.
- Koenker, R. and Bassett, G. W. (1978). Regression quantiles. Econometrica, 46:33–50.
- Koenker, R. and Hallock, K. F. (2001). Quantile regression. Journal of Econometric Perspectives, 15(4):143–156.

- Koenker, R. and Machado, J. A. (1999). Goodness of fit and related inference processes for quantile regression. Journal of the American Statistical Association, 94(448):1296–1310.
- Koenker, R. and Xiao, Z. (2006). Quantile autoregression. Journal of the American Statistical Association, 101(475):980–990.
- Lee, L. and Yu, J. (2009). Spatial nonstationarity and spurious regression: the case with a row-normalized spatial weights matrix. Spatial Economic Analysis, 4(3):301–327.
- Lee, L. F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. Econometrica, 72(6):1899–1925.
- Li, G., Li, Y., and Tsai, C.-L. (2015). Quantile correlations and quantile autoregressive modeling. Journal of the American Statistical Association, 110(509):246–261.
- Lütkepohl, H. (2005). New introduction to multiple time series analysis. Springer Science & Business Media.
- Newman, M., Barabasi, A. L., and Watts, D. J. (2011). The structure and dynamics of networks. Princeton University Press.
- Novy, D. (2013). Gravity redux: measuring international trade costs with panel data. Economic inquiry, 51(1):101–121.
- Portnoy, S. (1991). Asymptotic behavior of regression quantiles in non-stationary, dependent cases. Journal of Multivariate Analysis, 38(1):100–113.
- Portnoy, S. (1997). Local asymptotics for quantile smoothing splines. The Annals of Statistics, 25(1):414–434.
- Sewell, D. K. and Chen, Y. (2015). Latent space models for dynamic networks. Journal of the American Statistical Association, 110(512):1646–1657.
- Sewell, D. K., Chen, Y., et al. (2017). Latent space approaches to community detection in dynamic networks. Bayesian Analysis, 12(2):351–377.
- White, H., Kim, T.-H., and Manganelli, S. (2015). Var for var: measuring systemic risk using multivariate regression quantiles. Journal of Econometrics, 187(1):169–188.
- Zhang, D. and Wu, W. B. (2015). Gaussian approximation for high dimensional time series. arXiv preprint arXiv:1508.07036.

- Zhang, J. and Chen, Y. (2013). Sampling for conditional inference on network data. Journal of the American Statistical Association, 108(504):1295–1307.
- Zhang, X. and Cheng, G. (2014). Bootstrapping high dimensional time series. arXiv preprint arXiv:1406.1037.
- Zhao, Y., Levina, E., and Zhu, J. (2011). Community extraction for social networks. Proceedings of the National Academy of Sciences, 108(18):7321–7326.
- Zhou, J., Tu, Y., Chen, Y., and Wang, H. (2017). Estimating spatial autocorrelation with sampled network data. Journal of Business & Economic Statistics, 35(1):130–138.
- Zhu, X., Huang, D., Pan, R., and Wang, H. (2018). Multivariate spatial autoregression for large scale social networks. Working Paper.
- Zhu, X., Pan, R., Li, G., Liu, Y., and Wang, H. (2017). Network vector autoregression. Annals of Statistics, 45(3):1096–1123.
- Zou, T., Lan, W., Wang, H., and Tsai, C.-L. (2017). Covariance regression analysis. Journal of the American Statistical Association, 112(517):266–281.

Table 1: The detailed NQAR analysis results for the Chinese Stock dataset ($\tau = 0.05, 0.5, 0.95$). The parameter estimates ($\times 10^{-2}$) are reported for $\tau = 0.05, 0.5, 0.95$, where the standard error ($\times 10^{-2}$) is given in parentheses. The p-values are also reported.

	$\tau = 0.05$		$\tau = 0.3$		$\tau = 0.5$		$\tau = 0.7$		$\tau = 0.95$	
	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
$\hat{\beta}_0$	0.05 (0.00)	< 0.01	0.55 (0.02)	< 0.01	1.00 (0.04)	< 0.01	1.50 (0.05)	< 0.01	2.96 (0.13)	< 0.01
$\hat{\beta}_1$	0.00 (0.01)	0.99	-1.57 (0.60)	< 0.01	-0.04 (0.77)	0.95	2.29 (0.89)	0.01	6.09 (2.16)	< 0.01
$\hat{\beta}_2$	4.16 (0.14)	< 0.01	23.75 (0.39)	< 0.01	35.70 (0.47)	< 0.01	47.18 (0.54)	< 0.01	67.84 (1.13)	< 0.01
SIZE	0.00 (0.01)	0.98	-0.35 (0.07)	< 0.01	-1.00 (0.09)	< 0.01	-1.75 (0.11)	< 0.01	-4.10 (0.28)	< 0.01
BM	0.00 (0.01)	0.99	-0.10 (0.04)	0.02	-0.29 (0.04)	< 0.01	-0.56 (0.08)	< 0.01	-0.71 (0.25)	< 0.01
PR	0.00 (0.00)	1.00	-0.30 (0.02)	< 0.01	-0.30 (0.12)	0.01	-0.11 (0.13)	0.39	0.39 (0.38)	0.31
AR	-0.02 (0.03)	0.55	-0.30 (0.03)	< 0.01	-0.66 (0.11)	< 0.01	-0.24 (0.40)	0.54	-0.47 (0.36)	0.20
CASH	-0.01 (0.00)	0.01	-0.02 (0.05)	0.72	-0.14 (0.06)	0.01	-0.15 (0.08)	0.05	-0.05 (0.27)	0.86
LEV	0.00 (0.01)	0.97	-0.58 (0.05)	< 0.01	-0.79 (0.05)	< 0.01	-1.34 (0.02)	< 0.01	-2.42 (0.44)	< 0.01

Table 2: The NAR analysis results for the Chinese Stock dataset. The parameter estimates ($\times 10^{-2}$) and the standard error ($\times 10^{-2}$) are reported. The p-values are also given.

	Estimate	SE	<i>p</i> -value
BASELINE $\hat{\beta}_0$	1.19	0.04	< 0.01
NETWORK $\hat{\beta}_1$	0.22	0.67	0.74
MOMENTUM $\hat{\beta}_2$	36.79	0.27	< 0.01
SIZE $\hat{\gamma}_1$	-1.60	0.10	< 0.01
BM $\hat{\gamma}_2$	-0.35	0.08	< 0.01
PR $\hat{\gamma}_3$	-0.03	0.10	0.75
AR $\hat{\gamma}_4$	-0.19	0.15	0.19
CASH $\hat{\beta}_0$	0.02	0.11	0.82
LEV $\hat{\beta}_1$	-1.34	0.10	< 0.01

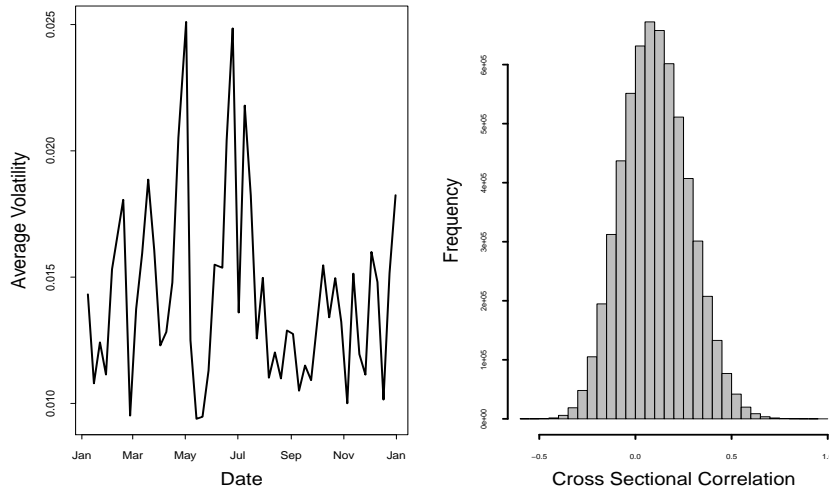


Figure 1: Left panel: the average stock volatility of Chinese A stock market in 2013. Higher volatility level can be captured in the first half of 2013; right panel: histogram of cross sectional correlations for the stocks. The average correlation level is 0.105.

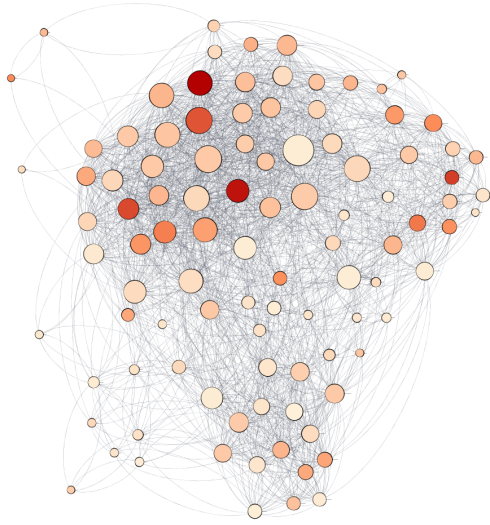


Figure 2: The common shareholder network of top 100 market value stocks in 2013. The larger and darker points imply higher market capitalization.

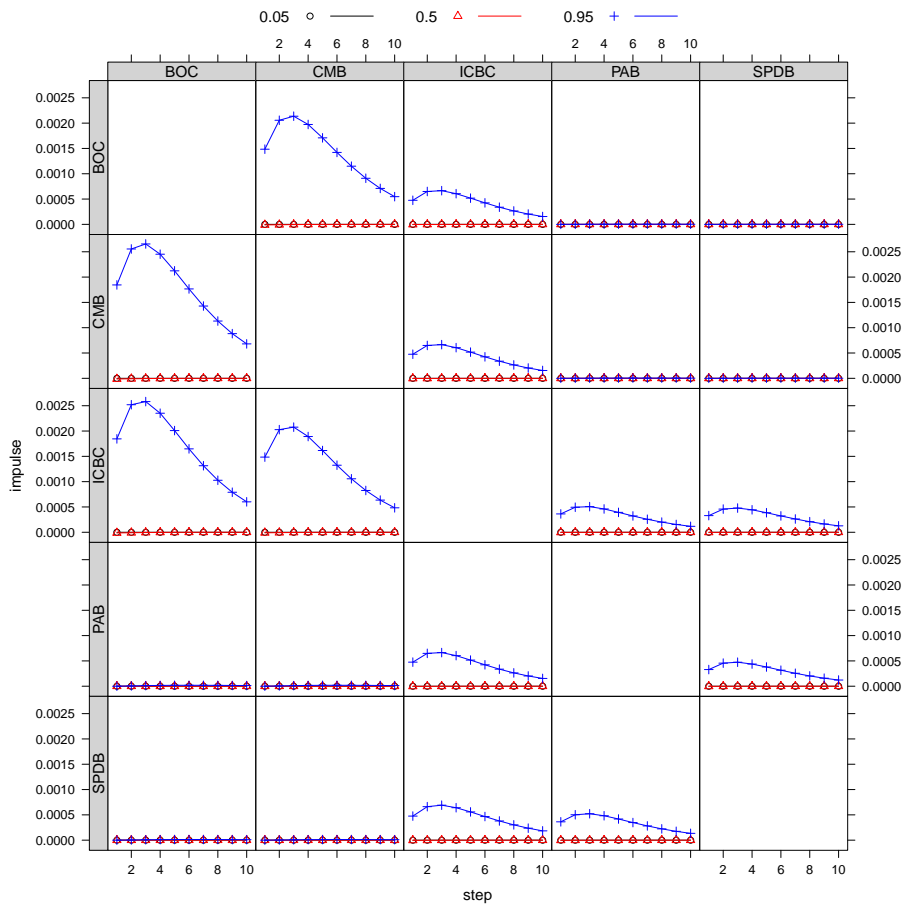


Figure 3: Impulse analysis for $\tau = 0.05, 0.5, 0.95$. The cross-sectional impulse effect intensity between BOC, CMB, ICBC, PAB, and SPDB are given. The (i, j) th panel denotes the impulse impact of the j th bank on the i th bank (i.e., $IEI_{\tau}^{(j,i)}(l)$).

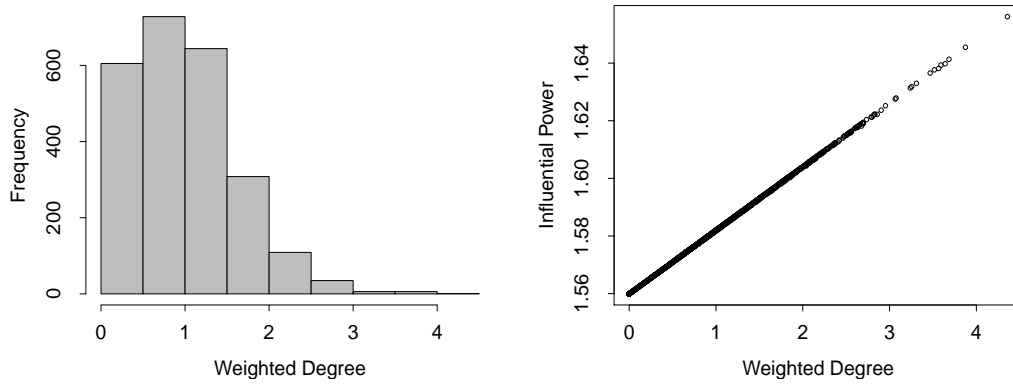


Figure 4: Left panel: the histogram of the weighted degrees; right panel: the influential power against weighted degrees.

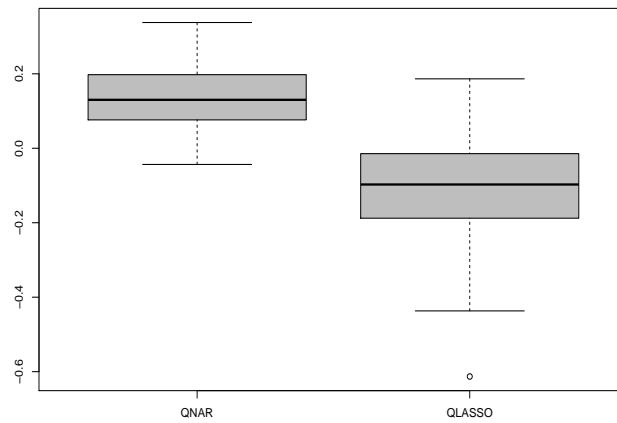


Figure 5: The prediction goodness-of-fit measure $R^1(\tau)$ at $\tau = 0.95$ for the QNAR model and the QLASSO methods. Better prediction performance can be detected for the QNAR model.



Dynamic credit default swap curves in a network topology

Xiu Xu, Cathy Yi-Hsuan Chen & Wolfgang Karl Härdle

To cite this article: Xiu Xu, Cathy Yi-Hsuan Chen & Wolfgang Karl Härdle (2019) Dynamic credit default swap curves in a network topology, *Quantitative Finance*, 19:10, 1705-1726, DOI: [10.1080/14697688.2019.1585560](https://doi.org/10.1080/14697688.2019.1585560)

To link to this article: <https://doi.org/10.1080/14697688.2019.1585560>



Published online: 15 Mar 2019.



Submit your article to this journal [↗](#)



Article views: 184



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 1 View citing articles [↗](#)

Dynamic credit default swap curves in a network topology

XIU XU  ^{†*}, CATHY YI-HSUAN CHEN ^{‡§} and WOLFGANG KARL HÄRDLE ^{‡¶||††}

[†]Dongwu Business School, Soochow University, Suzhou, China

[‡]Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Berlin, Germany

[§]Adam Smith Business School, University of Glasgow, Glasgow, UK

[¶]Sim Kee Boon Institute for Financial Economics, Singapore Management University, Singapore

^{||}Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen, China

^{††}Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

(Received 27 November 2017; accepted 1 February 2019; published online 15 March 2019)

Systemically important banks are connected and their default probabilities have dynamic dependencies. An extraction of default factors from cross-sectional credit default swap (CDS) curves allows us to analyze the shape and the dynamics of default probabilities. In extending the Dynamic Nelson Siegel (DNS) model to an across firm multivariate setting, and employing the generalized variance decomposition of Diebold and Yilmaz [On the network topology of variance decompositions: Measuring the connectedness of financial firms. *J. Econom.*, 2014, 182(1), 119–134], we are able to establish a DNS network topology. Its geometry yields a platform to analyze the interconnectedness of long-, middle- and short-term default factors in a dynamic fashion and to forecast the CDS curves. Our analysis concentrates on 10 financial institutions with CDS curves comprising of a wide range of time-to-maturities. The extracted level factor representing long-term default risk shows a higher level of total connectedness than those derived for short-term and middle-term default risk, respectively. US banks contributed more to the long-term default spillover before 2012, whereas European banks were major default transmitters during and after the European debt crisis, both in the long-term and short-term. The comparison of the network DNS model with alternatives proposed in the literature indicates that our approach yields superior forecast properties of CDS curves.

Keywords: CDS; Network; Default risk; Variance decomposition; Risk management

JEL classification: C32, C51, G17

1. Introduction

The probability of default (PD) is defined as the likelihood of a default with respect to obligations over a particular time horizon. A PD ‘curve’ maps PDs of different maturity horizons and typically shows an upward slope, although it may occasionally display a downward slope to signal a severe credit deterioration in the short-term relative to the long-term. The PD curve varies in time and, as in the analysis of term structure for credit default swaps (CDS) spreads, it carries information on co-movements and common factors. An extraction of the common factors from cross-sectional CDS spreads allows the shape and the dynamics of the PD to be analyzed. To be more specific, the shape of the PD curve can be parsimoniously inferred by projecting the cross-sectional

CDS spreads with different maturities on a few factors. The dynamics of the curves and their interplay, which is cast in the form of a network topology, reflects their interdependency with a controllable dimensionality.

With an understanding of the dynamics and interplay of PD curves, the next research question is how these lead to default risk contagion? The literature on this topic mainly studies two issues: How to quantitatively measure the contagion risk? and through which channel is contagion risk effected? The two issues are often intertwined. Concerning the transmission of default risk, the extant research has two focuses: the first is to examine the common determinants of CDS spread that may trigger contagion (Alexander and Kaeck 2008, Das *et al.* 2009, Galil *et al.* 2014). The second stream aims to directly measure the connectedness and contagion in a system. By using 5-year CDS spread data, Yang and Zhou (2013) propose a PCA framework to extract

*Corresponding author. Email: xiux@suda.edu.cn

common factors, and build up a network based on the constructed latent factors. Filipović and Trolle (2013) infer a term structure of interbank risk, and further conduct a decomposition for the default and non-default component. Bostanci and Yilmaz (2015) focus on connectedness of global sovereign credit risk. Puliga *et al.* (2014) measure the interdependency of financial institutions via a network and propose systemic risk stress tests. Getmansky *et al.* (2016) characterize the CDS market dynamics via volume-based statistics. However, as noted by Pan and Singleton (2008), the degree of connectedness or contagion depends upon which maturities of CDS contract are chosen. The term structure of contagion risk that can be characterized by the joint dynamics of CDS curves has not yet been fully studied in the current literature.

This study strives to analyze the term structure of default risk in order to quantify the connectedness and the resulting contagion across different horizons. It is different from the above cited earlier research in several aspects. First, we consider the CDS curve as a whole and model the joint dependency of the curve characteristics (e.g. level, slope and curvature) rather than the correlation of spreads. The CDS curve yields information on the risk-neutral default probability over different time horizons. The market participants rely on this curve to interpret the market's expectation of default risk at different time horizons, to manage credit risk and to design credit derivative contracts. Second, we derive the CDS dynamics through a DNS technique and produce short-term (ST), middle-term (MT) and long-term (LT) default factors that we stack into a multivariate time series. Third, by applying a VAR approach and the generalized variance decomposition machinery of Diebold and Yilmaz (2014), we obtain an interpretable network structure and total connectedness measures. Fourth, the proposed network, as an extension of the DNS model to a multivariate setting, allows one to analyze the interconnectedness in a dynamic fashion and to forecast the CDS curves, even for a 30-year horizon.

Individual default risks may build up into a systemic risk that impacts the stability of the architecture of the financial system (Longstaff *et al.* 2005, Benzoni *et al.* 2015). Indeed, a default event in a bank can grow in to a systemic failure, extreme turbulence and widespread distress triggering a contagion, as described in Duffie *et al.* (2009). Knowledge of correlated exposure holdings is therefore important to understand interbank obligations, and this is the very reason why we focus on the global systemically important banks (G-SIBs). To stabilize financial markets, the Financial Stability Board (FSB) strives to identify the systemically important financial institutions and rank them in terms of their systemic relevance. To match this point with our network analysis we extend earlier approaches to a network topology of variance decomposition, that enables us to quantify default spillover, contagion and interconnectedness. Given the ST, MT and LT default factors, it is of interest to know to what extent the credit condition of a bank is subject to its past, and also how it depends on the default risk of other banks. Is risk contagion driven by ST or LT credit horizons or both? Can we use this information

to foresee a contagion and evaluate risk transmissions among creditors?

It is worthwhile to relate these issues to CDS pricing that needs to take credit contagion into account. One may thus produce better out-of-sample forecast performance for a CDS curve, here for an entire curve instead of a point prediction. Precise forecasts are appreciated by counterparties. For buyers who wish to hedge credit risk, a good forecast is very decisive in gauging a CDS contract, since they can benefit through a relatively lower insurance payment. For sellers, the default network information helps them to avoid underestimating CDS spreads. We will detail the forecast performance in section 5. Our network formulation of the joint CDS structure also helps bondholders with different time horizons of credit exposure, policy makers with policies designed for the ST and the LT perspectives, portfolio managers who wish to diversify their bond portfolios, and credit agencies who wish to assign a credit rating to firms.

Our primary findings are: (1) The CDS curves of G-SIBs banks tightly comove and highly connect with each other, especially in the long-term. US banks contribute more to the long-term default contagion before 2012, whereas the European banks are major default transmitters during and after the European debt crisis in both the long-term and short-term. (2) The time-varying total default connectedness serves as an indicator of systemic risk, it especially identifies a clustering default sub-system. A rising total default connectedness signals a higher likelihood of default contagion. (3) The channels of credit contagion vary with credit horizon. In the long-term, credit contagion is more likely to be experienced through a credit market (TED and credit spread) and a change of investor risk attitude, while in the short-term, contagion conditions of the credit market are exclusively relevant. (4) The network-based DNS model, relative to the DNS model, yields better out-of-sample prediction for CDS curves.

The rest of this paper is organized as follows: the network-based DNS model is introduced in section 2 and model estimation is detailed in section 3. Section 4 describes the data and offers a preliminary analysis. We summarize the empirical results and detail the analysis of static and dynamic connectedness measures in section 5. Finally, section 6 concludes this paper.

2. Modeling framework

After discussing the information content of CDS spreads that provides the theoretical foundations, this section details the procedures of modeling the CDS curves with dynamic Nelson Siegel (DNS) framework, in which the three Nelson Siegel parameters (i.e. level, slope and curvature) can be viewed as the long, medium and short term factor of the CDS curves. Furthermore, to measure the systemic default caused by credit contagion, we use the Diebold-Yilmaz connectedness measures, which rely on the variance decompositions in a vector autoregressive (VAR) model. In addition, the out-of-sample forecasting framework with or without other banks' information are also introduced.

2.1. Implied default intensity in CDS spreads

The basic pricing formula for CDS contracts is to achieve a payoff balance between CDS buyers and CDS sellers. Consider a CDS contract with a maturity of M years and quarterly premium payments. $CDS_t(M)$ denotes the annualized spread at issue. L is the risk-neutral expected loss of the notional value in the event of default. We normalize the notional face value of the contract as 1. λ_t denotes the risk-neutral arrival rate of a credit event (i.e. default intensity). Then, at issue, the present value of CDS-provider side and that of CDS-buyer side should be equal,

$$\begin{aligned} & \frac{1}{4} CDS_t(M) \sum_{j=1}^{4M} E_t \left[\exp \left\{ - \int_t^{t+j/4} (r_s + \lambda_s) ds \right\} \right] \\ & = L \int_t^{t+M} E_t \left[\lambda_u \exp \left\{ - \int_t^u (r_s + \lambda_s) ds \right\} \right] du \end{aligned} \quad (1)$$

where r_t is the risk free rate. Pan and Singleton (2008) assumed the s years time discount factor as $\delta(s) = \exp(-\int_t^{t+s} r_u du)$, and presumed the conditional survival probability $q(s)$ as follows

$$q(s) = \exp(-\lambda_t s) \quad (2)$$

Then, (1) is transformed to

$$\begin{aligned} & \frac{1}{4} CDS_t(M) \sum_{j=1}^{4M} E_t \left\{ \delta \left(\frac{j}{4} \right) q \left(\frac{j}{4} \right) \right\} \\ & = L \sum_{j=1}^{4M} E_t \left[\delta \left(\frac{j}{4} \right) \left\{ q \left(\frac{j-1}{4} \right) - q \left(\frac{j}{4} \right) \right\} \right] \end{aligned} \quad (3)$$

Combing (3) and (2), we can directly imply the default intensity from CDS spreads,

$$\lambda_t = 4 \log \left\{ 1 + \frac{CDS_t(M)}{4L} \right\} \quad (4)$$

It is noteworthy that the explicit relationship between default intensity and CDS spreads, such as in (4), is only satisfied under certain assumptions, such as a constant loss given default L , and survival probability $q(s)$ in (2). Given that the implied default intensity is naturally dependent on the predetermined model set-up, we focus on CDS spreads as a direct indicator of default intensity (see equation (4)), to get rid of the potential model misspecification risk. CDS spreads investigation also permits us to perform an out-of-sample forecast and practice a convenient comparison.

In the assumption of continuous premium or coupon payments, the default intensity is represented as the solution of an integral equation that can be derived from the standard pricing approach for CDS (see equation (3)). The integral equation is shown to be equivalent to an ordinary linear differential equation of second order with time dependent coefficients, see Baranovski *et al.* (2009). For the special case of Nelson Siegel CDS terms structure models, this permits a closed form analytical solution to the second order differential equation.

2.2. Fitting the CDS curve via the dynamic Nelson Siegel model

2.2.1. A factor model representation. Nelson and Siegel (1987) propose a PCA based parsimonious three factor model for the cross-sectional yields at any point. Diebold and Li (2006) extend this into the DNS framework and find excellent forecasting properties for interest rates. Diebold *et al.* (2006) model the yield curve via a state space factorization and find strong influences from macro variables. Here we assume a similar term structure framework for the CDS curves with DNS modeling for several reasons. First, fitting the observed term structure of quotes for CDS spreads through the Nelson-Siegel type models is able to generate smooth and reliable curve fitting, particularly for a case where the quotes can sometimes be seen as outliers and discontinuous jumps (due to data quality issue). In this regard, the Nelson-Siegel type models are stable w.r.t outliers and noisy data. Second, the Nelson Siegel functions have the advantage of permitting a closed form analytical solution to the second order differential equation, which is directly derived from the pricing approach of CDS (Baranovski *et al.* 2009). Shaw *et al.* (2014) and Krishnan *et al.* (2010) present the application of CDS curves with a DNS framework.

Define $y_{it}(\tau)$ as the nominal CDS spreads of financial institution i on a vector comprised of τ -period maturities. $\tau = (\tau_1, \tau_2, \dots, \tau_k) = (6M, 1Y, 2Y, 3Y, 4Y, 5Y, 7Y, 10Y, 20Y, 30Y)$, $k = 10$. The DNS factorization for a single financial institute i at time point t is,

$$\begin{aligned} y_{it}(\tau) = & l_{it} + s_{it} \left\{ \frac{1 - \exp(-\delta_i \tau)}{\delta_i \tau} \right\} \\ & + c_{it} \left\{ \frac{1 - \exp(-\delta_i \tau)}{\delta_i \tau} - \exp(-\delta_i \tau) \right\} + v_{it}(\tau) \end{aligned} \quad (5)$$

$$\begin{aligned} \begin{pmatrix} l_{it} \\ s_{it} \\ c_{it} \end{pmatrix} = & \begin{pmatrix} \alpha_i^l & 0 & 0 \\ 0 & \alpha_i^s & 0 \\ 0 & 0 & \alpha_i^c \end{pmatrix} \begin{pmatrix} l_{i,t-1} \\ s_{i,t-1} \\ c_{i,t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{it}^l \\ \varepsilon_{it}^s \\ \varepsilon_{it}^c \end{pmatrix} \\ \begin{pmatrix} v_{it} \\ \varepsilon_{it} \end{pmatrix} \sim & i.i.d.N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} Q_i & 0 \\ 0 & \Sigma_i \end{pmatrix} \right] \end{aligned} \quad (6)$$

where the disturbance vector $v_{it} = [v_{it}(\tau_1), v_{it}(\tau_2), \dots, v_{it}(\tau_k)]^\top$ and $\varepsilon_{it} = (\varepsilon_{it}^l, \varepsilon_{it}^s, \varepsilon_{it}^c)^\top$. The parameter matrix α_i is diagonal in transition equation. δ the decay factor. We vary the decay factors and estimate them for each bank differently. $y_{it}(\tau)$ is the so-called term structure of CDS spreads or CDS curve of institution i at time t .

The three DNS parameters l_{it} , s_{it} , and c_{it} can be interpreted as LT, ST, and MT default factors. Since the factor loading on l_{it} is 1, which is a constant and the same for all maturities, l_{it} can be viewed as the long-term or level factor. Any increase in l_{it} will cause the entire curve to shift upwards, representing the 'level' and 'long-term' components of the credit curve. The factor loading of s_{it} is $(1 - \exp(-\delta_i \tau)) / \delta_i \tau$, starting from 1 and exponentially decreasing to 0, can be viewed as the ST factor. The 'slope' of the credit curve changes accordingly. The loading of c_{it} is $(1 - \exp(-\delta_i \tau)) / \delta_i \tau - \exp(-\delta_i \tau)$, which is a function that starts from 0 and then increases and decays to 0; hence, this is the MT factor. In summary, the shape of

the credit curve is captured by these three factors. A time-varying shape is reflected by the changing loads on the three factors.

For the purpose of portraying the interplay of default factors given a particular horizon among banks, one has the challenge of calibrating the dynamics of l_{it} , s_{it} , and c_{it} as, for example, a VAR(1) process. This motivates us to study dynamically evolving l_{it} , s_{it} , and c_{it} . Ideally, the DNS model for each bank immediately forms a state-space system, as expressed in (6). We assume the level factor, the slope, and the curvature factor are orthogonal, so that the parameter matrix α_i is diagonal in transition equation. For this state space system, we will resort to the Kalman filter estimation method. We, therefore, assume that the disturbance vector v_{it} and ε_{it} are independent and both follow a normal distribution with covariance matrix Q_i and Σ_i , respectively. By doing so, one can distill the entire CDS curves, period by period, into three dynamically evolving dimensional parameters and then model their interplay characterized by a VAR(1) process.

2.3. The network topology of DNS factors

Joint default may become systemic at the moment when banks call for bailout together or even go bankrupt sequentially. The fact of correlated default and default contagion draws more attention (see Duffie *et al.* 2009, Duan and Miao 2015). Due to interbank loans and shared credit exposures, a default risk at one bank can easily spread to others. The speed and scope of spread is subject to each bank's systemic importance. Here, we introduce a network topology of variance decomposition to measure the 'credit connectedness', which quantifies the scope of 'default risk transmission'. The embedded dynamics mechanism allows us to evaluate the speed of default risk transmission.

2.3.1. Approximating model - VAR. We endow the level, the slope, and the curvature factors with VAR(p) dynamics:

$$x_t = \sum_{k=1}^p A_k x_{t-k} + u_t, \quad u_t \sim i.i.d.N(0, \Sigma) \quad (7)$$

$$x_t = (x_{1t}, x_{2t}, \dots, x_{Nt})^\top \quad (8)$$

where x_t represents $l_t = (l_{1t}, l_{2t}, \dots, l_{Nt})^\top$, $s_t = (s_{1t}, s_{2t}, \dots, s_{Nt})^\top$, $c_t = (c_{1t}, c_{2t}, \dots, c_{Nt})^\top$. N is the number of banks, in our case, $N = 10$. The autoregression matrix A_k is $N \times N$ dimensional, p denotes the lag order of VAR. If the lag order is identically equal to 1 for l_{it} , s_{it} , and c_{it} , then equation (8) represents,

$$l_t = \begin{pmatrix} l_{1t} \\ l_{2t} \\ \vdots \\ l_{Nt} \end{pmatrix} = \begin{pmatrix} \alpha'_{11} & \alpha'_{12} & \cdots & \alpha'_{1N} \\ \alpha'_{21} & \alpha'_{22} & \cdots & \alpha'_{2N} \\ \vdots & \vdots & \cdots & \vdots \\ \alpha'_{N1} & \alpha'_{N2} & \cdots & \alpha'_{NN} \end{pmatrix} \begin{pmatrix} l_{1,t-1} \\ l_{2,t-1} \\ \vdots \\ l_{N,t-1} \end{pmatrix} + \begin{pmatrix} u'_{1t} \\ u'_{2t} \\ \vdots \\ u'_{Nt} \end{pmatrix} \quad (9)$$

$$s_t = \begin{pmatrix} s_{1t} \\ s_{2t} \\ \vdots \\ s_{Nt} \end{pmatrix} = \begin{pmatrix} \alpha^s_{11} & \alpha^s_{12} & \cdots & \alpha^s_{1N} \\ \alpha^s_{21} & \alpha^s_{22} & \cdots & \alpha^s_{2N} \\ \vdots & \vdots & \cdots & \vdots \\ \alpha^s_{N1} & \alpha^s_{N2} & \cdots & \alpha^s_{NN} \end{pmatrix} \begin{pmatrix} s_{1,t-1} \\ s_{2,t-1} \\ \vdots \\ s_{N,t-1} \end{pmatrix} + \begin{pmatrix} u^s_{1t} \\ u^s_{2t} \\ \vdots \\ u^s_{Nt} \end{pmatrix} \quad (10)$$

$$c_t = \begin{pmatrix} c_{1t} \\ c_{2t} \\ \vdots \\ c_{Nt} \end{pmatrix} = \begin{pmatrix} \alpha^c_{11} & \alpha^c_{12} & \cdots & \alpha^c_{1N} \\ \alpha^c_{21} & \alpha^c_{22} & \cdots & \alpha^c_{2N} \\ \vdots & \vdots & \cdots & \vdots \\ \alpha^c_{N1} & \alpha^c_{N2} & \cdots & \alpha^c_{NN} \end{pmatrix} \begin{pmatrix} c_{1,t-1} \\ c_{2,t-1} \\ \vdots \\ c_{N,t-1} \end{pmatrix} + \begin{pmatrix} u^c_{1t} \\ u^c_{2t} \\ \vdots \\ u^c_{Nt} \end{pmatrix} \quad (11)$$

It is well known that the VAR model (8) (if stationary) can be written as $x_t = \Theta(L)u_t$, $\Theta(L) = \Theta_0 + \Theta_1 L + \dots + \Theta_p L^p + \dots$, $\Theta_i = A_1 \Theta_{i-1} + A_2 \Theta_{i-2} + \dots + A_p \Theta_{i-p}$, where Θ_0 and A_0 are $N \times N$ identity matrix, $A_i = 0$ for $i < 0$. This representation allows us to extract connectedness information. Following Diebold and Yilmaz (2014), we resort to apply this variance decomposition to establish a network structure.

2.3.2. Pairwise directional default connectedness. Default connectedness measures the shares of forecast error variation due to shocks arising from others. This is captured by the variance decomposition, in which the forecast error variance of variable i is decomposed into parts attributed to the remaining variables in the system. The generalized variance decomposition (GVD) (Koop *et al.* 1996) yields $\tilde{d}_{ij}(H)$ as the ij th H -step component, which represents the fraction of bank i 's H -step forecast error variance due to credit shocks in bank j ,

$$\tilde{d}_{ij}(H) = \frac{\sigma_{jj}^{-1} \sum_{h=0}^{H-1} (e_j^\top \Theta_h \Sigma e_j)^2}{\sum_{h=0}^{H-1} (e_i^\top \Theta_h \Sigma \Theta_h^\top e_i)} \quad (12)$$

where σ_{jj} is the j th diagonal element in the covariance matrix Σ of the error vector u_t - the standard deviation of the error term of j th equation -, and $e_j = (0, 0, \dots, 1, \dots, 0)$, which is a zero vector except j th element unity. H denotes the forecast horizon.

Given that the sum of $\tilde{d}_{ij}(H)$ in each equation does not necessarily equal the unit - that is, $\sum_{j=1}^N \tilde{d}_{ij}(H) \neq 1$ -, we normalize as follows:

$$d_{ij}(H) = \frac{\tilde{d}_{ij}(H)}{\sum_{j=1}^N \tilde{d}_{ij}(H)} \quad (13)$$

We define the pairwise directional credit connectedness from bank j to bank i as $C_{i \leftarrow j} = d_{ij}(H)$, and note that in general $C_{i \leftarrow j} \neq C_{j \leftarrow i}$. This finally leads to the connectedness in table 1,

Note that when there is a higher H horizon in a prediction error variance, there will be a higher value of $C_{i \leftarrow j}$. When H is very small, it limits $C_{i \leftarrow j}$ to short periods. On the other hand, as H increases the corresponding $C_{i \leftarrow j}$ increases slightly due to the incrementally less valuable conditioning information. For the case $H \rightarrow \infty$, one obtains an unconditional variance decomposition. To strike a balance between these effects, we select $H = 12$. This is close to the time period

Table 1. Connectedness table.

	x_1	x_2	...	x_N	From others
x_1	d_{11}	d_{12}	...	d_{1N}	$\sum_{j=1}^N d_{1j}, j \neq 1$
x_2	d_{21}	d_{22}	...	d_{2N}	$\sum_{j=1}^N d_{2j}, j \neq 2$
...
x_N	d_{N1}	d_{N2}	...	d_{NN}	$\sum_{j=1}^N d_{Nj}, j \neq N$
To others	$\sum_{i=1}^N d_{i1}, i \neq 1$	$\sum_{i=1}^N d_{i2}, i \neq 2$...	$\sum_{i=1}^N d_{iN}, i \neq N$	$\frac{1}{N} \sum_{i,j=1}^N d_{ij}, i \neq j$

(10 days) of value at risk (VaR) required under the Basel accord and it is also in the practical rebalancing interval of portfolio management. In the empirical section, we also calculate the results for a range of horizons. It turns out that when H is larger than a certain value, around 10, $C_{i \leftarrow j}$ increases trivially. We can provide a robustness check for other H if required.

2.3.3. Interpreting the connectedness. Consider, for example, the first row of table 1. Here, the sum of the off-diagonal $d_{ij}, j \neq i$ of bank 1 accounts for shocks attributed to other banks, while as for the first column, the sum of $d_{i1}, j \neq 1$ indicates the risk contribution of bank 1. Total directional connectedness from others to i is

$$C_{i \leftarrow \bullet} = \sum_{j=1, j \neq i}^N d_{ij}(H) \tag{14}$$

Likewise, the total directional connectedness to others from j is

$$C_{\bullet \leftarrow j} = \sum_{i=1, i \neq j}^N d_{ij}(H) \tag{15}$$

The net default connectedness i is the difference ‘To’ and ‘From’:

$$C_i = C_{\bullet \leftarrow i} - C_{i \leftarrow \bullet} \tag{16}$$

The total default connectedness is:

$$C = \frac{1}{N} \sum_{i,j=1, j \neq i}^N d_{ij}(H) \tag{17}$$

Note that there are N ‘To’ and ‘From’ net connectedness adding up to total connectedness. Economically speaking, as C increases, the banks spread the default risk mutually. Hence, C in (17) is a quantitative measure of default spillover or contagion in a system.

3. Model estimation

The VAR approximating model of default intensity factor connectedness has a natural state-space model representation. If we pool all of the banks together, then the measurement and

transition equations are

$$Y_t = HX_t + v_t \tag{18}$$

$$X_t = FX_{t-1} + \varepsilon_t \tag{19}$$

where

$$Y_t = \begin{pmatrix} y_{1t}(\tau_1) \\ y_{1t}(\tau_2) \\ \vdots \\ y_{Nt}(\tau_k) \end{pmatrix}_{(Nk \times 1)}, \quad X_t = \begin{pmatrix} l_{1t} \\ s_{1t} \\ \vdots \\ c_{Nt} \end{pmatrix}_{(3N \times 1)},$$

$$v_t = \begin{pmatrix} v_{1t}(\tau_1) \\ v_{1t}(\tau_2) \\ \vdots \\ v_{Nt}(\tau_k) \end{pmatrix}_{(Nk \times 1)}, \quad \varepsilon_t = \begin{pmatrix} \varepsilon_{1t}^l \\ \varepsilon_{1t}^s \\ \vdots \\ \varepsilon_{Nt}^c \end{pmatrix}_{(3N \times 1)},$$

and

$$H = \begin{pmatrix} 1 & \frac{1 - \exp(-\delta_1 \tau_1)}{\delta_1 \tau_1} & \frac{1 - \exp(-\delta_1 \tau_1)}{\delta_1 \tau_1} - \exp(-\delta_1 \tau_1) & \dots \\ 1 & \frac{1 - \exp(-\delta_1 \tau_2)}{\delta_1 \tau_2} & \frac{1 - \exp(-\delta_1 \tau_2)}{\delta_1 \tau_2} - \exp(-\delta_1 \tau_2) & \dots \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \frac{1 - \exp(-\delta_N \tau_k)}{\delta_N \tau_k} & \frac{1 - \exp(-\delta_N \tau_k)}{\delta_N \tau_k} - \exp(-\delta_N \tau_k) & \dots \end{pmatrix}_{(Nk \times 3N)} \tag{20}$$

Meanwhile, when level, slope and curvature factors are orthogonal with autoregressive process of order one, the parameter matrix F is,

$$F = \begin{pmatrix} \alpha_{11}^l & 0 & 0 & \alpha_{12}^l & 0 & 0 \\ 0 & \alpha_{11}^s & 0 & 0 & \alpha_{12}^s & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \alpha_{N1}^c & 0 & 0 & \alpha_{N2}^c \\ \dots & \alpha_{1N}^l & 0 & 0 & & \\ \dots & 0 & \alpha_{1N}^s & 0 & & \\ \dots & \vdots & \vdots & \vdots & & \\ \dots & 0 & 0 & \alpha_{NN}^c & & \end{pmatrix}_{(3N \times 3N)} \tag{21}$$

which accommodates (6), (9), (10), and (11).

We introduce a two-step estimation method that couples the DNS model with the variance decomposition technique. In the first step, we estimate the dynamic level factor l_{it} , the slope factor s_{it} , and the curvature factor c_{it} for each bank i through the Kalman filter estimation of the state space model, based on equations (5) and (6). In the second step, by using the network framework based on variance decomposition in Diebold and Yilmaz (2014), we investigate the dynamics of LT, ST

Table 2. Banks.

	Institution	Ticker	Country
1	Bank of America	BAC	United States
2	Citygroup	C	United States
3	Goldman Sachs	GS	United States
4	J.P.Morgan	JPM	United States
5	Wells Fargo	WFC	United States
6	Deutsche Bank	DB	Germany
7	Commerzbank	CBG	Germany
8	Barclays Bank	BCS	United Kingdom
9	HSBC Bank	HBC	United Kingdom
10	UBS	UBS	Switzerland

Note: List of banks under study.

and MT default factors in a network perspective, based on equation (8). To uncover the dynamics of the network connectedness, we use a rolling window estimation for each of the factors.

4. Data

We will first introduce the used CDS spreads data and then perform the preliminary analysis to support the motivations in the subsequent study.

4.1. CDS spreads data

We focus our attention on the CDS spreads belonging to the Top 10 of the global systemically important banks (G-SIBs), i.e. $N = 10$, for their maturities in CDS contracts. In table 2, the CDS spreads with maturities ranging from 6 months and 1, 2, 3, 4, 5, 7, 10, 20, 30 years are available via DataStream. These 10 banks are selected out of the 34 G-SIBs based on the availability of sufficient maturities of CDS. Banks with less than 10 maturities of CDS are excluded from the study to ensure the representative ability of latent factors. The sample period selected is from 1 January 2008 to 31 December 2015, at daily frequency.

Figure 1 depicts a 3D plot of the time-varying CDS spread curves of Goldman Sachs and HSBC Bank. The display of the CDS curves for the banks is depicted in figure 1 in the Appendix. The CDS curves display apparently substantial level movements across time, and they also exhibit a clear commonality to support the notion of credit comovement. One can observe a simultaneous increase of credit curve into banks during the Eurozone debt crisis periods, which motivates us to analyse the connectedness across banks. Please note that the empirical results of this paper as well as the corresponding MATLAB programming codes can be found in the folder www.github.com/QuantLet/DDINetwork as well as www.quantlet.de.

4.2. Preliminary analysis of CDS spreads

In summary, there are 26 free parameters to be estimated: the 3×3 transition matrix of the three state variables containing nine free parameters, the mean state variables containing three free parameters, the one decay parameter in measurement equation, the diagonal of disturbance covariance matrix of transition equation containing 10 free parameters with each covariance for the counterpart of 10 maturities of CDS spreads, and the three free parameters constituting the diagonal of measurement disturbance covariance matrix with each for one of the three latent variables.

We use the Kalman filter to derive the state variables and CDS spreads of the next stage, after which we proceed to evaluate the unknown parameters with the maximum likelihood estimation under Gaussian distribution assumptions for the disturbance of measurement and transition equations. The initial parameter values are obtained by using the Diebold-Li two-step ordinary least squares regression. The estimated decay parameters vary across banks.

Table 3 reports summary statistics for the estimated DNS factors, which will be used in the subsequent network analysis. The level factors present least variance compared with the slope and curvature factors. The factor autocorrelations reveal that all of the factors display persistent dynamics, with the

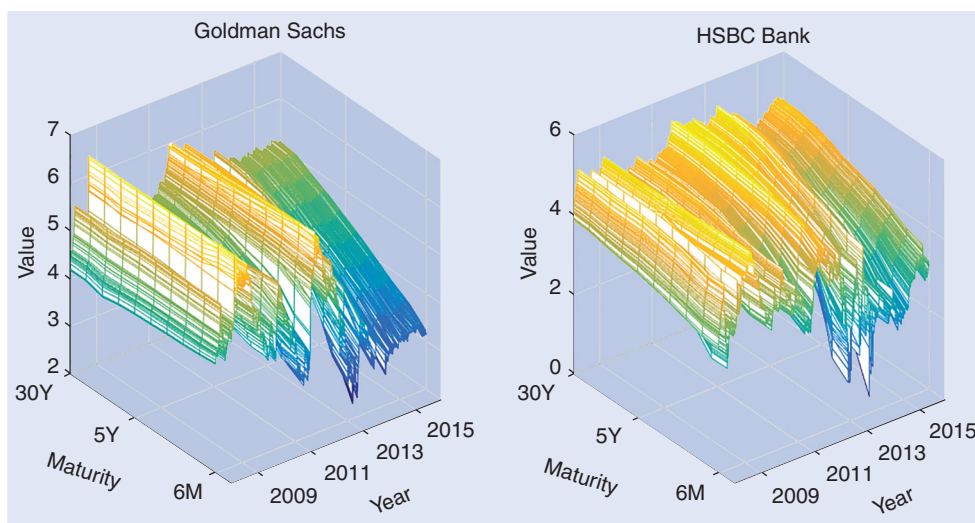


Figure 1. CDS spreads data. Note: (log) Credit default swap spreads 1 January 2008 to 31 December 2015 with daily data for Goldman Sachs and HSBC Bank across 10 maturities. The data in our study is presented after logarithm transformation.

Table 3. Descriptive statistics for the estimated DNS factor.

Factor	Mean	Std.dev.	Min	Max	$\rho(5)$	$\rho(10)$
Bank of America						
l_t	5.156	0.327	3.962	6.039	0.920	0.867
s_t	-1.475	1.160	-3.613	1.181	0.989	0.976
c_t	0.894	0.921	-1.150	3.414	0.976	0.946
Citygroup						
l_t	5.242	0.270	4.273	6.379	0.891	0.827
s_t	-1.486	1.146	-3.731	0.944	0.989	0.978
c_t	1.040	0.799	-0.823	3.276	0.972	0.944
Goldman Sachs						
l_t	5.175	0.323	3.867	6.498	0.914	0.854
s_t	-1.207	1.123	-3.273	0.914	0.989	0.977
c_t	0.751	0.822	-0.497	3.817	0.976	0.949
J.P.Morgan						
l_t	4.841	0.329	3.651	5.631	0.951	0.916
s_t	-1.717	1.029	-3.736	0.738	0.985	0.971
c_t	1.198	0.817	-0.083	4.275	0.966	0.932
Wells Fargo						
l_t	4.841	0.197	4.067	5.621	0.868	0.777
s_t	-1.868	1.232	-5.051	0.718	0.977	0.959
c_t	-0.932	1.228	-3.984	1.453	0.985	0.969
Deutsche Bank						
l_t	4.864	0.267	4.151	5.754	0.930	0.880
s_t	-1.558	0.974	-4.344	0.420	0.975	0.953
c_t	1.405	1.054	-0.589	5.226	0.954	0.916
Commerzbank						
l_t	5.027	0.439	3.927	5.898	0.966	0.945
s_t	-1.436	0.997	-3.893	0.307	0.984	0.969
c_t	0.996	0.885	-0.361	4.757	0.965	0.932
Barclays Bank						
l_t	4.933	0.353	4.048	5.764	0.938	0.902
s_t	-1.394	0.908	-3.464	0.385	0.979	0.959
c_t	1.345	0.934	-0.480	4.166	0.955	0.912
HSBC Bank						
l_t	4.778	0.321	3.948	5.329	0.972	0.945
s_t	-1.381	1.014	-6.551	0.180	0.931	0.887
c_t	-2.058	1.194	-6.017	0.472	0.959	0.927
UBS						
l_t	4.819	0.325	4.023	5.819	0.946	0.891
s_t	-1.540	1.101	-4.462	0.305	0.983	0.965
c_t	1.585	1.295	-0.277	6.928	0.964	0.927

Note: $\rho(5)$ and $\rho(10)$ denote the autocorrelation coefficients with a lag of 5 or 10 periods.

level more persistent than the slope. Although the level, the slope and the curvature factors behave distinctly, they clearly display a certain degrees of similarity across banks, as shown in figure 2.

The basis for the network analysis is the possible existence of co-movements in the factor dynamics across banks, implying the existence of potential spillover effects or an underlying transmission mechanism. To investigate this, we plot the estimated factors for all of the banks in figure 2. The upper, middle and bottom panels present the level, the slope and the curvature factor, respectively. To visualize the difference across the banks in the United States and Europe, we feature the factors of five US banks with solid line, while the counterparts of European banks are shown in dashed line. Figure 2 reveals clear evidence of co-movements in the factors dynamics, especially for the level and slope factor. One can observe that the slope/ST factors across 10 G-SIBs have climbed since the outbreak of the European debt crisis, indicating possible inverted credit curves (downward slope curves).

5. Empirical results

To characterize the evolution of the default risk connectedness among the US and European banks, we proceed to a four steps analysis: we first perform a full-sample analysis separately for level, slope, and curvature factor, to assess the unconditional or average connectedness. After this static connectedness analysis, a rolling-window sample analysis is conducted for the three factors to portray the dynamics of conditional connectedness. By doing so, one can monitor the dynamics of the spillover effect between the US and European banks over time. Using the total connectedness in the rolling-sample framework as an indicator of systematic default risk, we analyse the sources of the systematic risk. Finally, having the predefined connectedness among bank default risk, we report the forecasting performance when incorporating the factors from other banks.

5.1. Network: static

Systemic risk is not easy to define but the universally accepted characteristics are that it has a large impact, is widespread, and has a ripple effect that endangers the financial system. Network analysis enables us to cover three major concepts of systemic risk by portraying the interplay among financial institutions, measuring their interconnectedness and quantifying the spillover effect.

The interconnectedness of financial institutions on the interbank market is an absolute key to understanding systemic risk. Interconnectedness captures the situations when financial distress in one institution subsequently raises the likelihood of financial distress in other institutions because of their network of contractual relations and the interbank lending among them, leading to a ‘too-interconnected-to-fail’ situation. Therefore, the resulting connectedness parameters, such as C from (17), can be used to monitor systemic vulnerability.

In the following analysis, we examine the interconnectedness and spillover with respect to the default factors in the ST, MT and LT perspectives. This effort can help to answer the following questions: will credit spillover or contagion be evenly observed in the short-term and long-term credit horizon? And, can we use this information to foresee crisis and evaluate the tension embedded in the credit assets with different maturities?

5.1.1. Level factor. Table 4 reports the full-sample connectedness of the level factors. As the level factors capture the long-term component of CDS spreads, the entries in table 4 turn to the long-term directional connectedness measures. Many features are revealed. Blocks of high pairwise directional connectedness are notable, especially for the US banks. The values in the first five columns, which capture the spillover effect contributed to the US banks, are apparently higher than those of the European banks. The total connectedness for level factor is on average 85.50%. In addition, the ‘From’ degree distribution is noticeably less volatile than the ‘To’ degree distribution in the case of the US group but is not the case for the EU group. Through this table, one can find out

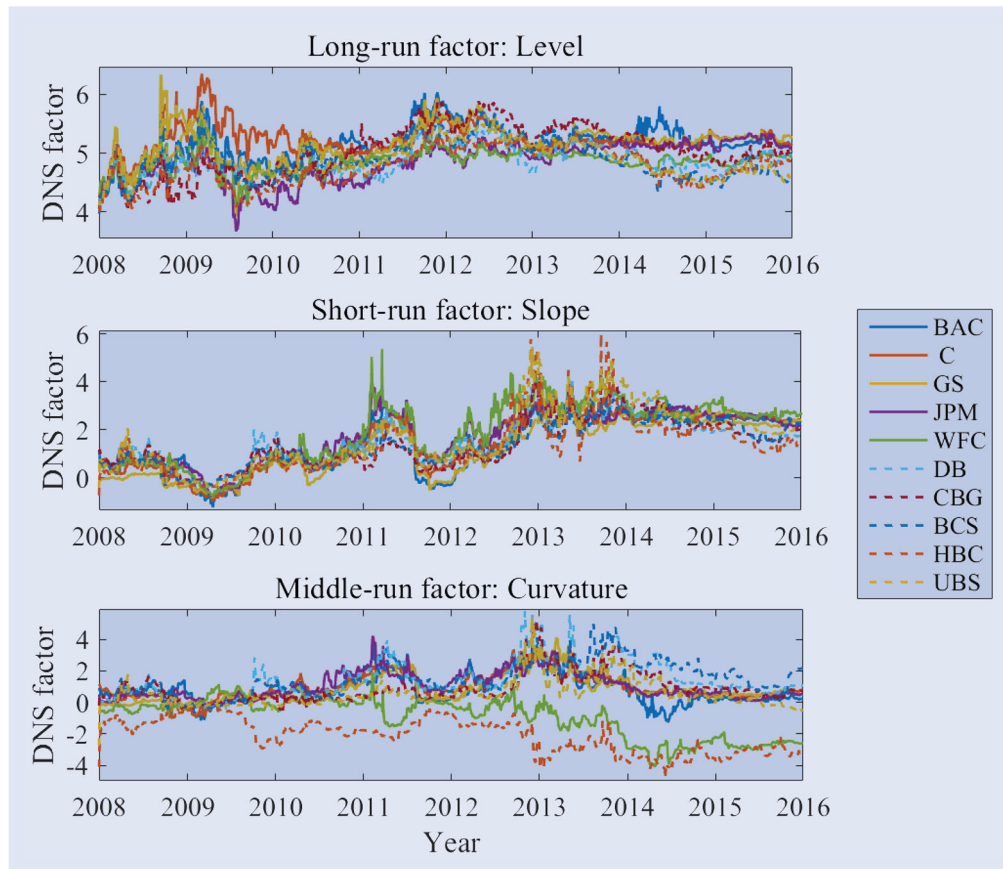


Figure 2. DNS factors. Note: Data period: 1 January 2008 to 31 December 2015 with daily data. The upper panel collects the level factors, middle panel collects the slope factors, and the bottom panel collects the curvature factors.

Table 4. Static Connectedness: Level factor.

	BAC	C	GS	JPM	WFC	DB	CBG	BCS	HBC	UBS	FROM
BAC	19.95	14.00	13.27	12.39	12.84	6.79	5.01	5.27	5.06	5.43	80.05
C	16.85	17.91	14.13	13.05	13.47	5.98	4.14	4.76	4.65	5.05	82.09
GS	16.46	18.54	20.03	14.03	14.95	3.58	2.39	2.74	3.48	3.81	79.97
JPM	17.42	17.73	16.22	15.67	15.17	4.12	2.96	3.11	3.76	3.83	84.33
WFC	16.16	16.35	15.05	14.57	15.88	5.27	3.81	3.96	4.38	4.56	84.12
DB	11.82	13.05	12.26	11.61	11.06	12.28	6.50	7.02	6.96	7.45	87.72
CBG	10.29	10.84	10.84	9.81	9.47	12.29	12.49	8.26	7.58	8.13	87.51
BCS	8.84	10.46	10.22	9.57	8.99	12.97	8.40	12.71	8.55	9.29	87.29
HBC	11.37	12.94	12.45	11.48	11.10	10.05	7.05	8.12	7.88	7.56	92.12
UBS	11.27	12.57	11.74	10.89	10.40	10.45	6.03	8.98	7.44	10.23	89.77
TO	120.47	126.48	116.17	107.40	107.44	71.49	46.30	52.23	51.87	55.11	85.50
NET	40.42	44.39	36.20	23.07	23.33	-16.23	-41.22	-35.07	-40.25	-34.65	-

Note: Data period: 1 January 2008 to 31 December 2015 with daily data.

that the US banks are exporting their LT default risk to the EU banks.

Let us discuss some of the features of the long-term connectedness table in more detail. The highest observed pairwise connectedness is from C to GS ($C_{GS \leftarrow C} = 18.54\%$), while, in return, the pairwise connectedness from GS to C ($C_{C \leftarrow GS}$) is slightly smaller 14.13%. The next highest pairwise connectedness is from C to JPM ($C_{JPM \leftarrow C} = 17.73\%$), which is slightly higher than the pairwise connectedness from BAC to JPM ($BAC_{JPM \leftarrow C} = 17.42\%$). Bank C had the largest market capitalization before the 2008 financial crisis periods, and it is reasonable that its total connectedness to others are the largest. Banks with bigger market capitalization are more capable of

offering interbank loans to other banks, which is the so-called ‘too big to fail’. On the other hand, the long-term pairwise directional connectedness among European banks is relatively small (less than 10%), except a few relatively large measures from DB to BCS ($C_{BCS \leftarrow DB} = 12.97\%$) and from DB to CBG ($C_{CBG \leftarrow DB} = 12.29\%$).

The ‘From’ column is the row sum of the pairwise connectedness except the own-effects (diagonal elements of the matrix). It reveals the total directional connectedness from others for each of the ten banks. In other words, it captures the contribution of credit shocks resulting from other banks to the total variance of the forecast error of bank i . While the total directional connectedness is distributed tightly, the

Table 5. Static Connectedness: Slope factor.

	BAC	C	GS	JPM	WFC	DB	CBG	BCS	HBC	UBS	FROM
BAC	46.65	11.91	8.89	7.71	4.36	4.94	4.66	2.97	4.20	3.72	53.35
C	12.66	15.97	7.80	9.58	9.54	8.19	6.87	7.91	16.05	5.44	84.03
GS	12.20	14.42	9.73	11.00	9.83	9.74	7.55	7.87	10.72	6.93	90.27
JPM	7.00	8.76	3.72	12.13	6.42	10.56	5.48	8.52	33.71	3.70	87.87
WFC	10.08	13.86	6.86	13.84	17.56	8.27	6.26	7.12	10.68	5.48	82.44
DB	8.04	10.09	5.74	9.11	6.73	21.65	10.49	10.29	9.18	8.67	78.35
CBG	6.71	8.88	4.34	9.00	12.05	10.43	10.56	8.14	24.48	5.40	89.44
BCS	6.86	8.93	5.14	7.67	5.07	17.13	13.26	17.49	8.59	9.86	82.51
HBC	3.04	3.73	2.07	3.41	3.00	15.07	9.32	13.14	39.23	8.00	60.77
UBS	6.22	8.11	4.53	7.16	5.41	15.48	11.06	12.36	17.83	11.85	88.15
TO	72.81	88.67	49.09	78.48	62.41	99.81	74.95	78.31	135.43	57.21	79.72
NET	19.46	4.64	-41.18	-9.39	-20.02	21.45	-14.49	-4.20	74.67	-30.94	-

Note: Data period: 1 January 2008 to 31 December 2015 with daily data.

'From' effects of US banks appear consistently smaller than those of the European banks, showing that the US banks are less impacted by the EU credit shocks.

The column sum of pairwise connectedness quantify the spillover effect of bank i to others. By definition, each bank's share in the forecast error variance of others is not compulsorily to add up to 1, therefore, elements in the 'To' row can exceed 100%. 'To' effect varies over banks, ranging from 126% to 46%. The largest commercial banks (as of 2008) were the ones that have the highest values of connectedness to others. C generated the largest default transmission, 126%, to others. This is consistent with the findings in Diebold and Yilmaz (2014), which is based on the volatility of stock returns among US financial institutes. In addition, the five US banks all generate significant (exceeding 100%) long-term default risk spillover to others when compared with European banks, which evidently implies the transmission of long-term default risk shocks from US financial institutions to their European counterparts.

Furthermore, the strong spillover effects between the connectedness of US and European banks are clearly observed in their 'Net' row. The difference between the total directional connectedness to others and the total directional from others results in the net total directional connectedness to others. C leads the highest net total directional connectedness (44.39%), followed by BAC (40.42%), with other positive effects of US banks. By contrast, the values of net total directional connectedness in European banks are all negative, indicating that the contributions of European banks' shock to other banks' forecast error variance are generally trivial in terms of the long run default risk.

5.1.2. Slope factor. The short-term connectedness is shown in table 5. The highest observed directional connectedness is $C_{JPM \leftarrow HBC} = 33.71\%$, followed by $C_{CBG \leftarrow HBC} = 24.48\%$. As one of the most vulnerable banks during the European debt crisis, HSBC Bank clearly spreads its financial stress to other banks.

It can be observed that Bank of America is weakly affected by the shocks from others, with only $C_{BAC \leftarrow} = 53.35\%$. Although, in the short term, banks in the same region still have relatively large connectedness compared with the cross-region connectedness, little evidence of consistent spillover

effects from US to Europe has been found compared with that in the long-term investigation. In contrast to table 4, HBC creates the strongest risk spillover, followed by DB, both were especially seriously hit during the European debt crisis. Consistently, in terms of 'Net' connectedness measures, HBC leads the group, 74.67%, while the next highest is from DB. Besides the positive values of BAC and C, negative values are generated from other banks. Finally, the total connectedness is 79.72%, which is slightly smaller than the 85.50% that was derived in the long-term total connectedness.

5.1.3. Curvature factor. The middle term directional connectedness is summarized in table 6. The total connectedness, 62.94%, is obviously smaller than the short term and long term connectedness. In terms of the pairwise directional connectedness, the values vary more widely; for example, the highest observed connectedness measure is from HBC to WFC ($C_{WFC \leftarrow HBC} = 33.18\%$) while, in return, the lowest one from WFC to HBC ($C_{HBC \leftarrow WFC} = 0.11\%$) is nearly zero. The spillover effect in this case is obviously 'asymmetric'. In the case of DB, one can find its spillover power in the MT or ST, but not in LT (see table 4). The default tensions emphasized on ST and MT imply that DB may hold more short-term risky loans, which endangers its short-term credit. Interestingly, the 'Net' directional connectednesses are uniformly positive among European banks compared with the consistent negative ones in the United States.

In a nutshell, the three DNS factors and their connectedness convey information w.r.t the default risk at the particular credit horizons. For banks like DB, the potential to experience credit deterioration and subsequently create spillover to others is more likely to happen in the short term. However, the credit condition becomes resilient in the longer term and has restrictive transmission, as shown through a reverse spillover in its level factor.

5.2. Network: dynamics

The DNS model coupled with a topological network can be seen as a means of monitoring systemic vulnerability. In the supervisory purpose, the updated assessment is even more in demand. For this purpose, one studies the dynamics of connectedness in which credit contagion can, therefore, be

Table 6. Static Connectedness: Curvature factor.

	BAC	C	GS	JPM	WFC	DB	CBG	BCS	HBC	UBS	FROM
BAC	22.08	10.89	6.26	6.82	1.19	16.77	9.25	10.36	0.18	16.20	77.92
C	6.40	19.77	7.55	8.62	1.73	16.19	9.16	14.00	2.64	13.94	80.23
GS	10.83	18.24	14.09	9.34	3.53	15.65	7.11	8.35	3.77	9.10	85.91
JPM	14.59	17.26	4.77	31.00	7.07	5.12	3.82	5.56	6.62	4.19	69.00
WFC	5.47	4.60	0.76	8.19	40.90	1.40	0.69	1.69	33.18	3.13	59.10
DB	1.50	5.91	3.35	4.34	0.33	52.11	10.00	11.48	2.57	8.39	47.89
CBG	0.68	2.17	1.16	1.18	0.10	17.45	40.38	10.49	5.35	21.03	59.62
BCS	1.10	5.07	2.60	3.53	0.23	27.17	13.86	29.46	0.12	16.86	70.54
HBC	0.12	2.83	2.80	1.55	0.11	1.71	0.53	2.10	84.39	3.85	15.61
UBS	0.83	5.29	3.53	3.35	0.09	20.97	12.03	17.13	0.35	36.41	63.59
TO	41.52	72.26	32.78	46.92	14.38	122.44	66.45	81.17	54.77	96.69	62.94
NET	-36.40	-7.97	-53.12	-22.09	-44.72	74.55	6.84	10.64	39.17	33.10	-

Note: Data period: 1 January 2008 to 31 December 2015 with daily data.

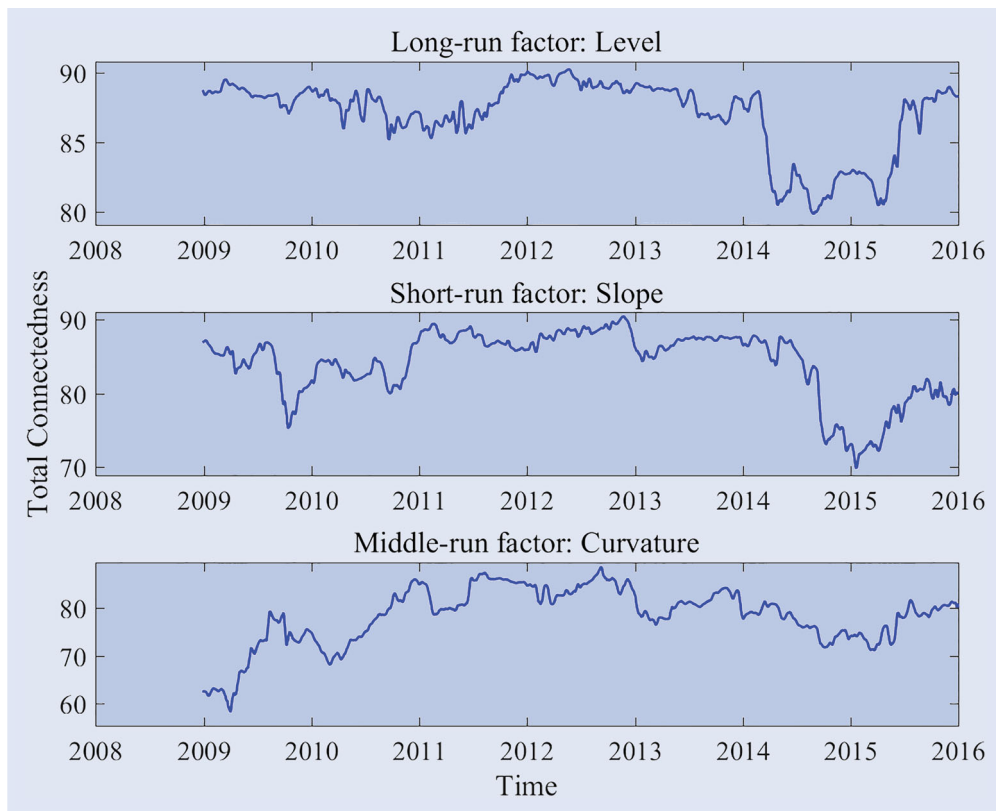


Figure 3. Rolling total connectedness. Note: Data period: 1 January 2009 to 31 December 2015 with daily data. The DNS factors and the corresponding dynamic of networks are computed by one-year rolling window estimation (260 observations). The variance decomposition is performed by the forecast horizon with 12 days. The red line denotes the smoothing line by one-week moving average.

identified in time. Accordingly, they will be asked by the FSB and Basel Committee on Banking Supervision (BCBS) for additional loss absorption capacities to ensure the sufficiency of their common equities in the case of a default.

5.2.1. Time-varying total connectedness. Figure 3 presents time-varying connectedness, C_t , estimated via C from (17) in a one-year (260 observations) rolling window size. It reveals clear default risk cycles. In the LT perspective, the period of 2009 to middle-2014 exhibits a long lasting cycle, coinciding with the outbreak of the European sovereign debt crisis. Despite short spells of recovery at the end of 2014, an

increasing connectedness signals an upcoming systemic risk starting from middle of 2015.

The long continuous cycle in the LT connectedness (upper panel) reveals similar patterns compared with the short term total connectedness by the slope factor (middle panel). The total connectedness declined from the high 90% to around 74% at the end of 2009, followed by a raising connectedness period during European debt crisis starting from 2012 to 2013. In this study, we point out that an increasing total interconnectedness signals a higher likelihood of credit contagion in the system. As the European debt crisis becomes widespread with a systemic danger, the short term total connectedness measures stay at the range of 85%–90% until the end of 2014.

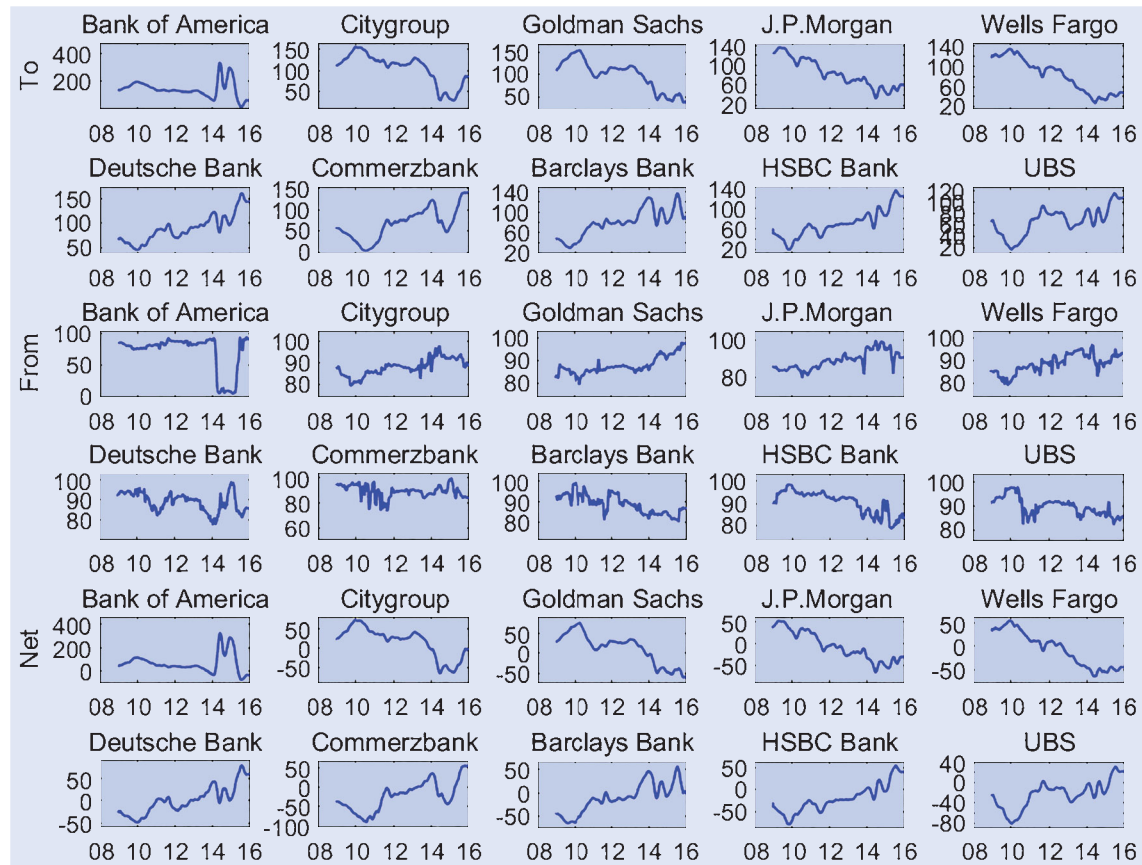


Figure 4. Rolling connectedness: level factors. Note: Data period: 1 January 2009 to 31 December 2015 with daily data. The DNS factors and the corresponding dynamic of networks are computed by one-year rolling window estimation (260 observations). The variance decomposition is performed by the forecast horizon with 12 days.

The banking industry suffered credit tension again as the Chinese stock market went out of control in June 2015. A third of the value of A-shares on the Shanghai Stock Exchange was lost within one month and this caused a large collapse in global financial markets, leading again to concerns about the stability of the global economy. The LT, ST and MT connectedness all reflect systemic fear in 2015.

5.2.2. Time-varying risk contribution. Figures 4, 5, and 6 present the dynamics of individual default risk contribution to total directional connectedness, which are quantified by equation (12) for level factor, slope factor and curvature factor. One can interpret that bank i has a higher marginal risk contribution in the long-term default risk, if the shock of level factor of bank i contributes more on the forecasting errors of level factors of the remaining banks. The upper panel depicts 'To' others, the middle panel displays 'From' others, and the bottom panel collects the 'Net' results. In each panel, the five US banks lie in the first row while other five European banks are in the second row.

Except for the Bank of America, in figure 4, the total directional connectedness 'To' others from US banks appears to be rising from 2008 until 2010; however, they show a downward trend after 2010. Conversely, the directional connectedness 'To' others from the five European banks tends to substantially rise during the period of 2010 to 2015. Overall, the long-term default risk connectedness analysis documents a

declined spillover effect from US banks to others and, in return, the default risk shocks resulting in the forecast error variance are increasingly remarkably transmitted from the European banks. This fact may reflect the supervisory effort of the US banking authority after the outbreak of the US subprime crisis.

Figure 5 reports a similar pattern of short term total directional connectedness dynamics. In the short term default risk connectedness, the 'To' effect from US banks, except for Bank of America, tends to fall roughly from 2012, after a rising trend during 2009 to 2012. In reverse, the short-term 'To' effect from European banks declined until 2012, subsequently followed by a rising trend. In summary, the 'Net' effect of US banks declines from 2012 after a consistent rise, while the 'Net' effect of European banks shows a reverse pattern. In other words, being analogous to the long term directional connectedness, in the short term, the shocks arising from European banks tend to be dominantly transmitted to others, especially after 2012, which coincides with the burst of the European sovereign debt crisis during 2012 to 2013.

In the three figures, one can see that Bank of America during the period of 2014 to 2015 creates a very promising 'Net' effect, regardless of default horizons. Obviously, BAC is an overwhelming default shock transmitter and it needs to be asked for an additional loss buffer. Due to the 2008 acquisition of Countrywide Finance, which was a high-flying mortgage company that fuelled many of the excesses of the housing boom, BAC took huge losses on

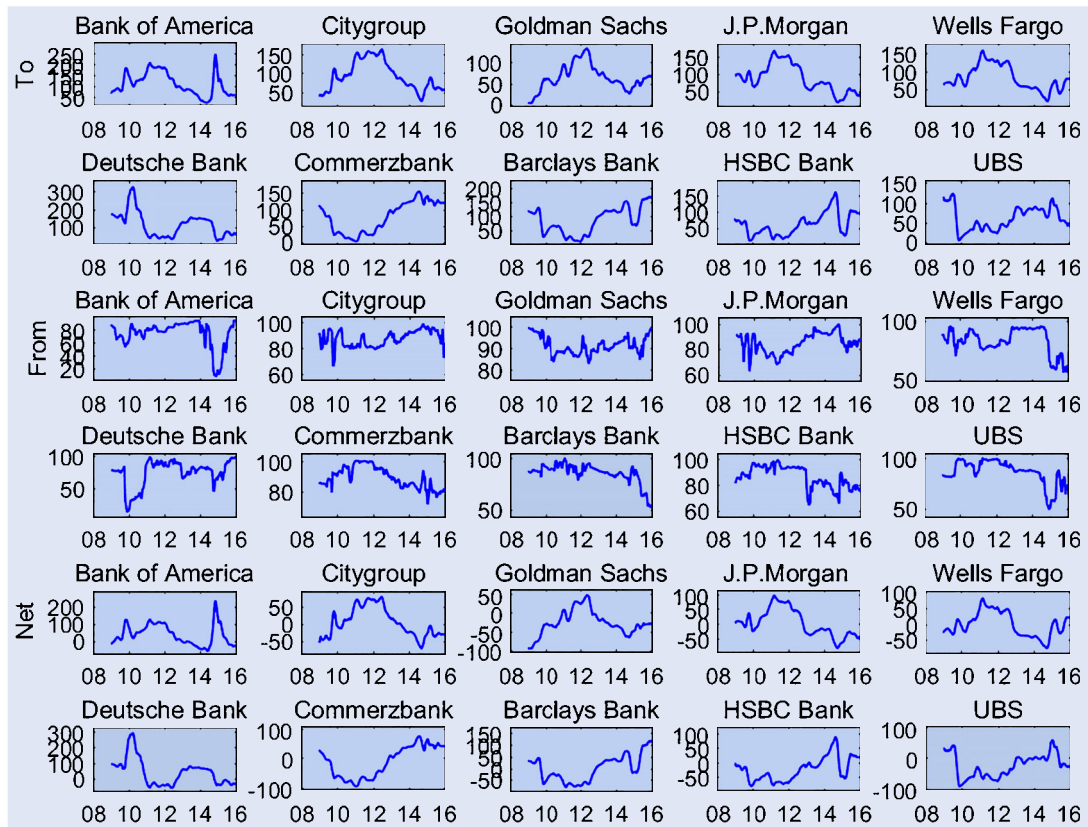


Figure 5. Rolling connectedness: Slope. Note: Data period: 1 January 2009 to 31 December 2015 with daily data. The DNS factors and the corresponding dynamic of networks are computed by a one-year rolling window estimation (260 observations). The variance decomposition is performed by the forecast horizon with 12 days.

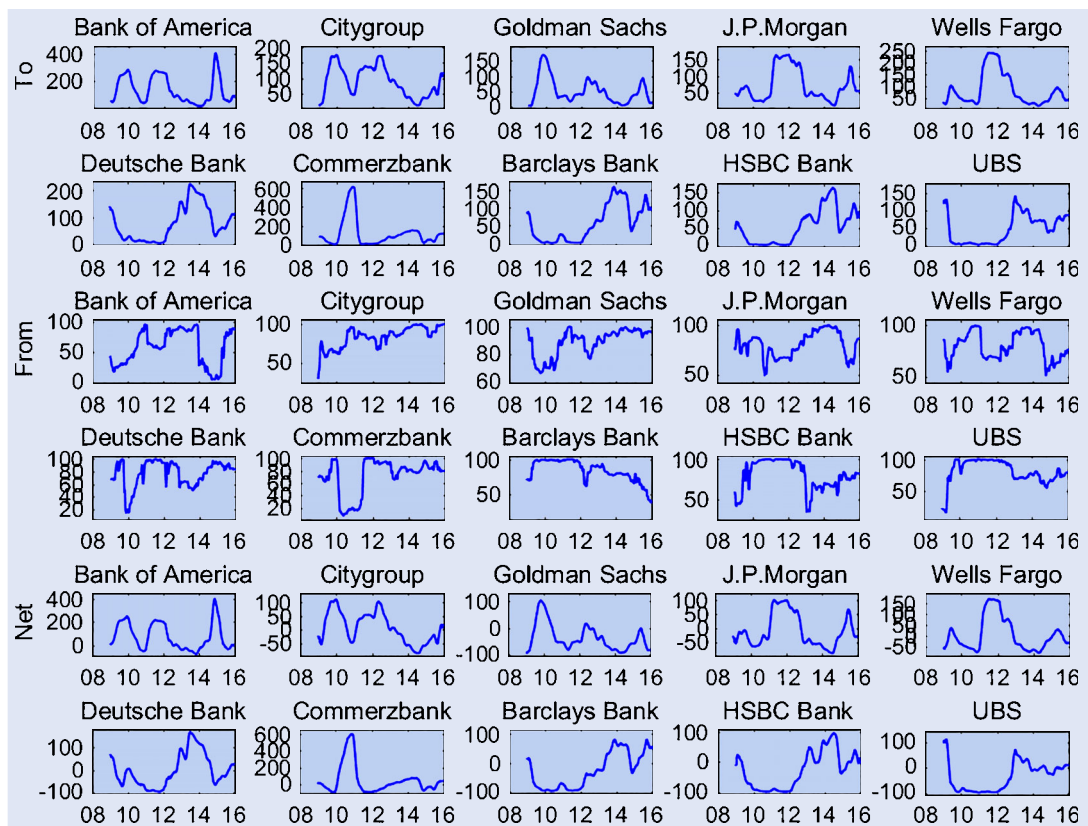


Figure 6. Rolling connectedness: curvature factors. Note: Data period: 1 January 2009 to 31 December 2015 with daily data. The DNS factors and the corresponding dynamic of networks are computed by one-year rolling window estimation (260 observations). The variance decomposition is performed by the forecast horizon with 12 days.

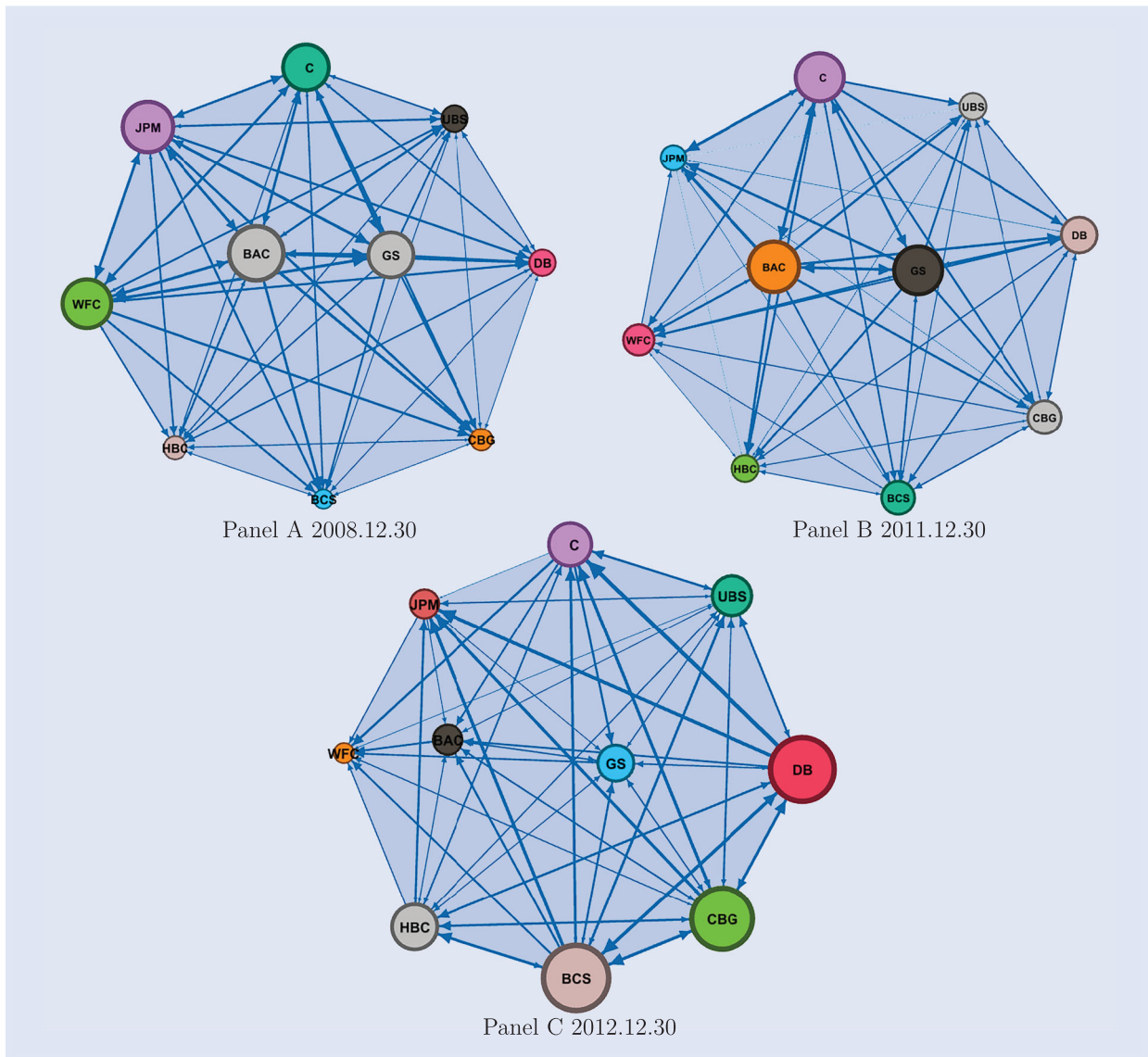


Figure 7. Level pairwise directional connectedness network. Notes: Node size and node color indicate $C_{\bullet \leftarrow i}$ of the bank's level factor. Edge thickness indicates the pairwise directional connectedness. Edge color does not vary with edge weight.

distressed Countrywide mortgages. In March 2014, the bank announced an unexpected \$6 billion in mortgage related legal expenses. Additionally, a further payment of more than \$16 billion in penalties to settle claims was also reported, which was formally announced in August. After these problems, in April, BAC disclosed a significant accounting error of \$4 billion capital loss, which has been undetected for several years. The capital error weighed heavily on the bank's shares, which fell by more than 6% on that trading day, wiping out \$10 billion in market value, which was far more than the actual losses. This indicated the collapse of trust from investors, which simultaneously triggered large shocks to other financial institutions.

5.2.3. Graphical representation. The network dynamics can be displayed graphically, where the node size and node color are designed to capture the 'To' effect. Meanwhile, directional edge thickness indicates the strength of pairwise directional connectedness, while edge color does not vary

with edge weight. Consider the three snapshots for 2008, 2011 and 2012 in figure 7.

In 2008, the node size of US banks is apparently larger than that of EU banks, indicating a credit spillover from the US. Besides, the thickness of the edge implies that the spillover effects are not only closely intertwined among US banks but are also transmitted to EU banks (e.g. from GS to DB or CBG, from BAC to CBG). This evidence had not yet fade away at the end of 2011 where BAC, JPM, and C are still sizable. However, the node sizes of European banks at the end of 2011 tend to enlarge compared with that in 2008, owing to the continuous negative impact from the Greece, Irish, and Portugal debt crisis. When the European sovereign debt crisis was at its peak at the end of 2012, the European banks' debts turned out to be enormously large, especially for those banks with a large market capitalization (such as DB, CBG, and BCS). Furthermore, the edge thickness indicates that the default risk shocks are mutually conducted not only among European banks but also substantially outflowed from EU banks to US Banks.

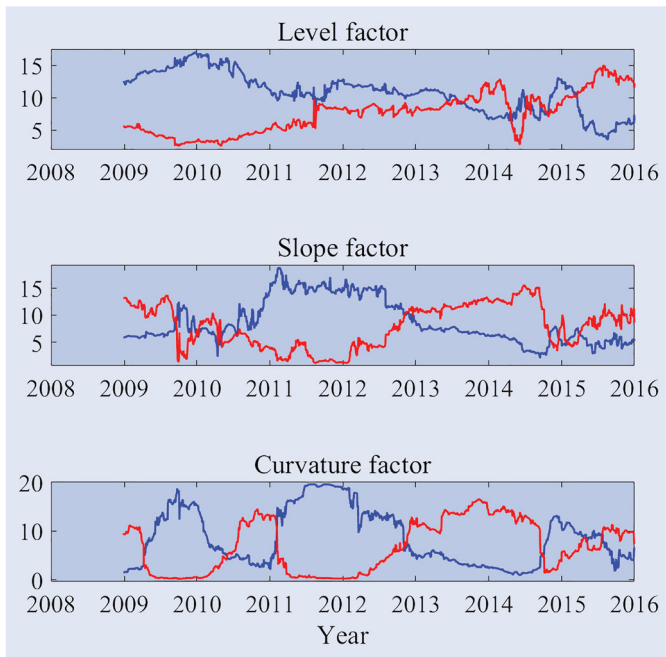


Figure 8. Dynamics of connectedness across the United States and Europe. The blue line presents the measures from US banks to European banks, the red line presents the measures from European banks to US banks.

5.3. The network between US and European banks

In order to have a clear picture on the cross-region spillover effects, we look at two groups - the US versus EU banks - and analyse the dynamics of their default risk transmission.

Figure 8 reveals that the transmission of overall default risk shocks, in the long term, has been increasing from European banks to US banks since 2010, while that of the short term has a crossing point in 2013 but falls down around 2014, seemingly followed by a new cycle of climbing periods, which is consistent with figures 3 and 5. The default shocks still originate from Europe even 5–6 years after the outburst of the European sovereign debt crisis.

5.4. The drivers of default connectedness

Having the dynamics of default connectedness in a system, we dive deeper to investigate the determinants of these dynamics which may give the insights into the channel of default contagion. In our analysis, the total connectedness can be viewed as an overall measure of default risk in the system and also as an indicator of credit contagion, in which a high value implies a widespread default risk. To understand the evolution of systemic default risk in the banking industry, and to control it further, the policy makers may rely on the model-implied indicators to monitor the frailty of default in a system.

Hence, in this part, we take the total connectedness derived from level factor, slope factor, and curvature factor in section 5.2.1 as a measure of long term, short term, and middle term systemic default risk respectively, and opt for a vector of state variables to analyse what drives the systematic default risk. We estimate the following regression on the

basis of daily data for the determinants of the connectedness of credit curves:

$$C_{\omega,t} = \alpha_{\omega} + \beta_{\omega}^{\top} M_{t-1} + \varepsilon_{\omega,t}, \quad \varepsilon_{\omega,t} \sim N(0, \sigma^2) \quad (22)$$

where $C_{\omega,t}$ denotes total connectedness of level factor, slope factor, and curvature factor at time t respectively, $\omega = \{l, s, c\}$. M_{t-1} denotes state variables at time $t - 1$.

Adrian and Brunnermeier (2016) propose to use the following macro published variables, e.g. (1) The change in the three-month yield; (2) The change in the slope of the yield curve, measured by the spread between the composite long-term bond yield and the three-month bill rate; (3) A short-term TED spread, defined as the difference between the three-month LIBOR rate and the three-month secondary market Treasury bill rate - this spread measures short-term funding liquidity risk; (4) The change in the credit spread between Moody’s Baa-rated bonds and the ten-year Treasury rate; (5) The daily market return computed from the S&P500; (6) The daily real estate sector return in excess of the market financial sector return; (7) VIX. In addition we employ common principal components (CPC), which is the average variance explained by the first principle component through the common principle component approach (CPCA), see Flury (1984), Fengler *et al.* (2003) and Chen and Härde (2015). The CPC factor is used here to capture a common factor that may not be directly observed.

Estimating PCs simultaneously in different groups (banks) can result in a joint dimension reduction transformation, as well as yielding a joint eigenstructure across groups (banks). The basic assumption of CPCA is that the space spanned by the eigenvectors is identical across several groups (banks), whereas the variances associated with the components are allowed to vary. The CPCA essentially tests whether the principal components for different banks are the same across different maturities. More formally, for the covariance matrix across K different maturities of bank i , Ψ_i , the hypothesis of CPCA is:

$$H_{CPC} : \Psi_i = \Gamma \Lambda_i \Gamma^{\top}, \quad i = 1, \dots, N \quad (23)$$

where Ψ_i is $K \times K$ positive definite covariance matrix; $\Gamma = (\gamma_1, \dots, \gamma_K)$ is an $K \times K$ orthogonal eigenvector matrix, which is identical for N banks; and, $\Lambda_i = \text{diag}(\lambda_{i1}, \dots, \lambda_{iK})$ is an diagonal eigenvalues matrix of bank i . For more details on the estimation, see Appendix A.2. Through averaging the variance explained by first principal component of each bank, we estimate the CPC variance explained variable using a fixed rolling window of 260 observations, which is in line with the similar procedure in section 5.2.1.

Table 7 provides summary statistics of the total connectedness and state variables. In line with the previous results in section 5.1, the mean value of total connectedness of level factor is larger than that of the slope factor, followed by that of the curvature factor. The negative skewness values, as well as kurtosis values of nearly 3, indicate that total connectedness measures seem to follow right-skewed asymmetric normal distribution. As for the CPC first factor variance explained variable, the range is roughly 13%–92%, which is reasonable with a large deviation of 15%.

Table 7. Summary of the variables.

	Mean	S.t.d.	Skew	Kurt	Min	Max
Total connectedness: Level	86.99	2.63	-1.26	3.47	79.61	90.26
Total connectedness: Slope	84.07	4.60	-1.15	3.49	69.52	90.44
Total connectedness: Curvature	78.37	6.09	-0.88	3.66	55.39	89.18
Three month yield change	0.01	1.18	0.22	7.84	-7.00	6.00
Term spread change	0.00	5.76	-0.15	6.24	-48.00	25.00
TED spread	29.35	19.21	2.85	11.67	8.76	133.50
Credit spread change	-0.15	2.78	0.34	11.46	-14.00	28.00
Market return	0.05	1.11	-0.27	7.69	-6.90	6.84
Real estate excess return	0.00	0.93	0.01	10.92	-6.00	6.98
VIX	20.17	8.04	1.58	5.27	10.32	56.65
CPC first factor variance explained	50.77	15.06	0.26	2.69	13.35	92.64

Note: Data period: 1 January 2009 to 31 December 2015 with daily data. The change, return data, and CPC factor variance explained are in percentage. We list the standard deviation (S.t.d.), skewness (Skew), kurtosis (Kurt) and so on.

Table 8. Determinants of Total connectedness.

	Level		Slope		Curvature	
	NW	HH	NW	HH	NW	HH
Three month yield change	0.010 (0.682)	0.010 (0.590)	-0.007 (-0.490)	-0.007 (-0.385)	-0.003 (-0.197)	-0.003 (-0.184)
Term spread change	-0.016 (-0.888)	-0.016 (-0.746)	-0.009 (-0.453)	-0.009 (-0.406)	0.001 (0.025)	0.001 (0.026)
TED spread	0.125*** (2.483)	0.125*** (1.906)	0.153*** (3.286)	0.153*** (2.289)	-0.363*** (-4.473)	-0.363*** (-5.811)
Credit spread change	0.056*** (3.035)	0.056*** (2.107)	0.044*** (2.108)	0.044*** (1.631)	0.042* (1.543)	0.042* (1.644)
Market return	-0.046*** (-2.550)	-0.046*** (-2.082)	-0.018 (-0.953)	-0.018 (-0.805)	-0.022 (-0.954)	-0.022 (-1.044)
Real estate excess return	-0.005 (-0.260)	-0.005 (-0.213)	-0.021 (-1.061)	-0.021 (-0.950)	-0.023 (-1.078)	-0.023 (-1.110)
VIX	0.298*** (4.886)	0.298*** (4.545)	-0.010 (-0.170)	-0.010 (-0.151)	-0.120* (-1.428)	-0.120*** (-1.924)
CPC factor variance explained	0.303*** (7.996)	0.303*** (8.543)	0.434*** (11.793)	0.434*** (11.994)	0.260*** (9.119)	0.260*** (7.677)
Adjusted R^2 (%)	23.04	23.04	19.91	19.91	29.79	29.79

Note: ***, **, * denotes the significance at the level of 1%, 5% and 10% respectively. 'NW' presents that the t -statistics displayed in parentheses are calculated by Newey-West standard errors allowing for up to five periods of autocorrelation. 'HH' represent the t -statistics displayed in parentheses are calculated by Hansen and Hodrick standard errors with five periods of lag.

After standardizing all the variables in table 7, we obtain the estimated parameters through (22) in table 8. In column 'NW' and column 'HH', the values in parentheses under the corresponding estimated parameters, present t -statistics based on Newey-West standard errors (Newey and West 1987), and Hansen-Hodrick standard errors (Hansen and Hodrick 1980) respectively, both with five periods of lag. In addition, the last line reports the adjusted R^2 value of the model.

In table 8, higher VIX, higher TED spread, higher credit spread, and lower market return result in high long term default risk. Consistent with Pan and Singleton (2008), we find that VIX, which measures investor appetite for risk, significantly triggers a credit comovement or contagion. Pan and Singleton (2008) argue that VIX is a key factor in an investor's appetite for global credit risk in credit markets. They find that there is a higher correlation between VIX and the first principle component (PC) of CDS spreads, we, however, link VIX to total connectedness. Total connectedness, as an aggregated measure, is quantified by a dynamic system, whereas the PC is more or less commonality and lacks of dynamics. The short-term connectedness is driven by the

TED spread and credit spread, while the medium-term connectedness is sensitive to the TED spread and VIX. The CPC factor is used to capture the latent common factor and is significant across the LT, MT and ST connectedness, implying that the existing latent determinants need to be further discovered. Overall, the average significance of the conditioning variables reported in table 8 show that the state variables do indeed contribute to the variation of total connectedness, particularly in the long-term. In summary, the channels of credit contagion vary with credit horizons. In the long-term credit contagion is more likely to experience through the credit market (TED and credit spread) and a change of investor risk attitude, while in the short-term contagion the conditions of credit market are exclusively relevant.

5.5. Out-of-sample forecasts

To evaluate the informativeness of the predefined network connectedness, it is of interest to compare the forecasting performance between the network-based DNS model and the

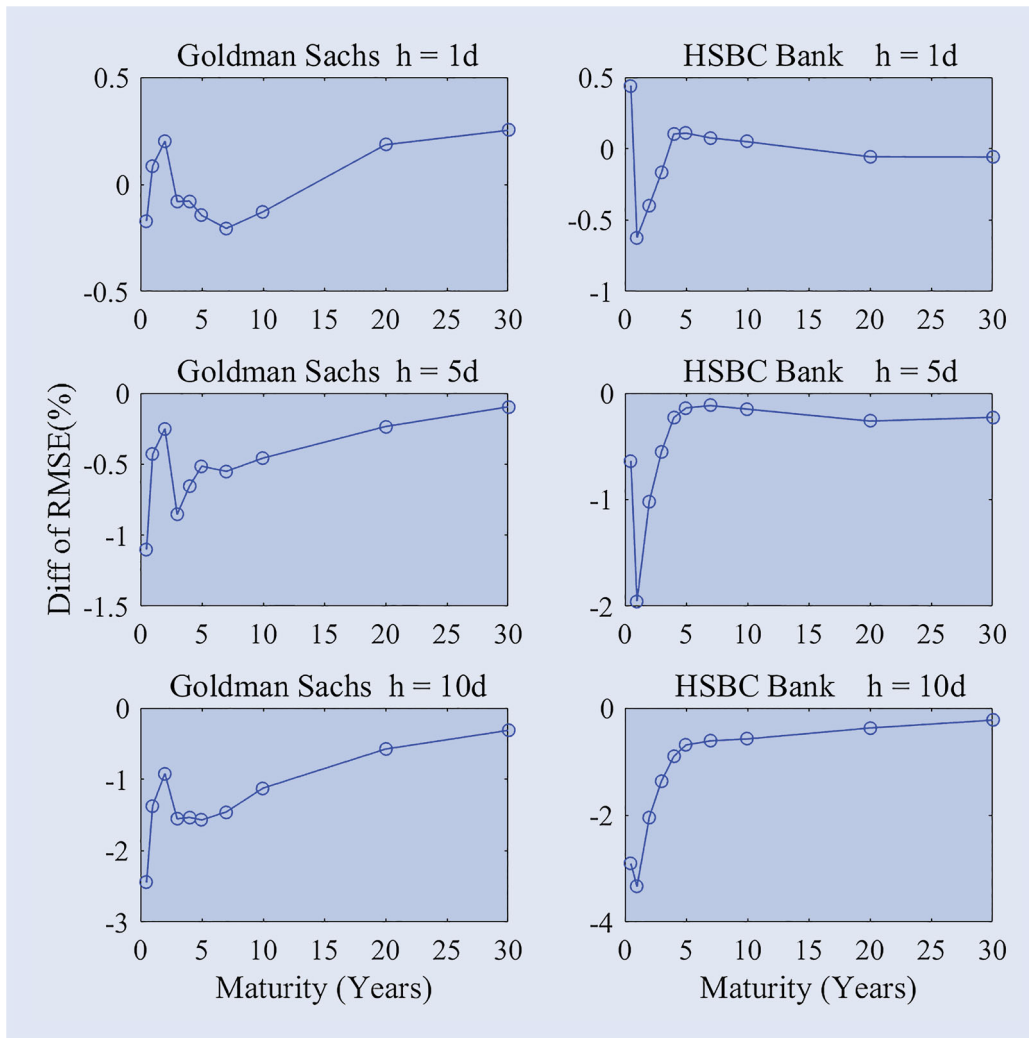


Figure 9. Difference of RMSE in percentage of CDS spreads forecast between network DNS model and DNS model, for horizon $h = \{1\text{day}, 5\text{day}, 10\text{day}\}$. The left panel lists the values of Goldman Sachs, the right panel for HSBC bank. The points correspond to different maturities. The forecast period is 1 January 2011 to 31 December 2013.

conventional DNS model. Using (3), one predicts:

$$\hat{y}_{i,t+h|t}(\tau) = \hat{\beta}_{i1,t+h|t} + \hat{\beta}_{i2,t+h|t} \left\{ \frac{1 - \exp(-\delta\tau)}{\delta\tau} \right\} + \hat{\beta}_{i3,t+h|t} \left\{ \frac{1 - \exp(-\delta\tau)}{\delta\tau} - \exp(-\delta\tau) \right\} \quad (24)$$

where $\hat{\beta}_{is,t}$, $s = 1, 2, 3$ denotes \hat{l}_{it} , \hat{s}_{it} , \hat{c}_{it} respectively, and can be estimated through (5) and (6).

(a) Network DNS model

To implement the forecast model comparison, the transition equation in a multivariate factor framework to undertake their interaction, named as DNS-VAR(1), is

$$\hat{\beta}_{is,t+h|t} = \hat{\gamma}_{is}^0 + \hat{\gamma}_{is} \hat{\beta}_{is,t} + \hat{\phi}_{js} \hat{\beta}_{js,t}, \quad j \neq i \quad (25)$$

where $\hat{\beta}_{js,t}$ is the latent factors from other banks j , which are estimated in the initial step, such as using (5) and (6). The parameter $\hat{\gamma}_{is}^0$, $\hat{\gamma}_{is}$ and $\hat{\phi}_{js}$ are estimated by the Kalman filter method. The technical details are specified in an appendix.

(b) DNS model

The autoregressive process of transition equation without the influence from the latent factors of other banks, named as DNS-AR(1), is

$$\hat{\beta}_{is,t+h|t} = \hat{\gamma}_{is}^0 + \hat{\gamma}_{is} \hat{\beta}_{is,t} \quad (26)$$

(c) VAR-elasticnet model

In addition, following Bostanci and Yilmaz (2015), we directly forecast CDS spreads of all the firms using elastic net estimation for each maturity respectively, named as VAR-elasticnet model,

$$\hat{y}_{t+h|t}(\tau) = \sum_{k=1}^p \hat{\theta} y_{t-k+1}(\tau)$$

$$y_t(\tau) = (y_{1t}(\tau), y_{2t}(\tau), \dots, y_{N_t}(\tau))^T$$

$$\hat{\theta} = \arg \max_{\theta} \left\{ \sum_{t=2}^T \left\{ y_t(\tau) - \sum_i \sum_{k=1}^p \theta_{i,k} y_{i,t-k}(\tau) \right\}^2 + \lambda \sum_{i=1}^K \sum_{k=1}^p \{ \alpha |\theta_{i,k}| + (1 - \alpha) \theta_{i,k}^2 \} \right\} \quad (27)$$

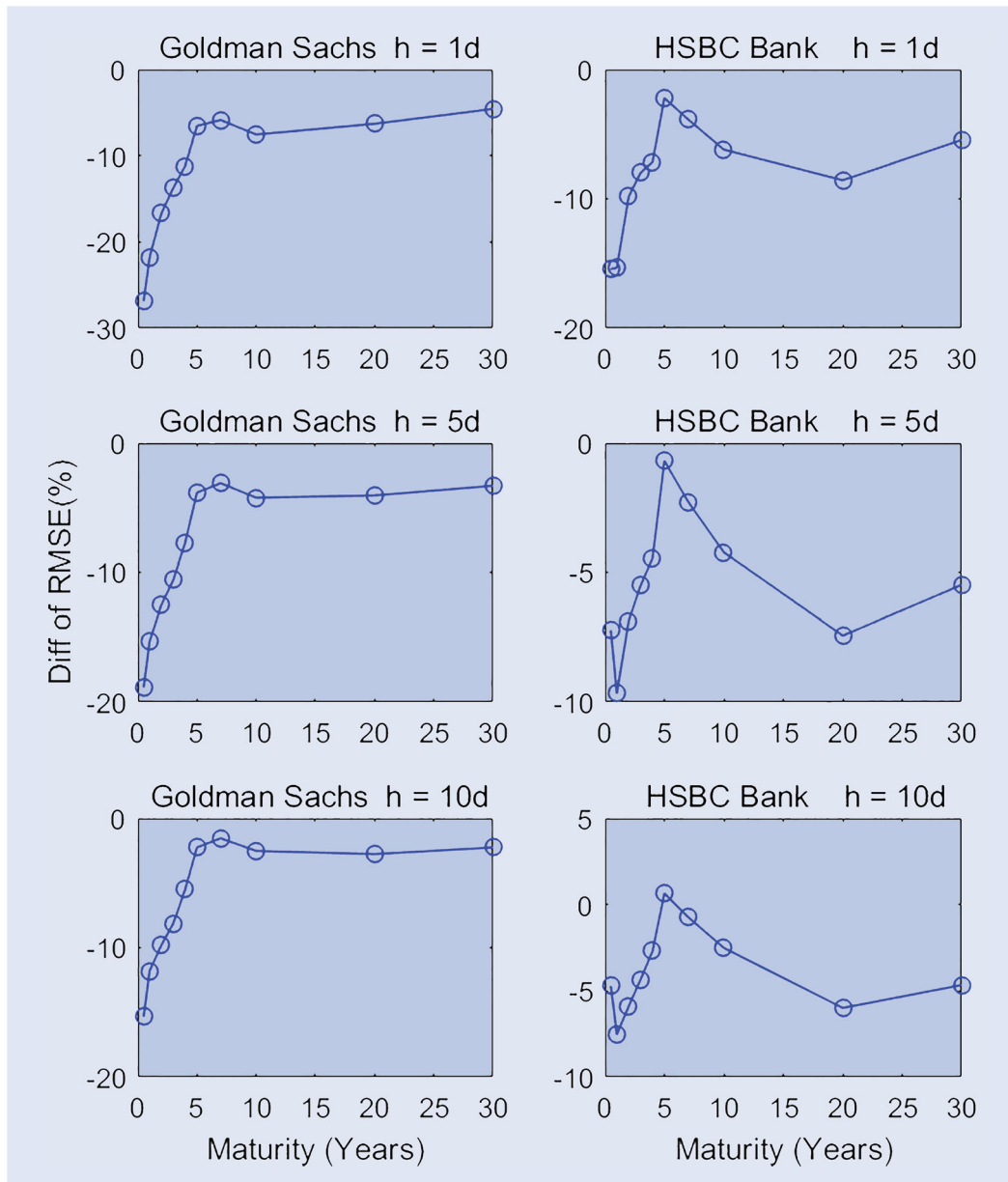


Figure 10. Difference of RMSE in percentage of CDS spreads forecast between network DNS model and VAR-elasticnet model, for horizon $h = \{1\text{day}, 5\text{day}, 10\text{day}\}$. The left panel lists the values of Goldman Sachs, the right panel for HSBC bank. The points correspond to different maturities. The forecast period is 1 January 2011 to 31 December 2013.

where p is the lag order in the VAR model. Following the literature, we choose $p = 2$.

We present the forecasting performance of model (a) in comparison with the alternatives model (b) and model (c). The forecast horizons are selected as 1 day, 5 days, and 10 days. We report the difference of root mean squared errors (RMSE) when comparing the network DNS model with conventional DNS model and VAR-elasticnet model. When taking the European crisis period, from 2011 to 2013, as the out-of-sample forecast period, we present the results of GS and HSBC bank as examples. Each point in figure 9 denotes the difference between the RMSE of network DNS model and that of DNS model while figure 10 represents the difference of RMSE between network DNS model and VAR-elasticnet model at each maturity period. The resulting negative values in the difference in both figures indicate the

superiority of the model with network perspective. One can observe the negative values in the difference of RMSE, which are homogeneously distributed in the majority of maturities, especially as the forecast horizons increase. This reveals that RMSE of forecasting CDS spreads in network DNS model is smaller than that of DNS model as well as VAR-elasticnet model, the network DNS model performs even better as the forecast horizon increases.

By overall summarizing the performance of European banks and US banks, we present the average value of RMSE difference of network DNS model with the alternative models in table 9. This reveals a similar pattern to what has been discovered in figures 9 and 10. More interestingly, the more frequent negative values shown in the EU group implies that the network information advantage in forecasting CDS curve seems to have more support in the EU area. To predict

Table 9. US and EU: average value of RMSE difference (%).

	<i>h</i>	6M	1Y	2Y	3Y	4Y	5Y	7Y	10Y	20Y	30Y
(a) Network DNS model versus DNS model											
US	1d	0.156	0.029	0.094	-0.031	-0.102	-0.076	-0.029	-0.026	0.158	0.198
	5d	-0.132	-0.235	-0.116	-0.356	-0.318	-0.108	0.140	-0.037	-0.031	0.092
	10d	-0.609	-0.637	-0.209	-0.480	-0.597	-0.566	-0.161	-0.290	-0.121	0.111
EU	1d	0.264	-0.162	-0.110	-0.027	0.091	0.022	-0.018	-0.041	0.048	0.081
	5d	-0.486	-0.973	-0.516	-0.318	-0.058	-0.098	-0.237	-0.354	-0.174	-0.011
	10d	-1.580	-1.388	-0.685	-0.446	-0.388	-0.524	-0.700	-0.804	-0.309	-0.014
(b) Network DNS model versus VAR-elasticnet model											
US	1d	-22.232	-19.466	-14.295	-11.950	-9.449	-4.581	-3.362	-5.238	-4.571	-2.492
	5d	-14.184	-13.604	-10.836	-9.138	-6.564	-2.607	-1.668	-3.152	-3.259	-2.174
	10d	-11.014	-10.747	-8.657	-7.262	-4.941	-1.617	-0.887	-2.178	-2.429	-1.633
EU	1d	-16.757	-14.500	-10.727	-8.799	-7.658	-3.258	-3.449	-5.357	-6.688	-4.479
	5d	-10.020	-9.878	-7.625	-6.145	-4.636	-1.365	-1.686	-3.240	-5.222	-3.998
	10d	-7.805	-7.842	-5.856	-4.642	-3.261	-0.382	-0.743	-2.114	-4.142	-3.296

Note: The average RMSE difference of CDS spreads forecast of network DNS model in comparison with DNS model in panel (a) and VAR-elasticnet model in panel (b). The values of US and EU banks are averaged in total. All the values are in percentage. The forecast period is 1 January 2011 to 31 December 2013.

Table 10. Forecast comparison between network DNS model and DNS model: DM test.

<i>h</i>	6M	1Y	2Y	3Y	4Y	5Y	7Y	10Y	20Y	30Y
1d										
BAC	1.036	2.356**	0.703	-2.514**	-2.105**	-2.954***	-2.389**	-1.286	3.402***	5.344***
C	3.259***	-2.933***	-1.270	-3.529***	0.156	3.543***	4.836***	4.593***	-0.968	-1.831*
GS	-3.290***	1.906*	4.159***	-2.043**	-2.034**	-3.774***	-6.257***	-3.802***	5.118***	7.390***
JPM	3.202***	-2.067**	-4.887***	-7.285***	0.754	4.240***	4.331***	1.772*	1.666*	4.398***
WFC	1.437	3.347***	7.637***	6.896***	-4.741***	-9.662***	-5.940***	-4.403***	10.599***	15.143***
DB	2.859***	-2.218**	-4.947***	-6.677***	2.902***	4.732***	4.827***	2.006**	-1.064	0.292
CBG	-2.258**	4.540***	4.045***	3.719***	0.560	-4.488***	-4.772***	-1.648*	4.138***	5.802***
BCS	7.258***	-6.434***	-6.738***	-0.524	8.014***	7.809***	2.624***	-1.156	-0.406	2.002**
HBC	3.041***	-6.318***	-5.654***	-2.964***	2.325**	2.404**	2.195**	1.939*	-3.762***	-4.773***
UBS	-0.533	2.535**	3.148***	2.773***	-3.367***	-4.665***	-6.062***	-7.829***	7.208***	9.688***
5d										
BAC	0.027	0.242	-0.763	-1.618	-0.896	-0.911	-0.362	-0.615	-0.479	-0.172
C	0.178	-2.284**	-2.459**	-2.490**	0.840	3.394***	4.498***	3.255***	-1.285	-2.572**
GS	-2.329**	-1.035	-0.656	-2.747***	-2.316**	-1.811*	-2.223**	-1.913*	-0.996	-0.412
JPM	0.348	-0.660	-1.559	-2.667***	-1.047	0.560	1.347	-0.155	-0.599	0.487
WFC	0.124	0.863	2.756***	2.576***	-1.383	-3.892***	-2.022**	-1.570	2.431**	4.420***
DB	-0.572	-1.703*	-2.158**	-2.799***	-0.073	0.421	0.415	-1.344	-2.286**	-1.383
CBG	-2.955***	-1.140	-1.377	-1.104	-1.050	-1.748*	-2.187**	-1.625	-0.613	-0.091
BCS	0.976	-2.924***	-2.126**	-0.141	1.372	1.414	-0.485	-1.840*	-0.852	0.335
HBC	-0.993	-3.938***	-2.475**	-1.690*	-0.944	-0.530	-0.559	-0.930	-1.994**	-2.198**
UBS	-0.891	0.571	1.185	0.847	-0.731	-1.839*	-2.878***	-2.929***	1.124	2.242**
10d										
BAC	-0.219	-0.130	-0.349	-0.778	-0.779	-0.958	-0.503	-0.583	-0.438	-0.212
C	-0.760	-2.301**	-1.992**	-1.955*	-0.107	1.855*	2.939***	2.075**	-0.405	-1.175
GS	-1.896*	-1.153	-0.888	-1.851*	-2.224**	-2.579***	-2.751***	-2.149**	-1.096	-0.605
JPM	0.031	-0.726	-1.051	-1.710*	-0.890	-0.301	0.041	-0.807	-0.771	-0.034
WFC	0.111	0.588	2.065**	2.056**	-0.140	-2.124**	-1.075	-0.871	1.405	2.838***
DB	-1.265	-1.098	-0.891	-1.489	-0.981	-0.972	-0.782	-2.014**	-1.792*	-0.995
CBG	-2.696***	-1.534	-1.379	-1.079	-0.985	-1.220	-1.464	-1.299	-0.787	-0.449
BCS	-0.295	-1.843*	-1.256	-0.170	-0.154	-0.789	-2.253**	-3.134***	-0.901	0.334
HBC	-2.296**	-3.564***	-2.595***	-2.037**	-1.758*	-1.426	-1.555	-1.750*	-1.405	-1.086
UBS	-0.964	0.029	0.614	0.609	-0.127	-1.086	-2.235**	-2.289**	0.780	1.726*

Note: This table gives the t-statistics of Diebold-Mariano test, that is, $H_0: \mu = 0$ in the regression $e_{t,net}^2 - e_t^2 = \mu + \varepsilon_t$ where $e_{t,net}$ and e_t denote the forecast error of network DNS model and DNS model respectively. The test is modified with robust Newey-West variances for heteroscedasticity and autocorrelation with the lags equal to the forecast horizon. * denotes a significance level of 10%, ** denotes a significance level of 5%, *** denotes a significance level of 1%. The forecast period is 1 January 2011 to 31 December 2013.

the CDS curve in EU banks, one has to opt for a network-perspective model. The prediction can then be used to decide the timing of entering a CDS contract to lock up the hedge cost in terms of default from the buyer side. For the seller

side, they are able to avoid underestimating CDS prices after incorporating default spillover through their network.

Tables 10 and 11 report the Diebold-Mariano test statistics (Diebold and Mariano 1995) to compare the network

Table 11. Forecast comparison between network DNS model and VAR-elasticnet model: DM test.

<i>h</i>	6M	1Y	2Y	3Y	4Y	5Y	7Y	10Y	20Y	30Y
1d										
BAC	1.895*	0.343	-2.398**	-4.424***	-7.715***	-4.766***	-3.274***	-5.760***	-5.278***	-0.514
C	4.581***	2.620***	-0.375	-0.716	-1.709*	1.763*	6.773***	5.278***	7.130***	11.808***
GS	3.356***	2.670***	-0.138	-1.435	-3.694***	0.809	1.154	-1.540	-1.011	3.034***
JPM	7.192***	7.370***	5.551***	5.020***	4.838***	9.188***	9.742***	5.142***	2.738***	6.167***
WFC	0.942	0.847	-0.920	-1.397	-1.160	5.361***	10.414***	9.758***	11.901***	14.973***
DB	4.966***	5.279***	4.870***	3.674***	2.769***	4.453***	4.364***	0.524	-1.801*	1.854*
CBG	7.986***	8.064***	5.465***	2.576***	-2.453**	0.231	-1.177	-6.083***	-7.597***	-2.815***
BCS	5.978***	6.884***	6.772***	6.468***	4.538***	5.837***	4.675***	-1.096	-4.122***	-0.884
HBC	4.471***	3.451***	4.580***	4.581***	2.821***	4.410***	1.603	-4.729***	-7.382***	-3.088***
UBS	-0.773	-1.977**	-3.097***	-2.222**	-0.750	5.250***	8.979***	6.609***	6.828***	10.772***
5d										
BAC	-0.751	-1.613	-3.364***	-4.761***	-6.574***	-5.275***	-4.167***	-5.646***	-5.221***	-2.942***
C	2.250**	0.462	-1.910*	-2.416**	-3.697***	-1.490	1.503	0.314	1.101	4.804***
GS	1.283	0.679	-1.763*	-2.712***	-4.124***	-1.321	-0.498	-2.180**	-2.225**	0.002
JPM	2.908***	2.859***	1.770**	1.325	1.111	4.820***	5.383***	2.527**	0.984	3.192***
WFC	-0.746	-0.918	-2.324**	-2.960***	-3.493***	0.300	3.167***	3.176***	5.018***	7.375***
DB	0.581	0.579	0.832	-0.111	-1.010	0.312	0.440	-1.925*	-2.673***	-0.651
CBG	3.332***	2.972***	1.318	-0.373	-2.992***	-1.691*	-1.969**	-4.153***	-4.814***	-2.419**
BCS	2.089**	2.662***	2.236**	1.474	-0.701	0.126	-0.630	-4.841***	-5.302***	-3.333***
HBC	0.560	-0.742	0.709	0.592	-0.726	0.898	-0.599	-4.421***	-5.454***	-2.877***
UBS	-3.327***	-4.038***	-4.787***	-4.364***	-3.564***	0.556	3.133***	1.430	0.818	4.002***
10d										
BAC	-1.690*	-2.235**	-3.460***	-4.252***	-5.322***	-4.700***	-3.908***	-4.748***	-4.377***	-3.009***
C	1.024	-0.830	-3.187***	-3.688***	-4.590***	-2.675***	-0.772	-1.622	-1.409	2.008**
GS	0.193	-0.603	-2.796***	-3.193***	-4.068***	-1.853*	-0.962	-2.214**	-2.345**	-0.772
JPM	1.330	1.352	0.500	0.078	-0.041	3.434***	4.236***	1.938*	0.424	2.199**
WFC	-1.340	-1.636	-2.945***	-3.600***	-4.119***	-1.509	0.405	0.610	2.423**	4.689***
DB	-1.169	-1.606	-1.260	-1.840*	-2.382**	-0.880	-0.817	-2.696***	-2.774***	-1.294
CBG	1.194	0.679	-0.154	-1.379	-3.379***	-2.608***	-2.597***	-4.284***	-4.675***	-2.781***
BCS	0.298	0.821	0.298	-0.586	-2.222**	-1.518	-1.580	-3.977***	-4.024***	-2.790***
HBC	-1.084	-2.461**	-1.238	-1.083	-1.903*	-0.291	-1.141	-4.302***	-4.627***	-2.533**
UBS	-4.071***	-4.768***	-5.329***	-5.551***	-5.047***	-1.596	0.514	-1.253	-2.552**	0.603

Note: This table gives the *t*-statistics of Diebold-Mariano test, that is, $H_0 : \mu = 0$ in the regression $e_{t,net}^2 - e_t^2 = \mu + \varepsilon_t$ where $e_{t,net}$ and e_t denote the forecast error of network DNS model and VAR-elasticnet model respectively. The test is modified with robust Newey-West variances for heteroscedasticity and autocorrelation with the lags equal to the forecast horizon. * denotes a significance level of 10%, ** denotes a significance level of 5%, *** denotes a significance level of 1%. The forecast period is 1 January 2011 to 31 December 2013.

DNS model versus DNS model and VAR-elasticnet model forecasts. In the Diebold-Mariano (DM) test, we conduct a pairwise test of the equality of the mean squared forecast errors by analyzing the difference between the squared forecast errors of the network DNS model and the DNS model, $e_{t,net}^2 - e_t^2 = \mu + \varepsilon_t$. The null hypothesis of equal performance is that $H_0 : \mu = 0$. We focus on the *t*-statistics of parameter μ , denoted as DM *t*-stat, which supports the network DNS model if it is significantly negative (significance level marked by asterisks). Regardless of the forecast horizons, we find that the negative DM *t*-stat are prevalent for the maturities less than 10Y lying on the CDS curves, either when comparing with DNS model or VAR-elasticnet model. In other words, the network DNS model is superior to the DNS model as well as VAR-elasticnet model in an out-of-sample forecast, especially for the EU CDS curves.

6. Conclusions

Systemic risk, the risk of disruption to financial services, can be teased out directly through an investigation of CDS curves, in comparison with the conventional market-based approach

that uses stock returns as alternatives, such as the CoVaR measure. This study shows the information content of a CDS curve for its term structure implications. The changing shape of a credit curve implies credit deterioration over different time horizons (ST, MT or LT), which can be used to manage credit exposures with different maturities.

The extant literature has paid insufficient attention to the shape of CDS curves, the dynamics of CDS curves and the co-movement or interplay between curves. We contribute by: (1) using the DNS model to extract the ST, MT and LT default factors from a CDS curve; (2) quantifying the co-movement of CDS curves through a total connectedness indicator to measure the tendency of credit contagion; (3) measuring the default spillover/contagion in the ST, MT and LT perspectives; (4) documenting the channel of credit contagion from the LT, ST and MT perspectives; and, (5) conducting out-of-sample predictions for CDS curves based on the network-based DNS model.

The evidence from G-SIB banks shows that their CDS curves co-move tightly with high connectedness, especially in the long-term. The US banks contributed more to the long-term default spillover before 2012, whereas the European banks were major default transmitters during and after the

European debt crisis, both in the long-term and the short-term. The time-varying default connectedness and spillover can be viewed as an indicator that can be used to monitor systemic default risk and credit contagion, especially to identify the trouble makers that could trigger a clustered default in the system. We find that the channels of credit contagion are conducted through the TED spread, credit spread, and VIX. In particular, the VIX, representing investors' appetite for global credit risk, is especially effective for finding the trigger for a long-term credit contagion. The outperformance of the network DNS model indicates that predictions based on CDS curves require network information.

Acknowledgments

The authors would like to thank the two anonymous referees and the editor of this journal for their valuable comments and suggestions.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

We gratefully acknowledge financial support through DFG via IRTG 1792 'High Dimensional Non Stationary Time Series', Humboldt-Universität zu Berlin. Xiu Xu acknowledges the support of the Natural Science Foundation of China (Grant No. 71803140). The work was also partially supported by GACR EXPRO project 19-28231X.

ORCID

Xiu Xu  <http://orcid.org/0000-0002-7742-9601>

References

- Adrian, T. and Brunnermeier, M.K., CoVaR. *Amer. Econ. Rev.*, 2016, **106**(7), 1705–1741.
- Alexander, C. and Kaeck, A., Regime dependent determinants of credit default swap spreads. *J. Bank. Financ.*, 2008, **32**(6), 1008–1021.
- Baranovski, A., von Lieres und Wilkau, C. and Wilch, A., New recipes for estimating default intensities, Technical report, SFB 649 discussion paper, 2009–004, 2009.
- Benzoni, L., Collin-Dufresne, P., Goldstein, R.S. and Helwege, J., Modeling credit contagion via the updating of fragile beliefs. *Rev. Finan. Stud.*, 2015, **28**(7), 1960–2008.
- Bostanci, G. and Yilmaz, K., How connected is the global sovereign credit risk network?, Available at SSRN 2647251, 2015.
- Chen, C.Y.-H. and Härdle, W.K., Common factors in credit defaults swap markets. *Comput. Stat.*, 2015, **30**(3), 845–863.
- Das, S.R., Hanouna, P. and Sarin, A., Accounting-based versus market-based cross-sectional models of CDS spreads. *J. Bank. Financ.*, 2009, **33**(4), 719–730.
- Diebold, F.X. and Li, C., Forecasting the term structure of government bond yields. *J. Econom.*, 2006, **130**(2), 337–364.
- Diebold, F.X. and Mariano, R.S., Comparing predictive accuracy. *J. Bus. Econ. Statist.*, 1995, **20**(1), 253–263.
- Diebold, F.X., Rudebusch, G.D. and Aruoba, S.B., The macroeconomy and the yield curve: A dynamic latent factor approach. *J. Econom.*, 2006, **131**(1), 309–338.
- Diebold, F.X. and Yilmaz, K., On the network topology of variance decompositions: Measuring the connectedness of financial firms. *J. Econom.*, 2014, **182**(1), 119–134.
- Duan, J.-C. and Miao, W., Default correlations and large-portfolio credit analysis. *J. Bus. Econ. Stat.*, 2015, **34**(just-accepted), 1–36.
- Duffie, D., Eckner, A., Horel, G. and Saita, L., Frailty correlated default. *J. Finance*, 2009, **64**(5), 2089–2123.
- Fengler, M.R., Härdle, W.K. and Villa, C., The dynamics of implied volatilities: A common principal components approach. *Rev. Deriv. Res.*, 2003, **6**(3), 179–202.
- Filipović, D. and Trolle, A.B., The term structure of interbank risk. *J. Financ. Econ.*, 2013, **109**(3), 707–733.
- Flury, B.N., Common principal components in k groups. *J. Am. Stat. Assoc.*, 1984, **79**(388), 892–898.
- Flury, B.N. and Gautschi, W., An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form. *J. Sci. Stat. Comput.*, 1986, **7**(1), 169–184.
- Galil, K., Shapir, O.M., Amiram, D. and Ben-Zion, U., The determinants of CDS spreads. *J. Bank. Financ.*, 2014, **41**, 271–282.
- Getmansky, M., Girardi, G. and Lewis, C., Interconnectedness in the CDS market. *Financ. Anal. J.*, 2016, **72**(4), 62–82.
- Hansen, L.P. and Hodrick, R.J., Forward exchange rates as optimal predictors of future spot rates: An econometric analysis. *J. Polit. Econ.*, 1980, **10**, 829–853.
- Härdle, W. and Simar, L., *Applied Multivariate Statistical Analysis*, 4th ed., 2015 (Springer: Berlin).
- Koop, G., Pesaran, M.H. and Potter, S.M., Impulse response analysis in nonlinear multivariate models. *J. Econom.*, 1996, **74**(1), 119–147.
- Krishnan, C., Ritchken, P.H. and Thomson, J.B., Predicting credit spreads. *J. Financ. Intermed.*, 2010, **19**(4), 529–563.
- Longstaff, F.A., Mithal, S. and Neis, E., Corporate yield spreads: Default risk or liquidity? New evidence from the credit default swap market. *J. Finance*, 2005, **60**(5), 2213–2253.
- Nelson, C.R. and Siegel, A.F., Parsimonious modeling of yield curves. *J. Bus.*, 1987, **10**, 473–489.
- Newey, W.K. and West, K.D., A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 1987, **55**(3), 703–08.
- Pan, J. and Singleton, K.J., Default and recovery implicit in the term structure of sovereign CDS spreads. *J. Finance*, 2008, **63**(5), 2345–2384.
- Puliga, M., Caldarelli, G. and Battiston, S., Credit default swaps networks and systemic risk. *Sci. Rep.*, 2014, **4**, 6822.
- Shaw, F., Murphy, F. and O'Brien, F., The forecasting efficiency of the dynamic Nelson Siegel model on credit default swaps. *Res. Internat. Bus. Financ.*, 2014, **30**, 348–368.
- Yang, J. and Zhou, Y., Credit risk spillovers among financial institutions around the global credit crisis: Firm-level evidence. *Manage. Sci.*, 2013, **59**(10), 2343–2359.

A. Appendix

A.1. CDS spreads

A.2. Common principal component analysis (CPCA)

Here we introduce the maximum likelihood estimation procedures of CPC under the hypothesis in (23). The theoretical proof and the

asymptotic properties of the estimates are referred to Flury (1984) and Flury and Gautschi (1986).

Let S_i be the (unbiased) sample covariance matrix of an underlying K -variate normal distribution $N_K(\mu, \Psi_i)$ with sample size n_i . Then $n_i S_i$ follows a Wishart distribution with degrees of freedom $n_i - 1$, Härdle and Simar (2015)

$$n_i S_i \sim \mathcal{W}_K(\Psi, n_i - 1) \tag{A1}$$

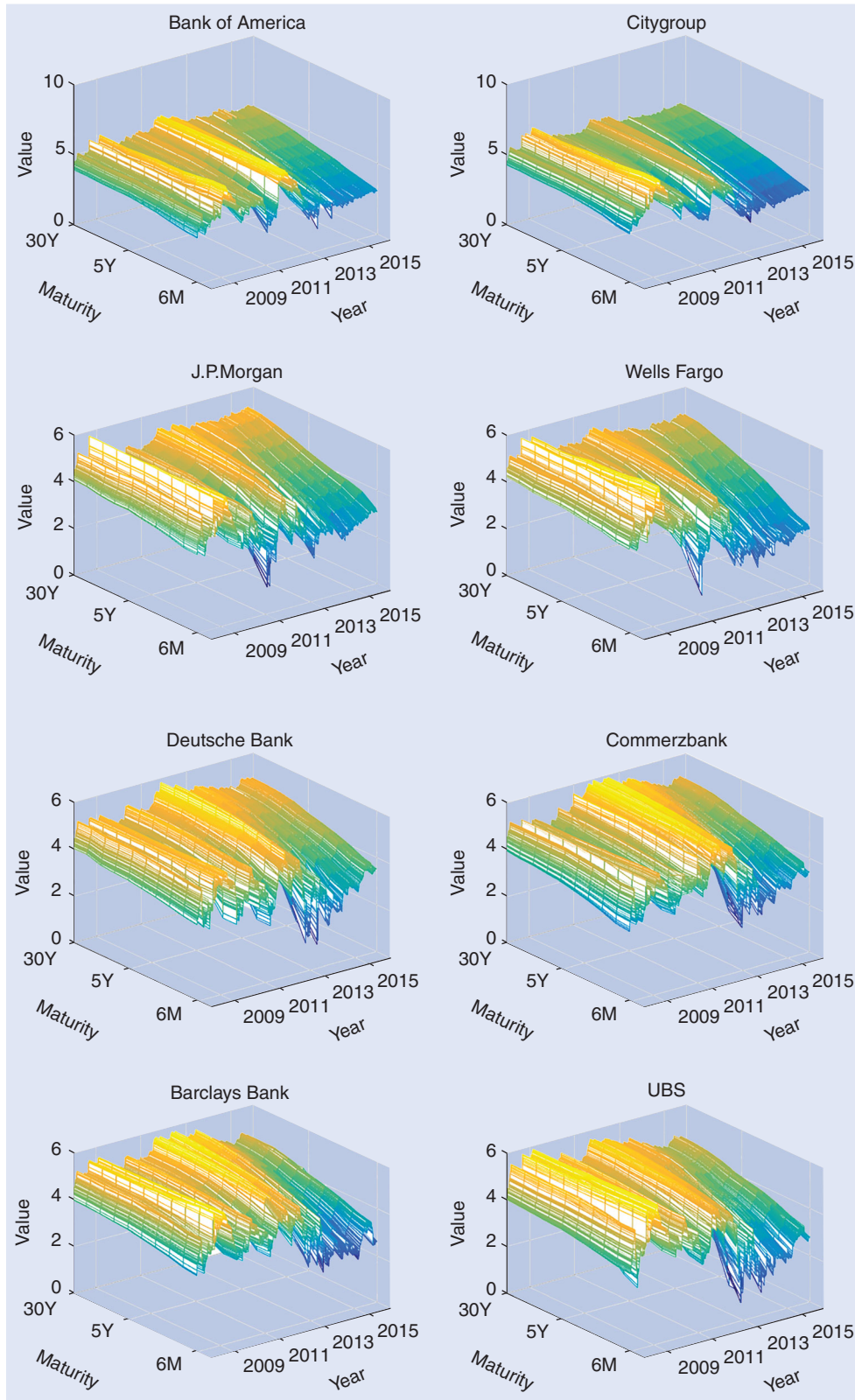


Figure A1. CDS spreads data.

Hence for N Wishart matrices S_i with sample size n_i , the likelihood function is

$$L(\Psi_1, \dots, \Psi_N) = C \prod_{i=1}^N \exp \left\{ \text{tr} \left(-\frac{1}{2} (n_i - 1) \Psi_i^{-1} S_i \right) \right\} |\Psi_i|^{-(1/2)(n_i - 1)} \quad (\text{A2})$$

where C is a constant not depending on the parameters Ψ_i . Maximizing the likelihood is equivalent to minimizing the function

$$g(\Psi_1, \dots, \Psi_N) = \sum_{i=1}^N (n_i - 1) \{ \log |\Psi_i| + \text{tr}(\Psi_i^{-1} S_i) \} \quad (\text{A3})$$

Assuming that H_{CPC} in equation (23) holds, replacing Ψ_i by $\Gamma \Lambda_i \Gamma^\top$, yields,

$$g(\Gamma, \Lambda_1, \dots, \Lambda_N) = \sum_{i=1}^N (n_i - 1) \sum_{j=1}^K \left(\log \lambda_{ij} + \frac{\gamma_j^\top S_i \gamma_j}{\lambda_{ij}} \right). \quad (\text{A4})$$

We impose the orthogonal constraints in Γ by using the Lagrange multipliers μ_j for the K constraints $\gamma_j^\top \gamma_j = 1$, and using the multipliers μ_{hj} for the remaining $K(K - 1)/2$ constraints $\gamma_h^\top \gamma_j = 0$ for

($h \neq j$). Hence the Lagrange function is

$$g^*(\Gamma, \Lambda_1, \dots, \Lambda_N) = g(\cdot) - \sum_{j=1}^K \mu_j (\gamma_j^\top \gamma_j - 1) - 2 \sum_{h < j}^K \mu_{hj} \gamma_h^\top \gamma_j.$$

Taking partial derivatives with respect to all λ_{im} and γ_m , the solution of the CPC model is given by the generalized system of characteristic equations, Flury (1984)

$$\gamma_m^\top \left\{ \sum_{i=1}^N (n_i - 1) \frac{\lambda_{im} - \lambda_{ij}}{\lambda_{im} \lambda_{ij}} S_i \right\} \gamma_j = 0, \quad m, j = 1, \dots, K, \quad m \neq j. \quad (\text{A5})$$

This is solved using

$$\lambda_{im} = \gamma_m^\top S \gamma_m, \quad i = 1, \dots, N, \quad m = 1, \dots, K$$

under the constraints

$$\gamma_m^\top \gamma_j = \begin{cases} 0 & m \neq j \\ 1 & m = j \end{cases}. \quad (\text{A6})$$

Flury (1984) proves existence and uniqueness of the maximum of the likelihood function, and Flury and Gautschi (1986) provide a numerical algorithm.



Dynamic semi-parametric factor model for functional expectiles

Petra Burdejová¹ · Wolfgang K. Härdle^{1,2}

Received: 30 November 2017 / Accepted: 7 March 2019 / Published online: 3 April 2019
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract

High-frequency data can provide us with a quantity of information for forecasting and help to calculate and prevent the future risk based on extremes. This tail behaviour is very often driven by exogenous components and may be modelled conditionally on other variables. However, many of these phenomena are observed over time, exhibiting non-trivial dynamics and dependencies. We propose a functional dynamic factor model to study the dynamics of expectile curves. The complexity of the model and the number of dependent variables are reduced by lasso penalization. The functional factors serve as a low-dimensional representation of the conditional tail event, while the time-variation is captured by factor loadings. We illustrate the model with an application to climatology, where daily data over years on temperature, rainfalls or strength of wind are available.

Keywords Factor model · Functional data · Expectiles · Extremes

1 Introduction

Statistical analysis of high-dimensional data nowadays plays a crucial role in various fields. Usually, one observes a high-dimensional vector evolving in time, that can be not only correlated to other variables but hide various types of inter-dependencies. One solution on how to analyze such data for further modelling is to treat it as discrete observations of functional times series. For example, observed as a function of

This research was supported by 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.

✉ Petra Burdejová
petra.burdejova@hu-berlin.de

¹ Humboldt-Universität zu Berlin, Berlin, Germany

² Sim Kee Boon Institute for Financial Economics, Singapore Management University, Singapore, Singapore

time over the year, the evolution of temperature curves; see e.g. Ramsay and Silverman (2005), the wind speed; see e.g. Burdejova et al. (2017), or pollution data; see e.g. Ignaccolo et al. (2008), can exhibit the trend for interesting periodical patterns supporting the thesis of climate change.

The same approach of functional data analysis was applied in health-care and clinical research, when Erbas et al. (2007) tested the trend in breast-cancer mortality, or Lee and Carter (1992) who performed the population analysis for mortality and fertility curves. Countless applications can be found in financial engineering, for example, yield curve modelling as Nelson and Siegel (1987) or Härdle and Majer (2016), modelling the collateralised debt obligations; see Choros-Tomczyk et al. (2016), analyzing the dynamics of limit order book or implied volatility; see e.g. Benko et al. (2009) or even the intraday price curves; see Kokoszka et al. (2014).

However, in most of the above-mentioned applications, one is interested in capturing the tail behaviour of the variables rather than variation around the mean. The majority of recent research in functional data has nonetheless focused only on the variation around the mean, as can be seen in the monographs of Ramsay and Silverman (2005), Ferraty and Vieu (2006) or Horváth and Kokoszka (2012).


For that purpose, in our work, we generalise one of the functional analytical models for expectiles. Expectiles, similar to quantiles, are tail measures, which uniquely characterise the conditional distribution of random variables. The same way as the quantile for the level of $\tau = 0.5$ corresponds to the median, and the $\tau = 0.5$ expectiles correspond to the mean. In case of the conditional expectiles on the other variable (e.g. time over year for the temperature or time over day for intra-day price curves) we refer to, so-called, expectile curves. Useful discussion about the advantages of such expectile regression and its comparison to quantile regression can be found for example in Kneib (2013).

Guo et al. (2015) modelled such expectile and quantile curves by rewriting them via Karhunen–Loève expansion. Tran et al. (2018) also presented analogue principal components of such tail event curves in an asymmetric norm. Both of these approaches assume the observations to be the independent realizations of a stationary stochastic process. Even though Hörmann and Kokoszka (2010) showed that the Karhunen–Loève expansion approach is suitable for the temporal dependence between functional observations, the question of modelling strong-dependent or non-stationary functional observations of extremes remains open. Therefore, the goal of our model is twofold: focus on modeling of the conditional extreme events for different expectile-levels, capture the dynamics of such tail event curves concerning any hidden pattern, dependence or non-stationarity.

In this work, we extend the generalized dynamic semi-parametric factor model for expectiles and provide the convergent algorithm for its estimation. This approach offers to focus on modelling the time-development of tail-event expectile curves with respect to possible strong-dependency or non-stationarity as well. Our work refers to the factor models as in Park et al. (2009) and Song et al. (2014), who did similar dynamic semi-parametric factor models for the L_2 -norm, which, for our model, corresponds as a specific case of expectile at 0.5-level.

As a motivation, let us assume that there is a need to estimate a collection of expectile curves, each coming from a separate data-set. Our first motivation example used

in Sect. 5 regards a set of daily observations of average temperature, i.e. the data as a vector $X_n \in \mathbb{R}^{365}$ for a year $n = 1, \dots, N$. In this situation, one needs to analyse jointly the time (over years for $n = 1, \dots, N$) as well as space dynamics (within the year for t th element of $X_n[t]$ for $t = 1, \dots, 365$) by simultaneous fitting. In the factor model, functional factors serve as a low-dimensional representation of the conditional tail event, while the time-variation is captured by factor loadings. All of them we approximate by a linear combination of basis-functions. Since the temporal dependencies and non-stationarity can arise from different sources, we start with an overparametrized model. That captures almost any behaviour or trend, including the cycles and the linear or quadratic trend. Then we seek a sparse solution and reduce the complexity of the model and the number of dependent variables by lasso penalization. Further, we apply the proposed model for estimation and forecasting of daily temperature. Other applicable examples can also be found in the analyses of customer demand planning, such as forecasting of electricity consumption as seen in López-Cabrera and Schulz (2017) via VAR-model or finance, economics, climatology or neurobiology as mentioned above.

The paper is organised as follows. After reviewing the concept of expectile curves in the first section, we present the dynamic semi-parametric factor model for expectiles and the algorithm to estimate this model based on the iterative weighted least squares. In Sect. 3 we examine the performance of our model. The performance of the algorithm is examined in the simulation study in Sect. 4. Finally, we apply the model to a Chinese temperature dataset in Sect. 5 and the dataset of the wind speed of hurricanes in Sect. 6. The last section summarises our findings. All codes used to obtain the results in this paper are available at  Quantlet, see details in Borke and Härdle (2017a, b).

2 Expectiles and expectile curves

The concept of expectiles was first presented by Newey and Powell (1987). Expectiles have a similar interpretation as quantiles, but are more efficient, and easier to compute due to the L_2 -norm in addition to being a coherent risk measure. Having a random variable Y the τ -expectile can be obtained by minimizing the expected loss:

$$e^\tau = \arg \min_{\theta} E \{ \rho_\tau (Y - \theta) \}$$

with asymmetric loss function

$$\rho_\tau (u) = |u|^\alpha \left| \tau - \mathbf{I}_{\{u < 0\}} \right|,$$

where $\alpha = 2$. In case of $\alpha = 1$ we get the quantiles. By generalization we can also get M-quantiles, see Breckling and Chambers (1988) or Jones (1994), who also showed that expectiles can be expressed as quantiles.

Expectiles can be understood intuitively in a similar way as quantiles. Though the τ -quantile can be defined as a value above the $\tau \cdot 100\%$ observations, expectile also takes the distance into account. τ -expectile is defined such that $\tau \cdot 100\%$ of the distance

of observations to it corresponds to the observations below it. Thus, expectiles are more sensitive to extreme observations and outliers.

However, in reality Y is usually associated with a vector of covariates X . For example, the variable X can express the development over time, i.e.:

$$e^\tau(x) = \arg \min_{\theta} E \{ \rho_\tau(Y - \theta) \mid X = x \}.$$

One is then interested in studying the conditional expectile as a function of x . For that purpose, we define the generalized regression τ -expectile as:

$$e^\tau(\cdot) = \arg \min_{f \in \mathcal{F}} E \{ \rho_\tau(Y - f(X)) \}.$$

where $f(\cdot)$ is a nonparametric function of covariates X from a set of functions \mathcal{F} , such that the expectation is well defined.

There are more possibilities on how to estimate such an expectile curve from an observed dataset. For example, expectile curve $e^\tau(\cdot)$ can be approximated by any basis and estimated iteratively. Schnabel and Eilers (2009) proposed to approximate the curves with P-splines and combine it with the LAWS (least average weighted squares) algorithm.

The aim of our work is to model a collection of N generalized expectiles curves $e_n^\tau(\cdot), n = 1, \dots, N$ with a semi-parametric factor model.

3 Dynamic semi-parametric factor model for expectile curves

Let us fix the level τ and assume the functional time series $e_n, n = 1, \dots, N$. We represent such a random process via factor model :

$$e_n(t) = \sum_{k=1}^K Z_{nk} m_k(t) = Z_n^\top m(t),$$

with time-varying factor loadings Z_{nk} and functional factors $m_k(t)$. Index t captures the spatial dependency while the index n express the evolution over time.

Suppose both sequences factorize over space and time with respect to some fixed bases. Thus, for some J -dimensional time basis $U^\top = (U_1, \dots, U_J)$ with $U_i = (U_i(1), \dots, U_i(N)), i = 1, \dots, J$ and L -dimensional space basis $\Psi = (\Psi_1, \dots, \Psi_L)$ with $\Psi_i = (\Psi_i(1), \dots, \Psi_i(T))$, we have the decomposition:

$$Z_{nk} = \sum_{j=1}^J \alpha_{kj} U_j(n) \quad \text{i.e. } Z_n = A_{K \times J} \cdot U(n)$$

and

$$m_k(t) = \sum_{l=1}^L \beta_{kl} \Psi_l(t) \quad \text{i.e. } m(t) = B_{K \times L} \Psi(t),$$

where $U^\top(n) = (U_1(n), \dots, U_J(n))$ and $\Psi(t) = (\Psi_1(t), \dots, \Psi_L(t))$. Those lead to the final dynamic semi-parametric factor model (DSFM):

$$e_n = (AU(n))^\top (B\Psi(t)) = U(n)^\top C\Psi(t), \tag{1}$$

where $C = A^\top B$ is a $J \times L$ matrix of coefficients needed to be estimated.

For the choice of both bases, one may employ various basis functions. To capture the periodic variation in time one can use the fourier basis, for the global trend over time, any orthogonal polynomial basis may be suitable. For the space basis, either B-splines, any polynomial basis or even principal components or their alternatives such as principal expectile components defined by Tran et al. (2018) may be used.

In order to estimate this model, we propose the iterative algorithm for minimising the penalized loss function. We define the weights in a similar manner as in Schnabel and Eilers (2009). Once the space and time basis are pre-specified, the choice of significantly loaded space and time basis functions is done via LASSO-penalization of the coefficient matrix C . As before, assume the fixed expectile level $\tau \in (0, 1)$ and for the observed discrete datapoints $Y_{n,t}, n = 1, \dots, N$ and $t = 1, \dots, T$:

1. Start with a setup for the weights $w_{n,t} = 0.5$. This corresponds to the mean curves.
2. Estimate the matrix \widehat{C} by minimising

$$\underset{C}{\operatorname{argmin}} \underbrace{\sum_{n=1}^N \sum_{t=1}^T w_{n,t} \left\{ Y_{n,t} - U(n)^\top C\Psi(t) \right\}}_{l(C)} + \lambda \sum_{j=1}^J \|C_{G_j}\|_2,$$

where the penalization term $\lambda \sum_{j=1}^J \|C_{G_j}\|_2$ is a group-Lasso penalization. The subsets for group-Lasso are pre-defined.

3. Update the weights

$$w_{n,t} = \begin{cases} \tau & \text{if } Y_{n,t} > U(s)^\top \widehat{C}\Psi(t), \\ 1 - \tau & \text{otherwise.} \end{cases}$$

4. Iterate via Steps 2. and 3. Recompute the weights until convergence, i.e. until there is no change in weights $w_{n,t}$.

Even though we can set the separate elements of C -matrix as the groups in LASSO-penalization, the group-LASSO would also allow us to give some specific structure or importance into the pre-defined basis if needed. In general, e.g. Huang and Zhang (2010) show and discuss the benefits of group-lasso. Note that function $l(C)$ in Step 2 is continuously differentiable and obtains a global minimum. Yang and Zou (2015) proposed the algorithm to solve such an optimization problem and proved its convergence for different types of “empirical loss + group lasso penalty” optimisation problems satisfying a quadratic majorization condition. Since for each iteration, the weights in Step 2 are fixed, it can be seen that our specific definition of weighted least squares combined with group-lasso penalization fulfills these requirements. Liao

et al. (2018) was also analyzing the similar optimization problem from the theoretical point of view, particularly the penalized expectile regression with an adaptive LASSO penalty and its asymptotic properties. The model proposed in this paper can be considered as an extension of work (Song et al. 2014). They proposed the generalized dynamic semi-parametric factor model in L_2 norm. Regarding prediction error, we refer to their work. Our model serves as an extension for asymmetric norm.

4 Simulation study

In order to evaluate the performance of the proposed model and the algorithm above we performed a simulation study. We followed the set up of Guo et al. (2015) and Tran et al. (2018), since they both proposed the alternatives for the modelling of tail event expectile curves. The data $Y_{n,i}$, $n = 1, \dots, N$ and $i = 1, \dots, T$, are simulated as:

$$Y_{n,i} = \mu(t_i) + \alpha_{1,n} f_1(t_i) + \alpha_{2,n} f_2(t_i) + \varepsilon_{n,i},$$

where t_i 's are the equidistant points on $[0, 1]$. We set the mean function $\mu(t)$ as $\mu(t) = 1 + t + \exp\{-t - 0.6\}^2/0.05\}$ and the principal component curves as $f_1(t) = \sqrt{2} \sin(2\pi t)$ and $f_2(t) = \sqrt{2} \cos(2\pi t)$.

Further, we consider the two following different scenarios for the scores of principal components $\alpha_{1,n}$ and $\alpha_{2,n}$ and 4 different error scenarios:

1. The scores set as $\alpha_{1i} \sim N(0, 36)$ and $\alpha_{2i} \sim N(0, 9)$ are both iid. The error term $\varepsilon_{n,i}$'s is: (1) iid $N(0, \sigma_1^2)$, (2) iid $t(5)$, (3) independent $N\{0, \mu(t_j)\sigma_1^2\}$ and (4) iid $\log N(0, \sigma_1^2)$. With $\sigma_1^2 = 1$.
2. The scores set as $\alpha_{1i} \sim N(0, 16)$ and $\alpha_{2i} \sim N(0, 9)$ are both iid. The error term ε_{ni} 's is: (1) iid $N(0, \sigma_2^2)$, (2) iid $t(5)$, (3) independent $N\{0, \mu(t_j)\sigma_2^2\}$ and (4) iid $\log N(0, \sigma_2^2)$. With $\sigma_2^2 = 0.5$.

For each of the parameter settings we ran the simulations 200 times. These scenarios allow us to analyse the different coefficient-to-coefficient-to-noise variations as well as the scenarios for fat tail errors (scenario of ε_2), heteroscedastic (scenario of ε_3) and skewed errors (scenario of ε_4). We analysed the performance for $\tau = 0.5, 0.6, \dots, 0.9$ based on the mean squared error (MSE) and its standard deviation (SD). Table 1 summarises the recorded MSE for the simulations. The standard deviations are given in brackets.

Regarding the choice of basis in all scenarios we chose to use $T/2$ b-spline curves for the space basis. The time basis was created as $N/2$ curves of the fourier basis and 3 trend curves: linear, quadratic and logarithmic. One has to be aware that the choice of basis, i.e. the number of basis functions can also have an impact on the results.

From the observed MSEs we concluded that whenever the error distribution is skewed, the model is likely to produce big MSEs. The model performs, in general, very well for different τ -levels and comparable to the already mentioned alternatives proposed by Guo et al. (2015) or Tran et al. (2018). However, since for higher tau levels fewer observations are taken into account with higher weights, the accuracy decreases and the MSEs increase with higher τ -s.

Table 1 Average MSE and its standard deviation in brackets by 200 simulation runs and different scenarios (DYTEC_simulation_all)

τ	ε	$N = 20, T = 100$		$N = 50, T = 150$	
		$\sigma_\varepsilon = 0.5$	$\sigma_\varepsilon = 1$	$\sigma_\varepsilon = 0.5$	$\sigma_\varepsilon = 1$
$\tau = 0.5$	$\varepsilon_1 \sim N(0, \sigma_\varepsilon^2)$	0.238 (0.011)	0.247 (0.012)	0.467 (0.023)	0.462 (0.013)
	$\varepsilon_2 \sim t(5)$	0.779 (0.053)	0.773 (0.028)	0.777 (0.054)	0.761 (0.028)
	$\varepsilon_3 \sim N(0, \mu(t_i)\sigma_\varepsilon^2)$	0.439 (0.022)	0.443 (0.013)	0.868 (0.054)	0.853 (0.023)
	$\varepsilon_4 \sim \log N(0, \sigma_\varepsilon^2)$	1.444 (0.058)	1.444 (0.028)	1.444 (0.592)	1.444 (0.271)
$\tau = 0.6$	ε_1	0.251 (0.013)	0.267 (0.018)	0.488 (0.024)	0.485 (0.016)
	$\varepsilon_2 \sim t(5)$	0.824 (0.059)	0.822 (0.034)	0.820 (0.059)	0.806 (0.033)
	$\varepsilon_3 \sim N(0, \mu(t_i)\sigma_\varepsilon^2)$	0.461 (0.023)	0.469 (0.019)	0.906 (0.046)	0.888 (0.026)
	$\varepsilon_4 \sim \log N(0, \sigma_\varepsilon^2)$	0.721 (0.097)	0.713 (0.051)	3.523 (1.055)	3.257 (0.501)
$\tau = 0.7$	ε_1	0.288 (0.011)	0.355 (0.012)	0.540 (0.023)	0.562 (0.013)

Table 1 continued

τ	ε	$N = 20, T = 100$		$N = 50, T = 150$	
		$\sigma_\varepsilon = 0.5$	$\sigma_\varepsilon = 1$	$\sigma_\varepsilon = 0.5$	$\sigma_\varepsilon = 1$
$\tau = 0.8$	$\varepsilon_2 \sim t(5)$	0.937 (0.079)	0.971 (0.066)	0.927 (0.079)	0.936 (0.055)
	$\varepsilon_3 \sim N(0, \mu(t_i)\sigma_\varepsilon^2)$	0.517 (0.028)	0.568 (0.061)	0.995 (0.050)	0.989 (0.043)
	$\varepsilon_4 \sim \log N(0, \sigma_\varepsilon^2)$	0.985 (0.148)	1.014 (0.096)	5.087 (1.655)	4.733 (0.800)
	ε_1	0.418 (0.073)	0.786 (0.365)	0.629 (0.053)	0.816 (0.204)
$\tau = 0.9$	$\varepsilon_2 \sim t(5)$	1.094 (0.108)	1.392 (0.320)	1.056 (0.103)	1.217 (0.187)
	$\varepsilon_3 \sim N(0, \mu(t_i)\sigma_\varepsilon^2)$	0.641 (0.069)	0.972 (0.347)	1.093 (0.068)	1.228 (0.182)
	$\varepsilon_4 \sim \log N(0, \sigma_\varepsilon^2)$	1.289 (0.205)	1.598 (0.367)	6.514 (2.204)	6.209 (1.089)
	ε_1	2.316 (1.247)	3.819 (1.648)	1.660 (0.689)	2.221 (0.779)
$\tau = 0.9$	$\varepsilon_2 \sim t(5)$	2.669 (1.069)	3.821 (1.527)	2.020 (0.675)	2.484 (0.751)
	$\varepsilon_3 \sim N(0, \mu(t_i)\sigma_\varepsilon^2)$	2.349 (1.087)	3.719 (1.577)	1.989 (0.652)	2.428 (0.741)
	$\varepsilon_4 \sim \log N(0, \sigma_\varepsilon^2)$	3.078 (1.154)	4.551 (1.542)	8.486 (2.627)	8.295 (1.519)
	ε_1				

5 Application for temperature curves

We apply our model to Chinese temperature data, which consists of daily average temperatures of 159 weather stations from the years 1957 to 2009. In this case $n = 1, \dots, 53$ corresponds to the year and $t = 1, \dots, 365$ corresponds to the day during the year. The model is applied for each station separately. It is obvious that while the factor loadings Z_{nk} vary over the years, the dependence within the year is captured by factors $m_k(t)$ themselves.

5.1 The choice of time basis U

The proper time basis allows us to capture any periodic variation as well as any trend. In the case of temperature data, we do not assume only linear trend, but also add quadratic and logarithmic function to the basis:

$$u_1(n) = \frac{n}{N}, \quad u_2(n) = \frac{n^2}{N^2}, \quad u_3(n) = \frac{\log n}{\log N}.$$

For the periodicity we use $N - 1$ fourier basis functions with period $N = 53$:

$$\begin{aligned} u_4 &= \text{constant} \\ u_5 &= \frac{1}{\sqrt{\frac{N}{2}}} \sin\left(\frac{2 \cdot \pi \cdot n}{N}\right), \\ u_6 &= \frac{1}{\sqrt{\frac{N}{2}}} \cos\left(\frac{2 \cdot \pi \cdot n}{N}\right), \\ &\dots \\ u_{52} &= \frac{1}{\sqrt{\frac{N}{2}}} \sin\left(\frac{\frac{(N-1)}{2} \cdot \pi \cdot n}{N}\right), \\ u_{53} &= \frac{1}{\sqrt{\frac{N}{2}}} \cos\left(\frac{\frac{(N-1)}{2} \cdot \pi \cdot n}{N}\right). \end{aligned}$$

In general, one may operate with various types of basis functions, such as higher-power polynomials, local polynomials, trigonometric or periodic functions, splines, etc., with regard to follow various types of non-linearity concerning the specific design of the given data. In this step the group-lasso approach could be beneficial. In case of any previous meteorological information about the periodicity and the cycles, the structure can be taken into account and implemented into the model thus an appropriate subset of bases is chosen.

5.2 The choice of space basis Ψ

In order to model the specific structure and pattern within the year, we set the space basis Ψ as simple B-splines, particularly $\frac{T-1}{2}$ functions of the order 5. That corresponds to the polynomials of degree 4. Classical approach is to arbitrary use at least degree 3 (the lowest degree that supports curves inflection) or higher. We wanted to increase the degree, however, avoiding the overfitting that leads to jumpy curves. One can also choose to use the first few principal components explaining 85% of variance or even more complex principal expectile components introduced by Tran et al. (2018) in order to not lose the specific information in tails.


5.3 Forecasting

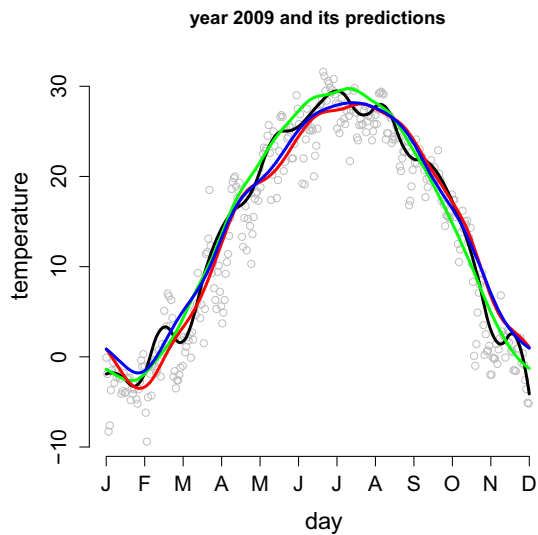
One of the traditional approaches for the forecasting of functional data can be done via Karhunen–Loève expansion. The functional time series is rewritten via principal components and their scores are consequently modeled separately with an appropriate model. The forecast, obtained by the model of the scores, together with the original principal components, are used for the prediction of functional time series. A similar approach can be done for expectiles; see e.g. López-Cabrera and Schulz (2017) who did a two-step approach. In the first step, for a fixed level of τ , the series of expectile curves is computed. Second, the principal component decomposition of the curves and the forecast of their scores are done via a vector auto-regressive model. However, one of the main restrictions of such an approach is the assumption of the weak-dependent data.

The proposed DSFM model for expectile curves provides us with the easy method of forecasting. Since the matrix of coefficients \widehat{C} is estimated once and the space basis is already predefined as well, one only needs to forecast the time basis. The basis consists of a set of functions, each of which can be simply prolonged for the value in time $N + 1$. For each $i = 1, \dots, J$ denote the values of prolonged basis function as $\widetilde{U}_i = (U_i(1), \dots, U_i(N), U_i(N + 1))$ and the new updated prolonged time basis as $\widetilde{U} = (\widetilde{U}_1, \dots, \widetilde{U}_J)$. We obtain the forecast as:

$$\hat{e}_{N+1}(t) = (A\widetilde{U})^\top (B\Psi) = \widetilde{U}^\top \widehat{C}\Psi.$$

Aiming to demonstrate the model, we applied the proposed DSFM-model to Chinese temperature data sets. The daily observations for a specific station No.1. from 1957 to 2008 were used for the estimation of the model and matrix \widehat{C} . Consequently, the model with prolonged time basis was used for the prediction of an expectile curve for the upcoming year 2009. We used two other benchmark approaches to compare the quality of our forecast. First, the VAR-model used the principal components of the pre-computed expectile curves from the years 1957–2008, forecasting their scores via VAR(4) model. The second model takes into consideration the possible non-stationarity and thus uses the ARIMA model to forecast the score of each component separately. Figure 1 shows the data for the year 2009, together with the expectile curve for fixed level $\tau = 0.8$. According to MSE, the DSFM-forecast (green) better

Fig. 1 Chinese temperature data for the year 2009 (grey) with its smoothed expectile curve (black). Predictions of expectile curve for level $\tau = 0.8$ by DSFM model (green), VAR model (red) and ARIMA model (blue)
 (DYTEC_temperature)
 (colour figure online)



predicts the expectile curve than the VAR-model (red) or the ARIMA-model (blue), which are constructed by forecasting the scores of principal components.

6 Application to wind speed data

As a second application, we used our DSFM-model for modeling expectile curves of the wind speed of hurricanes in a hurricane season across the North Atlantic basin from 1962 to 2011. As earlier, the observed data has the form $X_n(t_i)$, where the times t_i are here separated by six hours, and the index n stands for the year. The value $X_n(t_i)$ is the wind speed in knots ($1 \text{ kn} = 0.5144 \text{ m/s}$). The data is accessible on the Unisys Weather Information website, UNISYS (2015). We focused only on the hurricane-period from July till October, thus having $T = 400$ observations for every year $n = 1962, \dots, 2011$, i.e. $N = 50$. We treated time $0 \leq t \leq T$ within a year as continuous, and the observed curves as functional data.

Motivated by the work of Burdejova et al. (2017), who tested the hypothesis of linear trends for hurricanes, we modeled the hurricane data with our DSFM-model and focused primarily on the estimation of coefficients for different trend curves in time basis.

6.1 The choice of time basis U

For the periodicity we used also ten fourier basis functions with period $N = 50$ and a constant function. Since we were mainly interested not only in linear trends, we also added quadratic and logarithmic function to the basis, so we set as before:

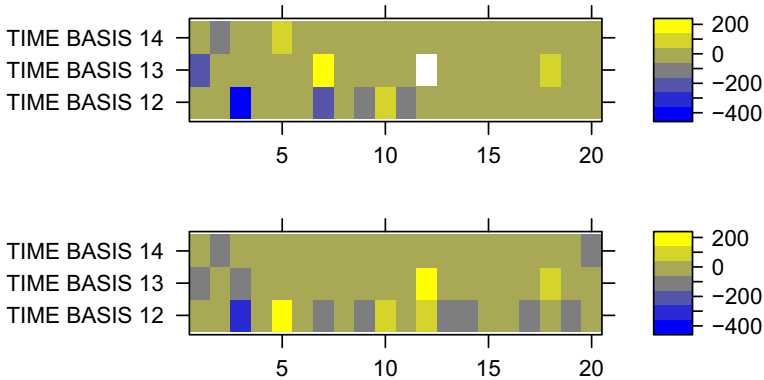




Fig. 2 Graphical representation of \hat{C} for hurricane data and level $\tau = 0.5$ (upper) and $\tau = 0.8$ (lower). The rows correspond to the last three time basis functions, i.e. logarithmic (time basis 12), quadratic (time basis 13) and linear (time basis 14). The

columns correspond to the space basis, i.e. 20 principal components ( [data_load_hurricanes](#)  [DYTEC_hurricanes](#))

$$u_{12}(n) = \frac{\log n}{\log N}, \quad u_{13}(n) = \frac{n^2}{N^2}, \quad u_{14}(n) = \frac{n}{N}.$$

6.2 The choice of space basis Ψ

In order to model the specific structure within the yearly period we use the principal components. We set the first 20 components as the space basis Ψ , since they explain 90% of the variance.

6.3 The estimates of matrix \hat{C}

The proposed algorithm was performed for two different τ -levels, $\tau = 0.5$ and $\tau = 0.8$. The estimations of matrices are shown in Fig. 2, for $\tau = 0.5$ (upper panel) and $\tau = 0.8$ (lower panel). Two conclusions are obvious from the estimations:

1. The coefficients related to the linear trend for all principal components are not as significant as the coefficients for quadratic and even logarithmic trend.
2. The linear trends have similar pattern for both τ -levels. However, this does not hold for the other two trends, whose coefficients related to all principal components differing with respect to τ , especially for logarithmic trend.

One could conclude that in the case of the historical observation of hurricanes, the question of testing models incorporating the trends other than linear is lying in the interest of future research.

7 Conclusion

In this paper, we propose the dynamic semi-parametric factor model for a joint estimation of expectile curves. For that purpose, we utilised a non-parametric series expansion for both factors and their (time-developing) scores. We have provided the convergent algorithm for its estimation that is based on the idea of iterative least squares. The presented model is thus a utile extension of the commonly known factor model for the mean, where L_2 norm is used.

This novel approach provides us with several advantages. One can easily directly estimate the extreme curves from the data without any need of pre-computing the expectile curves separately, i.e. not using two-step approach estimate each curve by expectile regression separately first and model the set of functions afterwards. Moreover, the method may be applied for non-stationary data as well. Any dynamics, hidden intra-dependencies, trends, or patterns of such tail event curves can be easily captured with the proper choice of time basis.

We have demonstrated good estimation properties in a simulation study for different setups of error terms and different expectile τ -levels as well. A method was applied to the Chinese temperature dataset of average daily temperatures over years in order to show its easy usability not only for modelling but also for forecasting, where it performs as well as any traditional approach predicting this type of functional data. The second application to the wind speed data of hurricanes showed not only the importance of considering various trends but also pointed out the fact of diverse factor structure for different τ -levels.

References

- Benko M, Härdle W, Kneip A (2009) Common functional principal components. *Ann Stat* 37(1):1–34. <https://doi.org/10.1214/07-AOS516>
- Borke L, Härdle WK (2017a) Q3-D3-LSA. In: Härdle WK, Lu HH, Shen X (eds) *Handbook of big data analytics*. Springer, New York. <https://doi.org/10.2139/ssrn.2871111>
- Borke L, Härdle WK (2017b) GitHub API based QuantNet mining infrastructure in R. Discussion Paper SFB2017-08, Humboldt-Universität zu Berlin. <https://doi.org/10.2139/ssrn.2927901>
- Breckling J, Chambers R (1988) M-quantiles. *Biometrika* 75:761–771. <https://doi.org/10.2307/2336317>
- Burdejova P, Härdle W, Kokoszka P, Xiong Q (2017) Change point and trend analyses of annual expectile curves of tropical storms. *Econom Stat* 1:101–117. <https://doi.org/10.1016/j.ecosta.2016.09.002>
- Choros-Tomczyk B, Härdle WK, Okhrin O (2016) A semiparametric factor model for CDO surfaces dynamics. *J Multivar Anal* 146:151–163. <https://doi.org/10.1016/j.jmva.2015.09.002>
- Erbas B, Hyndman RJ, Gertig DM (2007) Forecasting age-specific breast cancer mortality using functional data models. *Stat Med* 26(2):458–470. <https://doi.org/10.1002/sim.2306>
- Ferraty F, Vieu P (2006) *Nonparametric functional data analysis; theory and practice*. Springer, New York
- Guo M, Zhou L, Huang JZ, Härdle WK (2015) Functional data analysis of generalized regression quantiles. *Stat Comput* 25(2):189–202. <https://doi.org/10.1007/s11222-013-9425-1>
- Härdle WK, Majer P (2016) Yield curve modeling and forecasting using semiparametric factor dynamics. *Eur J Finance* 22(12):1109–1129. <https://doi.org/10.1080/1351847X.2014.926281>
- Hörmann S, Kokoszka P (2010) Weakly dependent functional data. *Ann Stat* 38(3):1845–1884. <https://doi.org/10.1214/09-AOS768>
- Horváth L, Kokoszka P (2012) *Inference for functional data with applications*. Springer, New York
- Huang J, Zhang T (2010) The benefit of group sparsity. *Ann Stat* 38(4):1978–2004. <https://doi.org/10.1214/09-AOS778>

- Ignaccolo R, Ghigo S, Giovenali E (2008) Analysis of air quality monitoring networks by functional clustering. *Environmetrics* 19(7):672–686. <https://doi.org/10.1002/env.946>
- Jones MC (1994) Expectiles and M-quantiles are quantiles. *Stat Probab Lett* 20:149–153. [https://doi.org/10.1016/0167-7152\(94\)90031-0](https://doi.org/10.1016/0167-7152(94)90031-0)
- Kneib T (2013) Beyond mean regression. *Stat Model* 13(4):275–303. <https://doi.org/10.1177/1471082X13494159>
- Kokoszka P, Miao H, Zhang X (2014) Functional dynamic factor model for intraday price curves. *J Financ Econ* 13:1–22. <https://doi.org/10.1080/1351847X.2014.926281>
- Lee RD, Carter LR (1992) Modeling and forecasting US mortality. *J Am Stat Assoc* 87(419):659–671. <https://doi.org/10.2307/2290201>
- Liao L, Park C, Choi H (2018) Penalized expectile regression: an alternative to penalized quantile regression. *Ann Inst Stat Math*. <https://doi.org/10.1007/s10463-018-0645-1>
- López-Cabrera B, Schulz F (2017) Forecasting generalized quantiles of electricity demand: a functional data approach. *J Am Stat Assoc* 112(517):127–136. <https://doi.org/10.1080/01621459.2016.1219259>
- Nelson CR, Siegel AF (1987) Parsimonious modeling of yield curves. *J Bus* 60(4):473–489. <https://doi.org/10.1086/296409>
- Newey W, Powell J (1987) Asymmetric least squares estimation and testing. *Econometrica*. <https://doi.org/10.2307/1911031>
- Park BU, Mammen E, Härdle W, Borak S (2009) Time series modelling with semiparametric factor dynamics. *J Am Stat Assoc* 104(485):284–298. <https://doi.org/10.1198/jasa.2009.0105>
- Ramsay J, Silverman B (2005) *Functional data analysis*. Springer, New York
- Schnabel SK, Eilers PH (2009) Optimal expectile smoothing. *Comput Stat Data Anal* 53(12):4168–4177. <https://doi.org/10.1016/j.csda.2009.05.002>
- Song S, Härdle WK, Ritov Y (2014) Generalized dynamic semi-parametric factor models for high-dimensional non-stationary time series. *Econom J* 17(2):S101–S131. <https://doi.org/10.1111/ectj.12024>
- Tran NM, Burdejova P, Osipenko M, Härdle WK (2018) Principal component analysis in an asymmetric norm. *J Multivar Anal*. <https://doi.org/10.1016/j.jmva.2018.10.004>
- UNISYS (2015) Data in atlantic and west pacific, 2015. Unisys Weather Information Systems. <http://weather.unisys.com/hurricane/index.php>. Accessed 20 Feb 2015
- Yang Y, Zou H (2015) A fast unified algorithm for solving group-lasso penalize learning problems. *Stat Comput* 25(6):1129–1141. <https://doi.org/10.1007/s11222-014-9498-5>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Investing with Cryptocurrencies—a Liquidity Constrained Investment Approach*

Simon Trimborn¹, Mingyang Li², and Wolfgang Karl Härdle³

¹National University of Singapore and Humboldt-Universität zu Berlin, ²Xiamen University, and

³Humboldt-Universität zu Berlin

Address correspondence to Mingyang Li, Wang Yanan Institute for Studies in Economics, Xiamen University, No. 422, Siming South Road, Xiamen, China, 361005 or e-mail: limingyang@stu.xmu.edu.cn

Received July 19, 2017; revised April 5, 2019; editorial decision April 7, 2019; accepted April 26, 2019

Abstract

Cryptocurrencies have left the dark side of the finance universe and become an object of study for asset and portfolio management. Since they have low liquidity compared to traditional assets, one needs to take into account liquidity issues when adding them to a portfolio. We propose a Liquidity Bounded Risk-return Optimization (LIBRO) approach, which is a combination of risk-return portfolio optimization under liquidity constraints. Cryptocurrencies are included in portfolios formed with stocks of the S&P 100, US Bonds, and commodities. We illustrate the importance of the liquidity constraints in an in-sample and out-of-sample study. LIBRO improves the weight optimization in the sense that it only adds cryptocurrencies in tradable amounts depending on the intended investment amount. The returns greatly increase compared to portfolios consisting only of traditional assets. We show that including cryptocurrencies in a portfolio can indeed improve its risk–return trade-off.

Key words: asset classes, blockchain, crypto-currency, CRIX, portfolio investment

JEL classification: C01, C58, G11

1 Introduction

With the emergence of cryptocurrencies, not only has a new kind of currency and transaction network arisen, but also a new kind of investment product. The cryptocurrency (CC) market has shown strong gains over the past years, which can be inferred from the CRIX, developed by [Trimborn and Härdle \(2018\)](#) and visualized at [thecrix.de](#). The CRIX index indicates a gain of the market of 500% over the past 2 years, which makes it attractive for investors. Simultaneously the market bears high risk in terms of price variations and operational risk. In the past years, users and exchanges were vulnerable in many ways, for

* The authors would like to thank the editor and two anonymous referees for their valuable comments to this article.

example, the traders on the exchange Mt. Gox experienced fraud and exchanges like Bitfinex got hacked. Also single users were subject to larceny. The situation has already improved a lot but there still remains a problem of trust, since the market is not fully developed. It has often been pointed out that a procedure called “cold-storage” should be used to secure one’s CCs. This refers to storing the access codes for the coins in such a way that they are disconnected from any device under threat of a hostile attack. While this source of risk can be managed comparably easily, the market risk is difficult to handle.

A natural question is why an investor should engage in such a risky market given these volatility effects. Advantages beside the opportunity for strong gains need to be present to make an investment worth the risk. An important perk is that CCs have a low linear dependency with each other and also the top 10 CCs (by market capitalization) have low correlations with traditional assets, [Elendner et al. \(2018\)](#). Since CCs have low correlations with each other and are uncorrelated with traditional assets, they are indeed interesting for investors due to the diversification effect. Making use of this advantage, [Brière, Oosterlinck and Szafarz \(2015\)](#) and [Eisl, Gasser, and Weinmayer \(2015\)](#) added Bitcoin to a portfolio of traditional assets and found an enhanced portfolio in terms of risk-return. Since alternative CCs (alt-coins, other than Bitcoin), have favorable properties too, we aim to construct portfolios consisting of traditional assets and several cryptocurrencies. [Lee Kuo Chuen, Guo, and Wang \(2017\)](#) worked in a related direction by investigating a portfolio mimicking the CRIX, the CRyptocurrency IndeX. Treating CRIX as a financial asset, [Chen et al. \(2017, 2018\)](#) investigated options pricing based on CRIX.

When investing in CCs, one is confronted with a higher volatility than for traditional assets, see [Figure 1](#). [Markowitz \(1952\)](#) developed a method which optimizes the portfolio allocation in terms of a minimum variance portfolio according to a target return. The approach was applied in a broad variety of applications and showed its usefulness especially in the case of Gaussian distributed data. But CCs are known to behave differently from the normal distribution, [Elendner et al. \(2018\)](#). In particular, the stronger tails come with higher risk, arising from higher moments, [Scaillet, Treccani, and Trevisan \(2018\)](#). Tail risk optimized portfolios might be worth considering in this market, for example, taking into account Conditional Value-at-Risk (CVaR), [Rockafellar and Uryasev \(2000\)](#). But there is another issue that cannot be handled by these risk optimization methods, namely the low liquidity of the CC market. [Figure 1](#) (right plot) presents a comparison of the liquidity measured by median daily trading amount of CCs and S&P 100 component stocks. It is obvious that the median daily trading amounts of CCs are all lower than the 25% quantile of S&P 100 stocks.

If we want to include CCs and stocks in the same portfolio, we need to avoid giving CCs too big a weight since this will induce a severe liquidity problem on adjusting the position when reallocating the portfolio. For example, if one holds a long position on an asset that is equal to twice its average daily trading amount, then one expects to take about two days to clear this position, following the same pace of the market. However, this may result in missing a trading opportunity. A proper way to deal with such a liquidity issue is the introduction of liquidity constraints on the weights. [Krokhmal, Palmquist, and Uryasev \(2002\)](#) utilized liquidity constraints in the sense of restricting the change in a position. [Darolles, Gourieroux, and Jay \(2012\)](#) choose a related approach by incorporating a penalty term into the optimization function, balancing the risk and change of positions in the portfolio. However we intend to be able to clear all positions at once, which is assumed to be in the interest of an investor engaging in a risky market like the CC market. Instead our definition

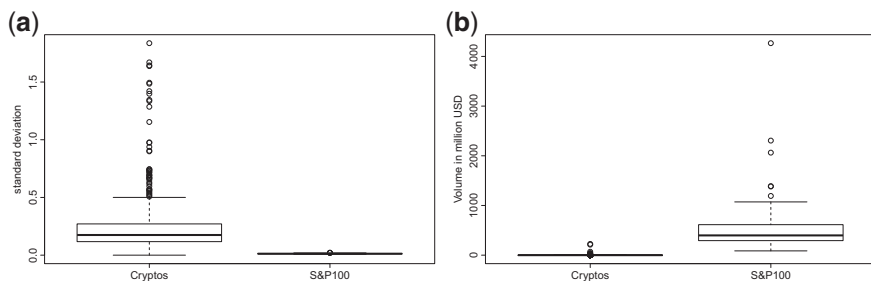


Figure 1 Boxplots of standard deviation and median trading volume (measured in US dollars) of all CCs and S&P 100 components, from April 22, 2014 to October 30, 2017. Obviously CCs have much lower daily trading volumes and higher volatilities than the stocks, highlighting the importance of volatility and liquidity risk management when investing in them.

(a) Comparison of standard deviation of CCs and S&P100 Equity Index components. (b) Comparison of median trading volume of CCs and S&P100 Equity Index components.

of liquidity constraints is concentrated on the entire weight given to a CC, rather than the allowed change in a position. Additionally, such an approach has the advantage of tackling a drawback of Markowitz portfolios. Minimum Variance optimized portfolios often suffer from extreme positive and negative weights, [Härde et al. \(2018\)](#). This may result from a single dominant factor in the covariance matrix, [Green and Hollifield \(1992\)](#). In an empirical study, [Jagannathan and Ma \(2003\)](#) find nonnegativity constraints on the weights to have an equal efficacy at removing the effect of a single dominant eigenvalue from the covariance matrix. [Fan, Zhang, and K. Yu. June \(2012\)](#) provide theoretical insights into their findings and find that constraining the weights from taking extreme positions is more effective than nonnegativity constraints. Thus introducing weight constraints gives us the opportunity to “kill two birds with one stone.”

Due to these challenges and the advantage from investing in cryptocurrencies, we aim at a portfolio optimization method which takes into account volatility or tail risk and low liquidity. We call it LIBRO—which is a combination of a risk optimization portfolio formation method and an additional restriction, which prevents big weights on assets with low liquidity. The portfolios are formed with Mean-Variance (Markowitz) and Conditional Value-at-Risk as risk measures. Due to the huge dimensionality of the asset universe and limited data availability, the sample covariance matrix may not be a well-conditioned estimator of its theoretical counterpart (well-conditioned in the sense that inverting the covariance matrix does not amplify the estimation error, [Ledoit and Wolf \(2004\)](#)). A well-conditioned and more accurate estimator was introduced by [Ledoit and Wolf \(2004\)](#), which we apply to the estimation of the Markowitz portfolios. Reduced factor model approaches were, for example, investigated by [Kozak, Nagel, and Santosh \(2017\)](#) and sparse estimation by, for example, [Friedman, Hastie, and Tibshirani \(2008\)](#). To investigate the robustness of the results, the reallocation dates in the out-of-sample study are set to be monthly and weekly. In order to overcome estimation difficulties driven by too short time series, we work under an extending window approach. Two datasets are compared in the application. The first one is a portfolio formed with S&P 100 components and CCs. The excess returns from the portfolio with CCs over the pure stock one range from 13.5 to 88% (gained over

3.5 years) in the in-sample analysis, and from 13.7 to 60% (gained over 2.75 years) in the out-of-sample analysis. When using stocks, bonds and commodities as the traditional assets, the results still range from 6 to 20.43% in-sample (3.5 years) and 6.7–24.38% out-of-sample (2.75 years). Summary statistics of the return series indicate that including CCs can increase the Sharpe Ratio, thus we show that including CCs can indeed improve the risk-return trade-off of the portfolio. Furthermore, the present article illustrates the importance of the liquidity constraints and their effects.

This article is organized as follows. Section 2 introduces the data. Section 3 presents the portfolio optimization methods. Section 4 introduces the liquidity constraints and Section 5 gives an in-sample and out-of-sample application with S&P 100 component stocks, Barclays Capital US Aggregate Index (US Bonds Index), S&P GSCI (Commodities Index), and CCs. The portfolios based on stocks are indicated by S, while the Stock, Bonds, and Commodities ones are indicated by SBC. The results are summarized in Section 6. The codes used to obtain the results in this article are available at www.quantlet.de, [Borke and Härdle \(2018\)](#) and [Borke and Härdle \(2017\)](#).

2 Data Description

In this article, 42 CCs are used to form portfolios together with traditional financial assets, with a sample period from April 22, 2014 to October 30, 2017. The daily prices (in USD) and trading volumes were downloaded from the CRIX cryptocurrencies database (the-crix.de), kindly provided by CoinGecko. The CCs were selected so that the average market cap during the sample period is no less than 10,000 US dollar. This criterion was applied since we target portfolios consisting of a reasonably high investment size, thus the CCs should have enough market capitalization to be added to a portfolio.

For the traditional financial assets, we chose the components of the S&P 100, Barclays Capital US Aggregate Index (US-Bonds Index), S&P GSCI (Commodities Index). The daily closing price (in USD) and trading volume, dated from April 22, 2014 to October 30, 2017, are downloaded from Datastream. To get the daily trading volume of stocks measured in US dollars, we multiplied the daily trading volume by the daily closing price. Three stocks were omitted: DowDuPont Inc., The Kraft Heinz Company, and PayPal Holdings, Inc., since they have a shorter sample period due to company mergers or spin-offs. For the indices we did not chose a particular Exchange Traded Fund, instead we assume perfect liquidity for both of them.

We show the summary statistics for the top 10 CCs over time in [Table 1](#). For the full list, see [Table A1](#) in the Appendix. In both tables, the CC statistics are arranged in decreasing order of their mean daily trading volume. For comparison, we list the summary statistics of stocks, a bond index, and a commodity index too. The summary statistics for the stocks are the average values for all stocks. The first 5 columns focus on the return series, while the remaining two list the mean trading volume and market capitalization. We will focus on [Table 1](#) to analyze the summary statistics of the CCs. Compared to the average annualized mean returns for stocks, which is 8%, and for the bond index, 5%, those for CCs can be quite shocking: except for PPC and BLK, all the other eight have returns that exceed 10%. What's more, five of them exceed 20%, three of them exceed 50%, and there is even one of them, DASH, that has an annualized average return that exceeds 100%. One can observe that three of the alt-coins, viz., Ripple (XRP), Dashcoin (DASH), and DigiByte

Table 1 Summary statistics of top ten CCs by trading volume. Ann.Ret and Ann.STD indicates annualized mean and standard deviation of the return of each CC, which are calculated by multiplying their daily counterparts by 250 and $\sqrt{250}$ respectively. For the purpose of comparison, we list the summary statistics of traditional financial assets at the bottom part of the table as well. "Bond" and "Commodity" indicate the summary statistics of the daily return of the bond index and commodity index, while "Stocks" indicates the average level of the summary statistics for the daily return of each individual stock

	Ann.Ret	Ann.STD	skewness	kurtosis	ρ	mean volume	market cap
BTC	0.49	0.55	-0.61	10.87	-0.01	4.71e+08	1.07e+10
LTC	0.29	0.87	0.34	23.70	0.02	9.04e+07	3.22e+08
XRP	0.68	1.09	2.72	37.76	0.01	4.88e+06	8.35e+08
DASH	1.16	1.37	0.62	50.60	-0.14	1.61e+06	1.52e+08
DOGE	0.11	0.96	0.97	15.54	0.01	3.87e+05	3.11e+07
NXT	0.17	1.08	0.80	8.42	-0.03	2.28e+05	2.08e+07
DGB	0.73	1.70	2.93	29.89	-0.03	1.63e+05	7.24e+06
PPC	-0.17	1.00	0.62	12.51	-0.05	1.05e+05	1.54e+07
BLK	-0.05	1.27	1.82	17.72	-0.08	7.15e+04	4.21e+06
VTC	0.33	1.76	1.84	17.83	-0.01	5.93e+04	2.42e+06
Stocks	0.08	0.20	-0.26	10.29	0.00	5.40e+08	
Bond	0.05	0.05	-1.91	12.38	-0.04		
Commodity	-0.20	0.20	0.05	4.84	-0.07		

(DGB), have a higher return than BTC, the dominant CC in the market, indicating that it is time to take these alt-coins into account for portfolio formation as well. However, the outstanding returns come at a price. Judging from [Table 1](#), CCs also have much higher volatility and tail risk. All CCs have an annualized volatility that exceeds 50%, with 7 of them even exceeding 100%. In contrast, S&P 100 stocks have an average annualized volatility of 20%, as does the commodity index, and the bond index has only 5%. Thus there is a trade-off in terms of high returns yet high standard deviations for the CCs. This finding is consistent with the size effect reported by [Elendner et al. \(2018\)](#). A similar picture is observed for the kurtosis: all the listed CCs, except for BTC and NXT, have a higher kurtosis than the stocks (10.29), bond index (12.38), and commodity index (4.84). Among these CCs, BTC has the lowest volatility of the 10, which is not surprising since it has the largest market capitalization, largest trading volume, and longest trading history, which makes it the most mature CC. Besides, BTC is the only CC that has a negative skewness, akin to stocks and bonds, while the other CCs all have positive skewness. This may imply that the other CCs are in a different developmental phase than BTC. For the auto-correlation, most of the top 10 CCs have negative or slightly positive ρ , like stocks and bonds.

The influence of liquidity is a major focus of this study. The evolution of the log of the trading volume of the top 10 CCs by trading volume is shown in [Figure 2](#), which, for each of the CCs, shows the daily changes. [Table 2](#) provides descriptive statistics of the log returns of the turnover value. For all of the 10 CCs one can see that they vary around a mean value of roughly 0. The median is slightly negative for all CCs, suggesting that there are more frequent decreases in the liquidity than increases. However the first and third quantiles show more or less opposing values, which hints that the variation around the

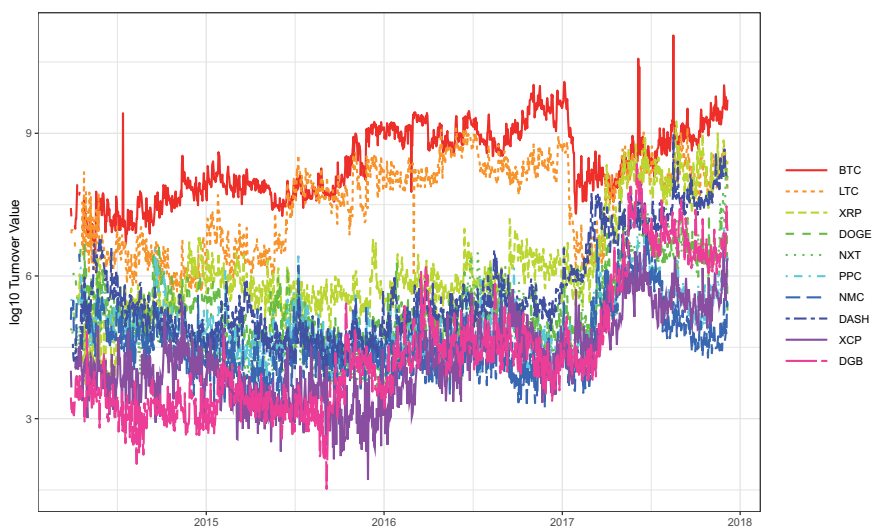


Figure 2 Common Logarithm of turnover value of top 10 CCs by their daily trading volume, BTC, LTC, XRP, DOGE, NXT, PPC, NMC, DASH, XCP, DGB.

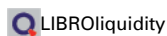


Table 2 Summary statistics of the trading volume of top 10 CCs

	BTC	LTC	XRP	DOGE	NXT	PPC	NMC	DASH	XCP	BLK
min	-5.50	-4.85	-3.65	-4.00	-2.06	-3.00	-2.95	-3.07	-2.64	-4.15
1st Quantile	-0.26	-0.36	-0.39	-0.37	-0.38	-0.50	-0.57	-0.35	-0.56	-0.50
mean	0.01	0.00	0.01	-0.00	0.01	0.00	-0.00	0.01	0.00	0.00
median	-0.03	-0.06	-0.02	-0.01	-0.02	-0.05	-0.04	-0.01	-0.04	-0.03
3rd Quantile	0.23	0.29	0.36	0.33	0.36	0.46	0.48	0.32	0.53	0.47
max	5.44	4.47	4.86	3.39	2.53	4.50	4.90	2.89	4.40	4.28
variance	0.26	0.41	0.48	0.41	0.38	0.63	0.84	0.33	0.83	0.67

median value is symmetric. In extreme cases, min and max values, this does not hold. The variance and extreme values suggest that fixing liquidity weights based on the mean turnover value could result in boundaries that are too high. Thus, an approach based on a robust measure, the median, will be employed. In the next section, we will introduce the portfolio optimization methods that we will use.

3 Constrained Portfolio Optimization

Markowitz (1952) introduced the theory of optimizing weights so that the variance of the portfolio would be minimized according to a certain target return. When the variance serves as a risk measure, this translates into risk minimization. Consider now N assets with T returns given by an $(N \times T)$ matrix X and let $\hat{\Sigma}$ be the estimated covariance

matrix of the respective assets. Then the Markowitz portfolio is defined as, [Härdle and Simar \(2015\)](#):

$$\begin{aligned} \min_w \quad & w^\top \hat{\Sigma} w \\ \text{s.t.} \quad & \mathbf{1}_N^\top w = 1, \mu \leq x^\top w \end{aligned} \quad (1)$$

where $w = (w_1, w_2, \dots, w_N)^\top$ are the weights on the assets, x is the $(N \times 1)$ vector of expected returns of the assets, $\mathbf{1}_N$ is an $(N \times 1)$ matrix (vector) with all elements equal to 1, and μ is the target return. The optimization problem is extended by a bound for each weight. The vector of constraints $a = (a_1, \dots, a_N)^\top$ with $a_i \in [0, \infty)$ for all $i = \{1, \dots, N\}$ is an $(N \times 1)$ vector and can be given (or estimated) upfront. Furthermore, an upper bound for the sum over the absolute values of the weights is introduced. Then, a constrained Markowitz portfolio is defined as

$$\min w^\top \hat{\Sigma} w \quad (2)$$

$$\begin{aligned} \text{s.t.} \quad & \mathbf{1}_N^\top w = 1, \mu \leq x^\top w, \\ & \|w\|_1 \leq c, |w_i| \leq a_i \forall i. \end{aligned} \quad (3)$$

The parameter c controls the amount of shortselling, $c \in [1, \infty)$. [Fan et al. \(2012\)](#) showed how the risk of the estimated portfolio is influenced by the choice of c while $a_i = \infty$ for all a_i . The estimation of $\hat{\Sigma}$ is crucial for the method, yet the huge dimensionality of the asset universe and limited data availability make the estimation of $\hat{\Sigma}$ difficult. Thus we employ the covariance estimator of [Ledoit and Wolf \(2004\)](#). It has been shown to be invertible, well-conditioned, and asymptotically more accurate than the sample covariance matrix. The estimator is a weighted average of the identity matrix and the sample covariance matrix. The identity matrix is a well-conditioned matrix and due to its being combined with the sample covariance matrix under a quadratic loss function, the resulting estimator has respective property and is more accurate than the sample covariance matrix, [Ledoit and Wolf \(2004\)](#). For more details, we refer to Section A.1 in the Appendix.

However, Markowitz portfolio optimization neglects the effect of higher moments when minimizing the risk. Due to the often occurring strong decreases in the CC market, portfolios optimized for Conditional Value-at-Risk (CVaR) will be employed to compare their performance with the Markowitz portfolio.

Define $y(w) = w^\top X$ as the returns of the portfolio with weights w and α as the probability level such that $0 < \alpha < 1$, the Value-at-Risk (VaR) is defined by

$$\text{VaR}_\alpha(w) = -\inf \{y | F(y|w) \geq \alpha\} \quad (4)$$

with $F(y|w)$ being the distribution function of the portfolio returns with weights w . $\text{VaR}_\alpha(w)$ is the corresponding α -quantile of the cdf, defining the loss to be expected $(\alpha \cdot 100)\%$ of the time. A negative sign is added to turn the negative return into a loss, which is defined on the positive domain. Considering the VaR only captures one quantile rather than the whole shape of the tails of the return distribution, it is silent on losses beyond that critical point. Overcoming the limitations of the VaR measure, the CVaR was introduced which measures the expected loss larger or equal to the $\text{VaR}_\alpha(w)$. Since CCs are at times subject to huge losses, the CVaR will give a more accurate impression of the investment at risk.

The Conditional Value-at-Risk is defined as, [Rockafellar and Uryasev \(2000\)](#),

$$CVaR_\alpha(w) = -\frac{1}{1-\alpha} \int_{y(w) \leq -VaR_\alpha(w)} yf(y|w)dy, \tag{5}$$

with $\frac{\partial}{\partial y} F(y|w) = f(y|w)$ the probability density function for the portfolio returns y with weights vector w . Thus $CVaR_\alpha(w)$ includes the expected value over the tail of the pdf left of the $VaR_\alpha(w)$.

The optimization problem is, then,

$$\min CVaR_\alpha(w) \tag{6}$$

$$\begin{aligned} \text{s.t. } \mathbf{1}_N^\top w &= 1, \mu \leq x^\top w, \\ \|w\|_1 &\leq c, |w_i| \leq a_i \forall i. \end{aligned} \tag{7}$$

So far we have taken into consideration the parameter c . However, short-selling is still rare in the CC market. Thus, the value of c is an issue. First exchanges started to offer the possibility of short selling for larger CCs, and the launch of Bitcoin Futures allows it too. But it is still not possible for most CCs to be shorted. Due to the inability to sell short in the CC market, the exposure is set to $c = 1$, which produces a no short-sell constraint combined with $\mathbf{1}_N^\top w = 1$. Surely it is only a matter of time until short selling is common in the CC market too. In that case, our optimization problem could be amended to allow for it, which would enable one to use hedging effects to decrease the historical risk. However, this approach can cause extreme weights on single assets, so that the position is not tradable in a real market situation.

4 LIBRO

So far, the actual measure for liquidity for the constraints was not further explained. Yet this is a central point of this study, because CCs have far lower daily trading amounts than traditional financial assets, causing a liquidity problem for any portfolio construction. To address this issue, one tries to avoid holding too many illiquid assets by employing weight constraints $|w_i| \leq a_i$ for all $i = 1, \dots, N$.

Many different liquidity measures have been proposed in the literature, tackling either one aspect of liquidity or several aspects at the same time, [Wyss \(2004\)](#). In the context of this paper, we are interested in

1. being able to trade the assets on the reallocation date,
2. being able to sell or buy between two reallocation dates, if necessary.

Naturally, a more liquid asset should be allowed to have a higher weight in the portfolio.

Our data set consists of daily price and turnover value observations, which enables us to use the turnover value as a proxy for liquidity. An even better measure would be Limit Order Book based measures since they allow a deeper look into the behaviour of the markets. As we do not possess of a sufficient history of these data to run an analysis, for the moment the liquidity measures using such information are not applicable. Since the time period of interest for a trading action is one day, daily closing data for the standard assets

and CCs are going to be used. The trading volume of asset i at date t , measured in USD, is defined to be¹

$$TV_{it} = p_{it} \cdot q_{it} \quad (8)$$

where p_{it} is the closing price of asset i at date t , and q_{it} is the number of shares of asset i traded at date t . The liquidity of asset i in a sample with time length T can be measured by, e.g. the mean or median of the daily trading volume. The mean has the disadvantage that days with outliers in the trading volume will distort the liquidity measure. The median is a robust measure and therefore better suited for this situation. We define the sample median of the trading volume to be

$$TV_{i,m} = \frac{1}{2} (TV_{i,u} + TV_{i,l}) \quad (9)$$

where $TV_{i,u} = TV_{i, \lceil \frac{T+1}{2} \rceil}$, $TV_{i,l} = TV_{i, \lfloor \frac{T+1}{2} \rfloor}$ and $\lceil \cdot \rceil$, $\lfloor \cdot \rfloor$ indicate the ceiling and floor operators respectively.

Next we construct the liquidity bound. Taking into account that investors enter the market with different sizes of portfolios, this has to be reflected in the liquidity bound. For example, an investor with a portfolio of 100 USD will not be as strongly affected by low liquidity as an investor with 10,000,000 USD. Denote by M the total amount we are going to invest and recall that w_i , $i = 1, \dots, N$ is the weight for asset i , so Mw_i is the market value of the position in asset i . Hence the constraint on w_i concerning the liquidity of asset i is

$$Mw_i \leq TV_{i,m} \cdot f_i, \quad (10)$$

where f_i is a factor controlling the maximum ratio of the position in asset i to its median trading volume, i.e. liquidity. The larger the f_i , the more bullish the investor is on asset i , and the more likely the position in asset i will suffer from a low liquidity problem when clearing or rebalancing. For example, setting $f_i = 0.1$ corresponds to a position in asset i not larger than 10% of the median trading amount of asset i . Dividing both sides of Equation (10) by M yields the bound for w_i :

$$w_i \leq \frac{TV_{i,m} \cdot f_i}{M} = \hat{a}_i.$$

Hence, the Markowitz portfolio optimization framework we will use in this article is

$$\begin{aligned} & \min w^\top \hat{\Sigma} w \\ \text{s.t. } & \mathbf{1}_N^\top w = 1, \quad \mu \leq x^\top w, \quad \|w\|_1 = 1, \\ & w_i \leq \frac{1}{M} \cdot \widehat{Li} q_i = \hat{a}_i \quad \forall i, \end{aligned} \quad (11)$$

4 Note this is an approximation of the actual trading volume. The trading volume of asset i on date t , measured in USD, is calculated as $TV_{it} = \sum_{j=1}^{N_t} p_{it,j} \cdot q_{it,j}$, where N_t is the number of times that asset i is traded on date t , and $p_{it,j}$ and $q_{it,j}$ are the trading price and number of traded shares for the j th trade, $j = 1, \dots, N_t$. However, due to the lack of sufficient historical data on the per-trade data of the CCs, the theoretical definition is not available for a sufficiently large sample size.

where $\widehat{Liq} = (TV_{1,m} \cdot f_1, \dots, TV_{N,m} \cdot f_N)^\top$. The CVaR optimization problem can thus be expressed as

$$\begin{aligned} & \min \text{CVaR}_\alpha(w) \\ & \text{s.t. } \mathbf{1}_N^\top w = 1, \mu \leq \mathbf{x}^\top w, \|w\|_1 = 1, \\ & \quad w_i \leq \frac{1}{M} \cdot \widehat{Liq}_i = \hat{a}_i \forall i. \end{aligned} \tag{12}$$

5 Application

For the application we treat three kinds of settings. First we perform an in-sample analysis without liquidity constraints to investigate whether including CCs in traditional financial portfolios increases the risk–return trade-off. Furthermore, we intend to determine whether including alt-coins (i.e. CCs other than Bitcoin) is profitable and confirm whether the introduction of liquidity constraints is necessary. Second, when liquidity constraints are included in the in-sample analysis, we set $f_i = 0.01$ for all $i = 1, \dots, N$, that is, we assume that our position in a certain asset cannot exceed 1% of its daily trading volume. This is a quite conservative setting, and investors who want to be more aggressive can enlarge this factor. We choose to be conservative because the CC market, especially the alt-coin market, exhibits swings in its daily trading volume. Thus we are securing our portfolio choice against these swings. Furthermore, trading on the entire daily trading volume would be a rather strict anticipation. In both cases, the portfolio and weights are formed over the entire time period. For the out-of-sample portfolio formation, we choose an extending window approach. The initial portfolio weights are derived for the time period from April 22, 2014 until December 31, 2014. Due to the limited data availability in the CC market, the April 2014 data are not omitted, so as to enhance the estimation. Two kinds of portfolio rebalancing frequencies are employed: weekly and monthly, while the underlying data period is extended. Thus for the monthly case, on the next re-evaluation date, which is February 1, 2015, the derivation period is extended to the period from April 22, 2014 until January 1, 2015. For the portfolio formation under CVaR, the quantile level in all cases is chosen to be $\alpha = 0.05$. For the liquidity constraint, we maintain the same setting for the whole of the analysis, which is either unbounded (without liquidity constraint), or bounded with investment amount equal to one of either 1.0×10^5 , 1.0×10^6 , or 1.0×10^7 US dollars, see Equations (11) and (12). For selecting the target return μ , the Sharpe Ratio is maximized for the Markowitz portfolio, and the Return-to-CVaR Ratio for the CVaR portfolio. The median over the trading volume, necessary for the constraints, is chosen in-sample over the entire sample and out-of-sample over the extending window. We compare two different data sets consisting of traditional assets by adding CCs to them: one based solely on stocks from the S&P 100 (S), and one on stocks plus US Bonds and Commodities (SBC). We denote the S and SBC plus CCs portfolios as S-CC and SBC-CC, respectively.

5.1 In-sample portfolio formation

5.1.1 Without liquidity bounds

The cumulative returns of the portfolios formed by the S&P 100 component stocks and SBC with or without CCs for both definitions of risk are shown in Figure 3. One can see

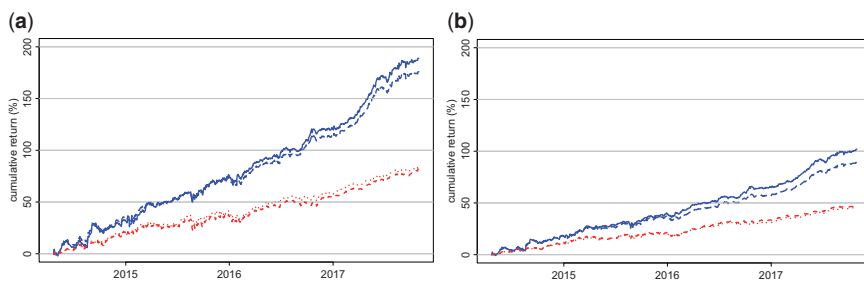


Figure 3 The panels 3a and 3b indicate the cumulative return performance of S/S-CC and SBC/SBC-CC portfolios. The dashed (---) and dotted (⋯) lines indicate the S/SBC portfolios formed with CVaR and Markowitz method respectively, while the solid (—) and long-dashed (— —) lines correspond to the S-CC/SBC-CC portfolios formed with CVaR and Markowitz method. The coloured version is available online.

(a) A comparison of the cumulative returns of the S-CC portfolios. (b) A comparison of the cumulative returns of the SBC-CC portfolios.

that the improvement in the returns is remarkable and consistent throughout the sample. Starting from the very beginning, the portfolio with CCs (S-CC/SBC-CC) outperforms the one without, with the difference becoming larger and larger as time goes on. At the end of the sample, the S-CC Markowitz portfolio gives a cumulative return of 173.3%, while the S portfolio ends at 85.3%, only one-half of the former. However, comparing the S-CC and the SBC-CC, the former outperforms the latter, with double the cumulative return. This is because the latter reaches the maximum Sharpe ratio at a lower target return. In fact, the optimal Sharpe ratio of the latter portfolio is always higher than those of the former portfolio, see Tables 3 and 4. The case is similar for the CVaR portfolio, with the difference being even larger. It is interesting to observe that the portfolio returns for the CVaR portfolio are higher than those for the Markowitz, since this suggests that the returns increase due to holding CCs rises more significantly than does the risk induced from the tails. The summary statistics of the Markowitz portfolio returns are given in Tables 3 and 4, where the S column is the one for the S portfolio and the SBC column is that for the SBC portfolio, and the “unbounded” column is for the S-CC and SBC-CC without implementing the liquidity constraints. As an example of the results of both strategies (Markowitz and CVaR) with S and SBC, we will now look in more detail at the S Markowitz portfolio results, Table 3. Coinciding with the previous finding, the annualized average returns of the S-CC (48%) is twice that of the S portfolio (23%). Though the volatility is a bit higher, the whole risk–return trade-off has improved after adding CCs, since the Sharpe Ratio increases from 0.12 to 0.18. The higher moments of portfolio returns are also improved: after adding CCs, the skewness changes from -0.31 to 0.14 , and the kurtosis decreases from 5.50 to 4.55. The Maximum drawdown, which measures the downside risk of the portfolio, stays the same. Similar results can be observed when forming a CVaR portfolio, see Tables 3 and 4. When using the SBC portfolio, the average returns roughly halve, while the standard deviation shrinks by more than one-half, resulting in an improvement in the Sharpe ratio.

Even though these are striking results, a check on the weights suggests that a liquidity constraint aiming at lowering the weight on illiquid assets is needed. The weights different

Table 3 Summary statistics of in-sample S/S-CC Markowitz/CVaR portfolio return. All indices are calculated using daily returns. Ann.Ret and Ann.STD indicate annualized mean return and standard deviation, ρ refers to the autocorrelation parameter

		S	S-CC			
			unbounded	$M = 1.0 \times 10^5$	$M = 1.0 \times 10^6$	$M = 1.0 \times 10^7$
Markowitz	Ann.Ret	0.23	0.48	0.33	0.27	0.27
	Ann.STD	0.13	0.17	0.14	0.13	0.13
	Sharpe-ratio	0.12	0.18	0.15	0.14	0.13
	Skewness	-0.31	0.14	-0.24	-0.54	-0.55
	Kurtosis	5.50	4.55	5.20	5.70	5.68
	Max drawdown	0.10	0.10	0.11	0.11	0.11
	Auto correlation	-0.02	-0.04	-0.01	0.00	0.01
CVaR	Ann.Ret	0.22	0.52	0.32	0.26	0.26
	Ann.STD	0.12	0.18	0.14	0.12	0.12
	Sharpe-ratio	0.12	0.18	0.15	0.14	0.13
	Skewness	-0.27	0.28	-0.00	-0.47	-0.48
	Kurtosis	5.50	4.81	4.89	5.60	5.54
	Max drawdown	0.10	0.09	0.10	0.10	0.10
	Auto correlation	-0.03	-0.05	-0.02	-0.02	-0.02

Table 4 Summary statistics of in-sample SBC/SBC-CC Markowitz/CVaR portfolio return. All indices are calculated using daily returns. Ann.Ret and Ann.STD indicate annualized mean return and standard deviation, ρ refers to the autocorrelation parameter

		SBC	SBC-CC			
			Unbounded	$M = 1.0 \times 10^5$	$M = 1.0 \times 10^6$	$M = 1.0 \times 10^7$
Markowitz	Ann.Ret	0.13	0.25	0.18	0.15	0.15
	Ann.STD	0.06	0.08	0.07	0.06	0.06
	Sharpe-ratio	0.14	0.19	0.17	0.16	0.15
	Skewness	-0.34	0.01	-0.27	-0.46	-0.48
	Kurtosis	4.95	4.33	4.81	5.04	5.20
	Max drawdown	0.04	0.05	0.05	0.05	0.05
	Auto correlation	0.03	-0.01	0.01	0.03	0.04
CVaR	Ann.Ret	0.13	0.28	0.19	0.16	0.15
	Ann.STD	0.06	0.09	0.07	0.06	0.06
	Sharpe-ratio	0.13	0.19	0.16	0.16	0.15
	Skewness	-0.26	0.29	0.03	-0.39	-0.38
	Kurtosis	4.77	4.59	4.69	4.90	4.86
	Max drawdown	0.05	0.05	0.05	0.05	0.05
	Auto correlation	0.01	-0.02	-0.02	0.00	0.01

from 0 given to CCs in the unbounded case for the Markowitz and CVaR portfolios are shown in Tables 5–8.

The key information conveyed from the tables is that the more liquid (traded) CCs are not generally given a larger weight. The biggest weight on a CC is given to NLG (Gulden),

Table 5 Weights (in %) given to CCs in in-sample S-CC Markowitz portfolios. Only CCs that have a positive weight in at least one portfolio are shown in the Table. The “unbounded” refers to the portfolio formed without liquidity constraint included; the remaining three are all formed under liquidity constraint with different investment amounts M . Underline/bold indicates that the weight equals to its liquidity upper bound

	Unbounded	$M = 1.0 \times 10^5$	$M = 1.0 \times 10^6$	$M = 1.0 \times 10^7$
BTC	0.00	5.37	8.28	9.01
XRP	3.72	4.53	<u>0.80</u>	<u>0.08</u>
DASH	3.17	<u>2.13</u>	<u>0.21</u>	<u>0.02</u>
DGB	0.95	<u>0.13</u>	<u>0.01</u>	<u>0.00</u>
VTC	0.18	<u>0.10</u>	<u>0.01</u>	<u>0.00</u>
NLG	5.30	<u>0.02</u>	<u>0.00</u>	<u>0.00</u>
FLO	1.53	<u>0.01</u>	<u>0.00</u>	<u>0.00</u>
RBY	0.84	<u>0.03</u>	<u>0.00</u>	<u>0.00</u>
NOTE	0.18	<u>0.01</u>	<u>0.00</u>	<u>0.00</u>
CBX	0.12	<u>0.00</u>	<u>0.00</u>	<u>0.00</u>
total	16.00	12.32	9.32	9.11

Table 6 Weights (in %) given to CCs in in-sample SBC-CC Markowitz portfolios. Only CCs that have a positive weight in at least one portfolio are shown in the Table. The “unbounded” refers to the portfolio formed without liquidity constraint included; the remaining three are all formed under liquidity constraint with different investment amounts M . Underline/bold indicates that the weight equals to its liquidity upper bound

	Unbounded	$M = 1.0 \times 10^5$	$M = 1.0 \times 10^6$	$M = 1.0 \times 10^7$
BTC	0.00	2.11	3.14	3.68
XRP	1.47	1.77	<u>0.80</u>	<u>0.08</u>
DASH	1.34	1.41	<u>0.21</u>	<u>0.02</u>
DGB	0.43	<u>0.13</u>	<u>0.01</u>	<u>0.00</u>
VTC	0.13	<u>0.10</u>	<u>0.01</u>	<u>0.00</u>
NLG	2.26	<u>0.02</u>	<u>0.00</u>	<u>0.00</u>
FLO	0.64	<u>0.01</u>	<u>0.00</u>	<u>0.00</u>
RBY	0.37	<u>0.03</u>	<u>0.00</u>	<u>0.00</u>
total	6.65	5.57	4.18	3.79

which has a medium level of liquidity. However, Bitcoin (BTC), the CC that has the largest trading volume, is given a zero weight. Considering the CCs in the unbounded S-CC Markowitz portfolio, their weights account for 16.3% of the whole portfolio. The top three CCs by weight are NLG (Gulden), XRP (Ripple), and DASH (Dash coin), at 5.3%, 3.7% and 3.2% respectively. This shows that alt-coins are more appealing than BTC in terms of variance minimization, at least during the period covered by this paper. Furthermore, the inclusion of liquidity constraints appears necessary, since the highly weighted CCs have, partly, low liquidity compared to BTC. For instance, with an investment amount of $M = 1.0 \times 10^6$ US dollars, and considering S-CC, one needs to hold a position of 5, 300 US

Table 7 Weights (in %) given to CCs in in-sample S-CC CVaR portfolios. Only CCs that have a positive weight in at least one portfolio are shown in the Table. The “unbounded” refers to the portfolio formed without liquidity constraint included; the remaining three are all formed under liquidity constraint with different investment amounts M . Underline/bold indicates that the weight equals to its liquidity upper bound

	Unbounded	$M = 1.0 \times 10^5$	$M = 1.0 \times 10^6$	$M = 1.0 \times 10^7$
BTC	0.00	3.82	8.71	9.61
XRP	3.30	5.57	<u>0.80</u>	<u>0.08</u>
DASH	3.72	<u>2.15</u>	<u>0.22</u>	<u>0.02</u>
DGB	0.00	<u>0.12</u>	<u>0.01</u>	<u>0.00</u>
NLG	6.22	<u>0.02</u>	<u>0.00</u>	<u>0.00</u>
FLO	2.79	<u>0.01</u>	<u>0.00</u>	<u>0.00</u>
RBY	1.11	<u>0.03</u>	<u>0.00</u>	<u>0.00</u>
MAX	0.50	<u>0.01</u>	<u>0.00</u>	<u>0.00</u>
CBX	0.28	0.00	<u>0.00</u>	<u>0.00</u>
ZEIT	0.12	<u>0.00</u>	<u>0.00</u>	<u>0.00</u>
total	18.03	11.73	9.74	9.71

Table 8 Weights (in %) given to CCs in in-sample SBC-CC CVaR portfolios. Only CCs that have a positive weight in at least one portfolio are shown in the Table. The ‘unbounded’ refers to the portfolio formed without liquidity constraint included; the remaining three are all formed under liquidity constraint with different investment amounts M . Underline/bold indicates that the weight equals to its liquidity upper bound

	Unbounded	$M = 1.0 \times 10^5$	$M = 1.0 \times 10^6$	$M = 1.0 \times 10^7$
BTC	0.00	1.27	4.38	4.73
XRP	1.32	2.33	<u>0.80</u>	<u>0.08</u>
DASH	1.80	<u>2.15</u>	<u>0.22</u>	<u>0.02</u>
DGB	0.00	<u>0.12</u>	<u>0.01</u>	<u>0.00</u>
NLG	3.18	<u>0.02</u>	<u>0.00</u>	<u>0.00</u>
FLO	1.35	<u>0.01</u>	<u>0.00</u>	<u>0.00</u>
RBY	0.53	<u>0.03</u>	<u>0.00</u>	<u>0.00</u>
MAX	0.26	<u>0.01</u>	<u>0.00</u>	<u>0.00</u>
total	8.44	5.94	5.40	4.84

dollars in NLG, which is equal to 47.3% of its average daily trading volume. Taking into consideration the price impact, this position is neither easy to obtain nor to clear. **Table 7** shows the weights for the CVaR $p_{1.0 \times 10^6}$ portfolio, providing similar results. In fact, the influence of the CCs and NLG in particular is even higher. However, when relying on the SBC portfolio, the weights given to CCs shrink markedly, **Table 8**. It seems the inclusion of bonds and commodities shifts the Mean–Variance and Mean–CVaR Frontier so strongly that the resulting portfolio favors having fewer CCs. However one can see that this harms the returns achieved from the portfolio.

5.1.2 Including liquidity constraints

The cumulative returns of portfolios with liquidity constraints included are shown in Figure 4, with the summary statistics of these returns in Table 3. When liquidity constraints are imposed, the cumulative returns shift downward compared to those without liquidity constraints. The larger the amount to invest, the lower the cumulative returns. This is not surprising, since adding a liquidity constraint makes the global optimal Sharpe ratio point and Return-to-CVaR point unreachable, and the larger the investment amount, the tighter the liquidity constraint. Hence, the constrained optimal Sharpe ratio point is further from the unconstrained one. When the investment amount is set to 1.0×10^5 or US dollars, the cumulative returns of the Markowitz portfolios still outperform the one formed by only traditional assets throughout the sample, and by the end of the sample period, the cumulative returns are 128.5% and 117.0% respectively, still 43.2% and 31.7% higher than the S portfolio, by considering this setting as an example. When the investment amount is increased to 1.0×10^7 US dollar, the portfolio does not outperform the one containing only stocks until 2017, however it ends at 98.8%, still 13.5% higher. For the CVaR portfolios, the constrained portfolios still outperform the S and SBC portfolios, however only after 2017 and with excess returns of 12.2% and 5.8%, respectively.

Summary statistics also favour the portfolios with CCs, see the last three columns of Tables 3 and 4. When the investment amounts are set to 1.0×10^5 , 1.0×10^6 or 1.0×10^7 US dollars, the liquidity constrained portfolios have higher average return and higher Sharpe-ratio than the S/SBC portfolios, with, however, a slightly higher standard deviation and almost the same downside risk. Kurtosis and Skewness show a mixed picture. Mostly the absolute value of either increases, making it less of a Gaussian distribution. Yet this observation is not surprising, taking into account the strong deviations of CC return series from normality. Overall, the summary statistics provide support for adding CCs to portfolios consisting of the chosen traditional assets.

To see how the liquidity constraint affects the optimization procedure, we turn to an analysis of the weights. For the comparison, we focus on the weights given to the CCs, since the impact of CCs on the portfolio is of greater interest and no liquidity bound on the traditional assets is binding in any situation. The weights are shown in Tables 5–8, where we show the weights for the CCs included in the portfolio in any considered situation. The first column shows the weights when no liquidity constraint is implemented, while the remaining three columns show those when liquidity constraints are included with the three different investment amounts. The weights coloured red indicate that its liquidity upper bound is binding, i.e. the weight given to this CC just equals its liquidity upper bound. Before implementing the liquidity upper bound, the CCs account for 16% of the total position for the S-CC Markowitz portfolio, with the largest weight 5.3% given to NLG, and zero weight to Bitcoin. After including a liquidity constraint, the total weights on the CCs decrease to 12.3%, 9.3% and 9.1% as the investment amount increases from 1.0×10^5 to 1.0×10^7 US dollars. This is not surprising, since the liquidity upper bounds limit the weight given to the CCs. When the investment amount equals 1.0×10^5 US dollars, the lower 8 CCs are binding, including NLG, which only has a weight of 0.02%. When the investment increases to 1.0×10^6 and 1.0×10^7 US dollars, only the bound on Bitcoin is not binding.

As the liquidity constraints tighten, the weight on Bitcoin increases from 5.4% to 9.0%, which shows the great investment potential of the Bitcoin market, since it can account for about 9% of the portfolio when formed together with S&P 100 stocks, while not being

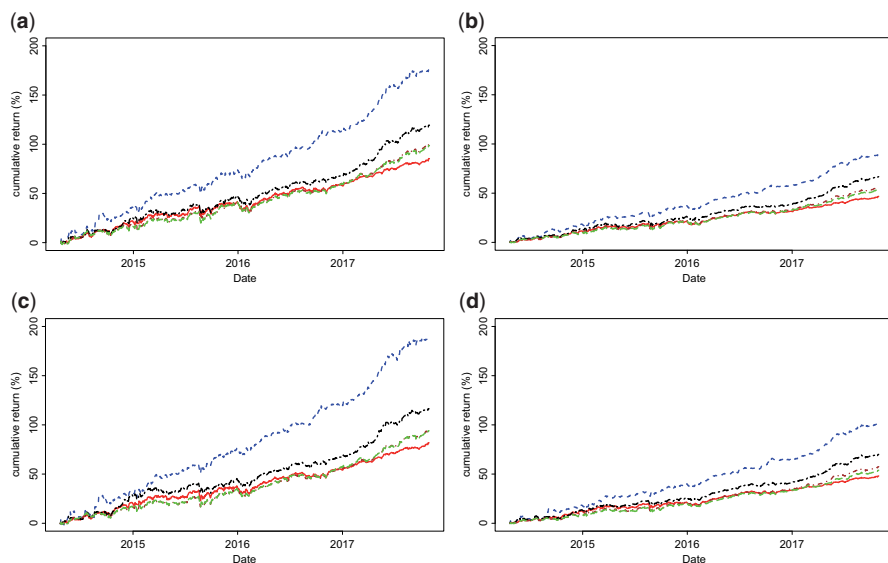


Figure 4 The solid and dashed lines indicate the cumulative return performance of Markowitz and CVaR portfolios. The solid line (—) and dashed line (- -) stand for S/SBC and S-CC/SBC-CC without liquidity constraints respectively. The remaining 3 portfolios are S-CC/SBC-CC ones containing the bounds $M = 1 \times 10^5$ USD (— · —), $M = 1 \times 10^6$ USD (- - -), $M = 1 \times 10^7$ USD (- · -). The coloured version is available online.

 LIBROinsample

(a) In-sample cumulative returns of S-CC Markowitz portfolios (b) In-sample cumulative returns of SBC-CC Markowitz portfolios (c) In-sample cumulative returns of S-CC CVaR portfolios (d) In-sample cumulative returns of SBC-CC CVaR portfolios

constrained. When considering the CVaR portfolio, the cumulative weight on the CCs in the unbound case is 18%, thus even larger, however in the constrained cases this shrinks to 9.71%. For the SBC portfolios, with either definition of risk (Markowitz or CVaR), the weight given to CCs is considerably lower, so the portfolios favour bonds and commodities. Still no weight is given to BTC in the unbounded case. However, when optimizing with liquidity constraints, BTC receives a weight and the constraints become active on various CCs. This shows that the constraints are necessary for achieving the goal set in this study, namely constructing portfolios with CCs in which the positions in the CCs can be easily cleared. After illustrating the potential effect of CCs on portfolio performance and the effect of liquidity constraints on the in-sample analysis, we turn next to an out-of-sample study to investigate the performance of the portfolios under pseudo-real conditions.

5.2 Out-of-sample portfolio formation

After having analysed the potential of CCs for enhancing the performance of a combined portfolio, an out-of-sample analysis remains to justify their applicability in real-world investments. The S portfolio and the SBC portfolio will be constructed with monthly rebalanced weights, calculated using all the sample data before the rebalancing day. First, the

portfolio is formed on January 1, 2015 and held to January 31, 2015, with weights calculated using the sample from April 22, 2014 to December 31, 2014. Then, data from April 22, 2014 to January 31, 2015 is used to calculate the new weights, and the portfolio is rebalanced accordingly at February 1, 2015. For subsequent periods, the portfolio will be rebalanced at the first day of each month, with the weights calculated using all the sample data before that day. We choose this extending window approach to calculate the rebalancing weights due to the limited amount of data in the sample.

The cumulative returns of the Markowitz portfolios are illustrated in [Figure 5](#), where panel 5a shows S/S-CC portfolios, and panel 5c shows S/SBC-CC portfolios. For each panel, it is obvious that the S-CC/SBC-CC portfolio, no matter whether liquidity bounded or not, outperforms its counterpart without CCs at the end of the sample. For the S-CC portfolio, when no liquidity constraint is applied, the cumulative returns exceed the S portfolio at the very beginning of 2015, with the difference continuing to increase until the end of the sample. In the whole out-of-sample period, the cumulative returns of the S-CC portfolio without liquidity constraints is over 80%, while that of the S portfolio is 15.5%, a quite substantial improvement. The SBC-CC portfolio without liquidity constraints outperforms the SBC portfolio from March 2016 onwards, and the difference keeps enlarging in the remaining periods. At the end of the sample, the cumulative returns of the SBC-CC portfolio reach 37.4%, which is substantially higher than the 12.1% of the SBC portfolio. The big improvement of the S-CC/SBC-CC portfolio over the S/SBC one indicates the huge potential investment gain that can be obtained by including CCs in a portfolio. However, as stated in the in-sample cases, only when these improvements persist when liquidity constraints are included, can one infer that the profits are feasible in practice.

Now comes the situation when liquidity constraints are included. In [Figure 5](#), we label the cumulative returns calculated with liquidity constraints with black, brown and green, indicating an investment amount equal to 1×10^5 , 1×10^6 , and 1×10^7 US dollars, respectively. For all three investment amounts, the S-CC portfolios exceed the S one starting from March of 2016, and the difference does not become large until 2017. At the end of the sample, the liquidity bounded portfolios with investment amounts 1×10^5 , 1×10^6 , 1×10^7 US dollars end with a cumulative return equal to 40.0%, 28.6% and 29.2%, which is 24.5%, 13.1% and 13.7% higher than the pure stock portfolio. When bond and commodity indexes are included, the liquidity bounded cumulative return under an investment amount of 1×10^5 , 1×10^6 , 1×10^7 US dollars reaches 21.2%, 17.7% and 16.4%, all of which outperform the SBC portfolio. All in all, adding CCs into portfolios is profitable even after controlling for low liquidity by imposing constraints.

Summary statistics also favour the portfolios with CCs added, see [Tables 9](#) and [10](#). In either the S-CC or SBC-CC case, the portfolios with CCs always dominate the one without, regarding returns and Sharpe ratio. When no liquidity bound is applied or when the investment amount is 1×10^5 US dollars, the portfolio show less negative skewness and less heavy tails. Although the skewness and kurtosis may get worse under a tighter liquidity constraint when the investment amount gets larger, the maximum drawdown improves after CCs are added, which is somehow surprising, since the CC market is considered highly risky. Interestingly, the mean returns on SBC/SBC-CC are lower than for S/S-CC, yet in combination with a lower volatility as well. The SBC-CC portfolio outperforms S-CC in two ways: first, it only has about one-half the max drawdown, which is a substantial

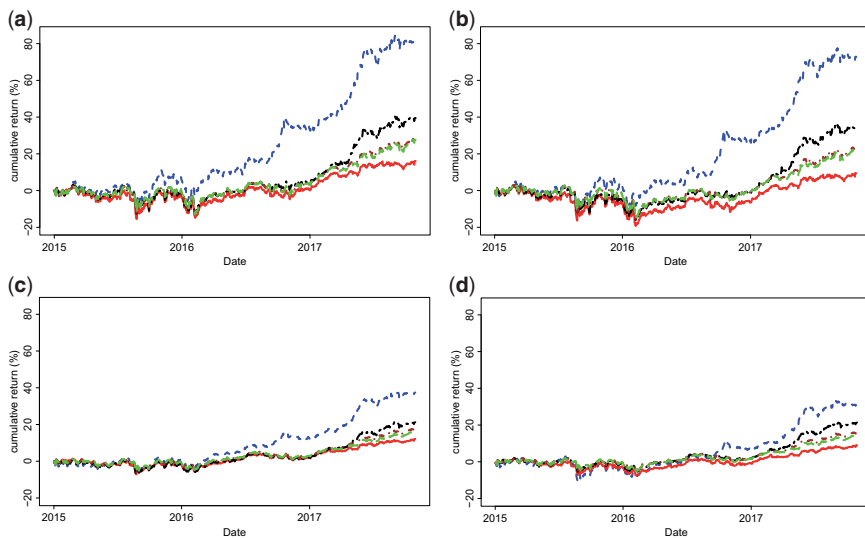


Figure 5 Out-of-sample cumulative returns with monthly and weekly adjusted Markowitz portfolios. The solid line (—) and dashed line (- -) stand for S/SBC and S-CC/SBC-CC without liquidity constraints respectively. The remaining 3 portfolios are S-CC/SBC-CC ones containing the bounds $M = 1 \times 10^5$ USD (■ ■ ■), $M = 1 \times 10^6$ USD (— —), $M = 1 \times 10^7$ USD (— —). The coloured version is available online.

(a) Out-of-sample cumulative returns for monthly adjusted S-CC Markowitz portfolios (b) Out-of-sample cumulative returns for weekly adjusted S-CC Markowitz portfolios (c) Out-of-sample cumulative returns for monthly adjusted SBC-CC Markowitz portfolios (d) Out-of-sample cumulative returns for weekly adjusted SBC-CC Markowitz portfolios

Table 9 Summary statistics of out-of-sample monthly rebalanced S/S-CC Markowitz/CVaR portfolio return. All indices are calculated using daily returns. Ann.Ret and Ann.STD indicate annualized mean return and standard deviation, ρ refers to the autocorrelation parameter

		S-CC				
		S	Unbounded	$M = 1.0 \times 10^5$	$M = 1.0 \times 10^6$	$M = 1.0 \times 10^7$
Markowitz	Ann.Ret	0.05	0.28	0.13	0.10	0.09
	Ann.STD	0.13	0.19	0.15	0.14	0.14
	Sharpe-ratio	0.03	0.09	0.06	0.04	0.04
	Skewness	-0.38	0.09	-0.29	-0.44	-0.44
	Kurtosis	5.98	5.23	5.95	6.20	6.15
	Max drawdown	0.17	0.17	0.16	0.16	0.16
	Auto correlation	0.03	-0.02	0.03	0.04	0.04
CVaR	Ann.Ret	0.07	0.28	0.12	0.08	0.16
	Ann.STD	0.14	0.21	0.16	0.17	0.18
	Sharpe-ratio	0.03	0.08	0.05	0.03	0.06
	Skewness	-0.21	0.12	-0.48	-0.31	-0.11
	Kurtosis	5.83	5.13	6.39	9.36	9.44
	Max drawdown	0.16	0.19	0.18	0.25	0.23
	Auto correlation	-0.00	-0.05	0.04	0.10	0.12

Table 10 Summary statistics of out-of-sample monthly rebalanced SBC/SBC-CC Markowitz/CVaR portfolio return. All indices are calculated using daily returns. Ann.Ret and Ann.STD indicate annualized mean return and standard deviation, ρ refers to the autocorrelation parameter

		SBC-CC				
		SBC	Unbounded	$M = 1.0 \times 10^5$	$M = 1.0 \times 10^6$	$M = 1.0 \times 10^7$
Markowitz	Ann.Ret	0.04	0.13	0.07	0.06	0.06
	Ann.STD	0.07	0.10	0.08	0.07	0.07
	Sharpe-ratio	0.04	0.08	0.06	0.06	0.05
	Skewness	-0.33	-0.10	-0.34	-0.37	-0.38
	Kurtosis	5.98	5.63	6.33	6.31	6.40
	Max drawdown	0.08	0.10	0.08	0.07	0.07
	Auto correlation	0.08	0.03	0.07	0.08	0.08
CVaR	Ann.Ret	0.03	0.17	0.07	0.05	0.11
	Ann.STD	0.07	0.11	0.11	0.14	0.14
	Sharpe-ratio	0.03	0.10	0.04	0.02	0.05
	Skewness	-0.36	0.30	-0.94	-0.92	-0.51
	Kurtosis	5.47	5.87	14.16	14.93	15.48
	Max drawdown	0.11	0.06	0.14	0.22	0.23
	Auto correlation	0.05	-0.02	0.09	0.19	0.22

decrease in the downside risk; second, at larger investment amounts, it has higher Sharpe ratios.

Looking at the weights of the CCs gives an answer for how CCs influence the performance of the portfolios. For the monthly rebalanced weights, see [Figures 7 and 8](#). Obviously the CCs taking strong positions in the portfolios in the unbounded case, for example, NLG, which even reaches 8% in the fourth quarter of 2016, get ruled out when liquidity constraints are added. Furthermore, by the red rectangles we indicate that the weight reaches its respective upper bound on that reallocation date. The constraints are mostly in place, giving support for their introduction into the method. For the S-CC Markowitz portfolios, when the investment amount equals 1.0×10^5 US dollars, 6 CCs are included over time. When the investment amount increases to 1.0×10^6 and 1.0×10^7 US dollars, only 3 and 1 CCs are included over time, respectively, and BTC becomes the only one that is not affected by the liquidity bound. A further observation is the absence of Bitcoin, the largest and most liquid CC, in the unbounded portfolio. Yet with liquidity constraints, it is always part of the portfolio and does not reach its upper bound. Additionally, it becomes apparent that Bitcoin and also Ripple receive higher weights in 2017, due to their better Sharpe ratios, thus adding value to the portfolios due to the strong gains in the CC market in this period.

Shifting our analysis to the CVaR portfolios provides partly different observations. Still the unbounded portfolios with CCs clearly outperform those without CCs, and in the constrained portfolios, the cumulative returns perform better at the end of the sample, [Figure 6](#). However, for the constrained case with investment amounts $M = 1.0 \times 10^5$ and $M = 1.0 \times 10^6$ US dollars, the improvements are not consistent throughout the sample: the pure S portfolio still outperforms until March of 2017 and August of 2017. It is interesting to observe that in this case the portfolio having the highest investment amount, and therefore the strongest constraints, performs the best among the constrained ones, with a stable

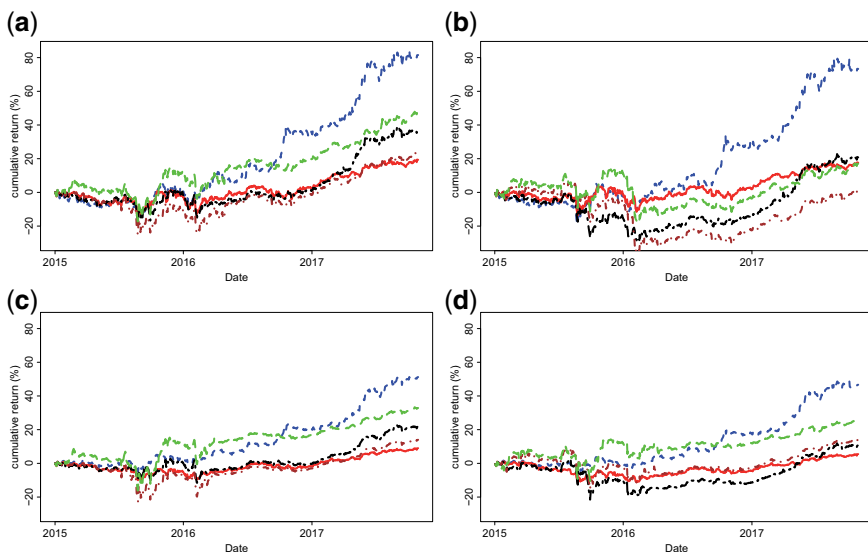


Figure 6 Out-of-sample cumulative returns with monthly and weekly adjusted CVaR portfolios. The solid line (—) and dashed line (---) stand for S/SBC and S-CC/SBC-CC without liquidity constraints, respectively. The remaining 3 portfolios are S-CC/SBC-CC ones containing the bounds $M = 1 \times 10^5$ USD (—■—), $M = 1 \times 10^6$ USD (---), $M = 1 \times 10^7$ USD (—). The coloured version is available online.

(a) Out-of-sample cumulative returns for monthly adjusted S-CC CVaR portfolios (b) Out-of-sample cumulative returns for weekly adjusted S-CC CVaR portfolios (c) Out-of-sample cumulative returns for monthly adjusted SBC-CC CVaR portfolios (d) Out-of-sample cumulative returns for weekly adjusted SBC-CC CVaR portfolios

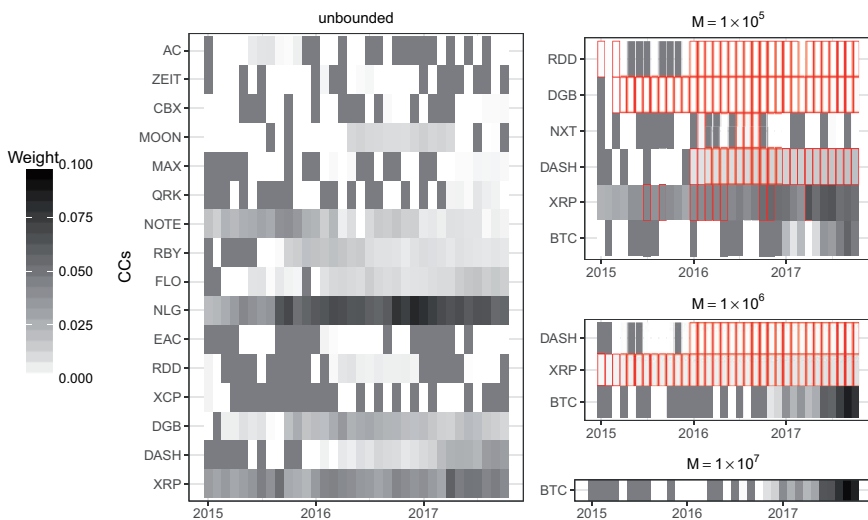


Figure 7 Weights given to CCs for S-CC Markowitz portfolios at each monthly rebalancing date under the three different investment amounts. Only CCs that have a non-zero weight on at least one rebalancing date are given. The darker the color, the larger the weight. Weights in are bounded by their upper bounds.

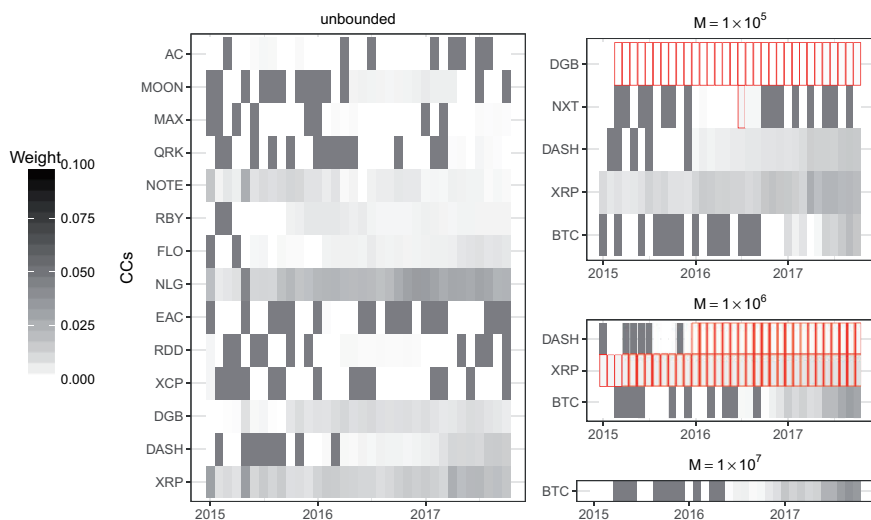


Figure 8 Weights given to CCs for SBC-CC Markowitz portfolio at each monthly rebalancing date under the three different investment amounts. Only CCs that have a non-zero weight on at least one rebalancing date are given. The darker the color, the larger the weight. Weights in are bounded by their upper bounds.

improvement compared to the pure S portfolio, and the highest cumulative returns at the sample end. A similar observation can be made for the SBC-CC portfolio. Looking at the weights for the monthly reallocation, [Figures 9 and 10](#), one observes only Bitcoin being included in the most strongly restricted portfolio. Since CVaR gives less weight to assets having high tail risk, it can be inferred that for larger investment amounts, only the tail risk of Bitcoin is sufficiently low to be appropriate for the portfolio. Interestingly, this causes the portfolio to outperform the other constrained ones, although the unbounded one still outperforms.

Looking at the summary statistics, [Tables 9 and 10](#), the SBC-CC unbounded portfolio performs better, with an annualized return of 0.17, than the corresponding Markowitz portfolio, with an annualized return 0.13. Also the Sharpe Ratio is enhanced. It is quite remarkable that the annualized return is high for $M = 1.0 \times 10^7$ US dollars, yet the Sharpe Ratio (0.05) is the same as for the corresponding Markowitz portfolio. For the S-CC portfolios, again they perform better than their SBC-CC CVaR counterparts, both in terms of annualized return and Sharpe Ratio. Again the $M = 1.0 \times 10^7$ US dollars portfolio shows a remarkably high annualized return. Comparing the weights, [Figures 7–10](#), one observes that only BTC was included, yet for the CVaR portfolios with higher values, only from mid-2016 onwards. For the Markowitz portfolios, BTC was also included in 2015.

5.3 Robustness: Weekly versus monthly rebalancing

Since the CCs' market has a great deal of variation, it is an interesting question to ask how will the portfolio perform if we readjust the portfolio more frequently, for example weekly

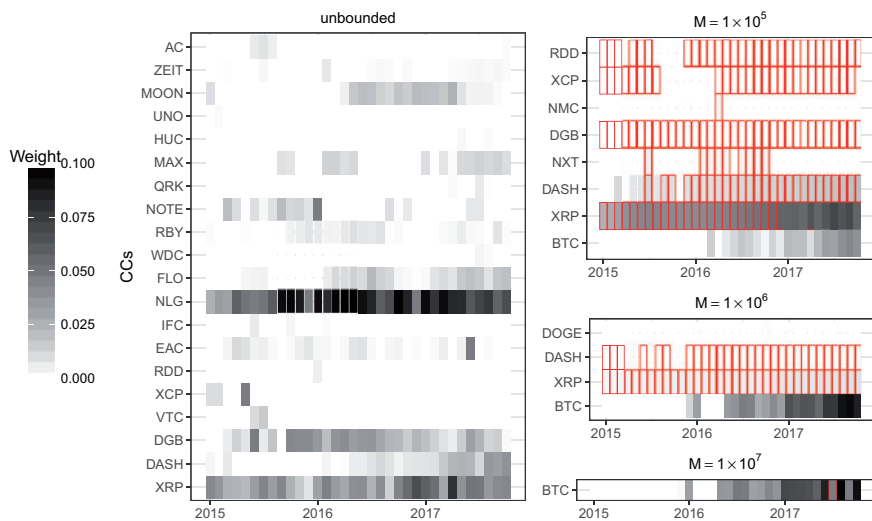


Figure 9 Weights given to CCs for S-CC CVaR portfolio at each monthly rebalancing date under the 3 different investment amounts. Only CCs that have a non-zero weight on at least one rebalancing date are given. The darker the color, the larger the weight. Weights in are bounded by their upper bounds.

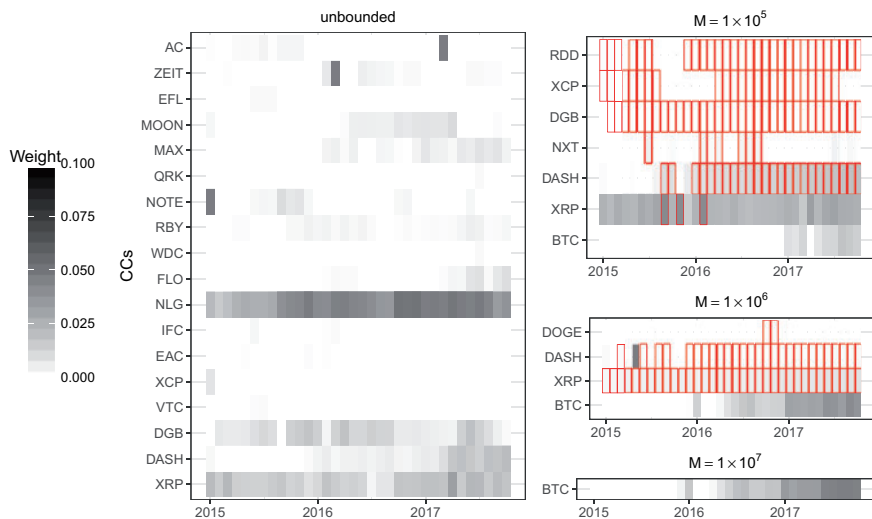


Figure 10 Weights given to CCs for SBC-CC CVaR portfolio at each monthly rebalancing date under the three different investment amounts. Only CCs that have a non-zero weight on at least one rebalancing date are given. The darker the color, the larger the weight. Weights in are bounded by their upper bounds.

rather than monthly. In this section, a weekly rebalancing portfolio is constructed, with weights updated every Wednesday. Again, the weights are calculated using an extending window approach: all the data before the readjustment date are used for the calculation, and the first portfolio is formed on January 1, 2015.

The cumulative return plots for the Markowitz and CVaR portfolios are shown in panels 5 b, 5d, 6 b and 6d. An overview of the results leads to the conclusion that the cumulative returns of the weekly rebalanced portfolios show almost the same pattern as those of the monthly ones; however, in most cases, they perform worse. For the Markowitz method, the cumulative returns at the end of the sample of the S and S-CC portfolio are 9.4% and 72.6% respectively, which are 6.7% and 8.5% smaller than those of their monthly readjusted counterparts. When liquidity constrained, the cumulative returns are 4.9%, 4.7% and 5.3% lower than for the monthly readjusted case at investment amounts 1×10^5 , 1×10^6 , 1×10^7 US dollars. The same deterioration happens when bond and commodity indices are included. For the case with CVaR portfolios, the situation is similar: in both the S-CC and SBC-CC cases, all the portfolios have a smaller cumulative return than their monthly counterpart.

Furthermore, rebalancing the portfolio weekly instead of monthly harms the performance. Apart from this issue, the return curves appear almost similar, suggesting robust results regarding the reallocation frequency. However the better performance with monthly rebalancing gives support for the interpretation that at times, swings in the return series of CCs have to be endured to ensure a better performance at the end of the day.

6 Conclusion

In this article, we have explored the potential gain of including CCs into risk optimized portfolios, taking into consideration the low liquidity of the CC market. On the one hand, the rapid rise in CCs make them promising investment assets, while on the other hand, they are more volatile, have heavy tails, and relatively low liquidity, so investing in them is somewhat challenging. To control the risk as well as the liquidity problem, we have proposed LIBRO method, which extends the framework employed in [Fan et al. \(2012\)](#) to contain an additional liquidity constraint, depending on the intended investment amount. Applying this method to monthly and weekly re-allocated Markowitz and CVaR portfolios consisting of S&P 100 component stocks, Barclays Capital US Aggregate Index (US Bonds Index), S&P GSCI (Commodities Index), and adding CCs to them, the results show a strong improvement in terms of volatility/quantile risk to returns. However, it is worth noting that Bitcoin (BTC), the earliest and most dominant CC, is given a zero weight when no liquidity constraint is included, and this under both definitions of risk: volatility and quantile risk. Two key conclusions can be inferred from this result: first, although the one most discussed in the literature, BTC is not the most appealing CC in terms of risk–return optimization, at least during the period covered by the present paper, which highlights our contribution to include CCs other than BTC for portfolio formation. Second, the inclusion of liquidity constraints appears necessary, since some high-weight CCs are less liquid than BTC. In this situation, one can no longer assume that positions in these CCs would not distort the market or be tradable in the necessary amounts. To improve the applicability of the portfolio formation strategy,

including an upper bound becomes necessary. It should be correlated with the daily trading volume of these assets.

The results of the in-sample analysis are already remarkable, while the out-of-sample analysis provides impressive results too. In the Markowitz portfolio the increase in cumulative returns reaches up to 80%. When including the liquidity upper bounds, the S-CC and SBC-CC portfolios still outperform the ones without constraints. For the Markowitz portfolios with monthly and weekly reallocation, the cumulative excess returns range from 10 to 22% with investment amounts equal to 1×10^5 , 1×10^6 , 1×10^7 US dollars. Over an investment period of roughly three years, this is a substantial gain. For the largest investment amount, which is, by construction, the most strongly restricted portfolio, one observes that the only CC included is BTC. Since CVaR gives less weight to assets having high tail risks, it can be inferred that for larger investment amounts, only the tail risk of BTC is sufficiently low to be appropriate for the portfolio. Furthermore, the monthly reallocated portfolios clearly outperformed the weekly adjusted ones. The better performance with monthly rebalancing gives support for the interpretation that, at times, swings in the return series of CCs have to be endured to ensure a better performance at the end of the day.

The main implications of this article are that including CCs into a portfolio can bring huge gains for the investor, even under the situation with the largest investment amount, which incurs the tightest liquidity constraint. In addition investing in alt-coins can provide much higher gains in the returns than just including BTC, but is more likely to encounter a liquidity problem; thus we have proposed LIBRO to tackle the low liquidity issue of certain CCs.

Supplementary Data

Supplementary data are available at *Journal of Financial Econometrics* online.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft via IRTG 1792 “High Dimensional Non Stationary Time Series”, Humboldt-Universität zu Berlin, and the China Scholarship Council. We gratefully acknowledged their support.

Appendix

A.1 A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices

For an $(N \times T)$ de-meaned matrix X with T iid observations, define S to be the sample covariance matrix $S = XX^T/T$, I_N to be the N -dimensional identity matrix, and $\text{tr}(\cdot)$ to indicate the trace of a matrix. Define $\|\cdot\|$ to be the Frobenius norm normalized by the dimension N , that is, $\|X\| = \sqrt{\text{tr}(XX^T)/N}$. Further define $\zeta = \text{tr}(\Sigma I_N)/N$, $\gamma^2 =$

Table A1 Summary statistics of CCs. Ann.Ret and Ann.STD indicate annualized mean return and standard deviation, which are calculated by multiplying 250 and $\sqrt{250}$ by their daily counterparts. Mean volume and market cap are measured in US dollars.

	Ann.Ret	Ann.STD	Skewness	kurtosis	ρ	mean volume	market cap
BTC	0.49	0.55	-0.61	10.87	-0.01	4.71e+08	1.07e+10
LTC	0.29	0.87	0.34	23.70	0.02	9.04e+07	3.22e+08
XRP	0.68	1.09	2.72	37.76	0.01	4.88e+06	8.35e+08
DASH	1.16	1.37	0.62	50.60	-0.14	1.61e+06	1.52e+08
DOGE	0.11	0.96	0.97	15.54	0.01	3.87e+05	3.11e+07
NXT	0.17	1.08	0.80	8.42	-0.03	2.28e+05	2.08e+07
DGB	0.73	1.70	2.93	29.89	-0.03	1.63e+05	7.24e+06
PPC	-0.17	1.00	0.62	12.51	-0.05	1.05e+05	1.54e+07
BLK	-0.05	1.27	1.82	17.72	-0.08	7.15e+04	4.21e+06
VTC	0.33	1.76	1.84	17.83	-0.01	5.93e+04	2.42e+06
NMC	-0.17	1.07	1.30	25.81	-0.09	5.26e+04	8.60e+06
POT	0.44	1.57	0.79	19.40	-0.05	3.95e+04	2.30e+06
XCP	0.36	1.46	1.16	9.52	-0.10	3.69e+04	7.51e+06
RDD	0.90	2.38	0.36	14.59	-0.25	2.55e+04	2.83e+06
XPM	-0.39	1.29	0.85	16.66	-0.07	2.05e+04	1.54e+06
NVC	-0.07	1.09	2.39	24.07	-0.01	1.73e+04	1.58e+06
EMC2	0.56	1.93	1.63	14.47	-0.02	1.57e+04	7.16e+05
EAC	-0.04	2.70	1.08	25.71	-0.27	1.50e+04	7.92e+05
IFC	-0.09	2.15	1.73	19.11	-0.13	1.26e+04	8.37e+05
NLG	1.00	1.46	0.90	12.35	-0.08	1.12e+04	4.28e+06
FTC	-0.00	1.76	0.59	13.66	-0.04	1.02e+04	1.84e+06
FLO	0.77	1.83	1.35	10.87	-0.07	8.35e+03	7.74e+05
ZET	-0.35	2.03	0.96	19.83	-0.16	5.80e+03	7.54e+05
WDC	-0.33	1.86	1.45	29.43	-0.12	5.65e+03	7.94e+05
RBY	1.16	2.18	0.98	28.41	-0.26	5.16e+03	2.83e+06
NOTE	0.79	1.60	1.59	15.01	-0.10	4.76e+03	1.11e+06
QRK	-0.27	2.91	-0.05	36.12	-0.35	4.38e+03	1.29e+06
MAX	-0.48	4.37	0.54	96.94	-0.38	3.69e+03	4.13e+05
HUC	-0.04	2.31	0.42	8.39	-0.19	3.25e+03	2.31e+05
SLR	0.72	1.92	0.44	9.59	-0.21	3.01e+03	2.39e+06
AUR	-0.03	1.51	1.08	19.12	-0.08	2.55e+03	1.17e+06
UNO	0.57	1.53	-0.24	15.64	-0.21	2.10e+03	9.14e+05
DMD	0.66	1.47	0.85	12.79	-0.20	1.57e+03	6.89e+05
GRS	0.99	2.86	1.52	15.80	-0.22	1.45e+03	4.72e+05
MINT	-0.01	3.63	0.27	7.76	-0.32	1.04e+03	9.71e+05
DGC	-0.43	1.92	0.71	27.86	-0.15	8.42e+02	2.95e+05
MOON	0.66	2.76	0.24	18.84	-0.16	6.83e+02	1.09e+06
EFL	0.58	1.87	-0.56	18.73	-0.16	6.19e+02	2.24e+05
NET	1.76	5.84	7.84	201.18	-0.22	5.45e+02	2.84e+05
CBX	0.17	3.33	0.81	28.53	-0.33	2.45e+02	1.80e+05
ZEIT	0.15	4.47	0.24	63.85	-0.25	2.10e+02	3.78e+05
AC	-0.30	3.05	1.17	26.52	-0.20	7.66e+01	4.13e+05
S&P stocks average	0.08	0.20	-0.26	10.29	0.00	5.40e+08	
Bond index	0.05	0.05	-1.91	12.38	-0.04		
Commodity index	-0.20	0.20	0.05	4.84	-0.07		

$\|\Sigma - \zeta I_N\|^2, \beta^2 = E(\|S - \Sigma\|^2)$ and $\delta^2 = E(S - \zeta I_N)$. Assuming $E(X^4) < \infty$, the optimization problem considered is

$$\begin{aligned} \min_{\rho_1, \rho_2} &= E(\|\hat{\Sigma} - \Sigma\|^2) \\ \text{s.t. } &\hat{\Sigma} = \rho_1 I_N + \rho_2 S \end{aligned}$$

which solves to $\rho_1 = \frac{\beta^2}{\delta^2} \zeta$ and $\rho_2 = \frac{\gamma^2}{\delta^2}$, thus the estimator is

$$\hat{\Sigma} = \frac{\beta^2}{\delta^2} \zeta I_N + \frac{\gamma^2}{\delta^2} S.$$

Since the estimator depends on the true covariance matrix Σ , a consistent estimator $\hat{\Sigma}^*$ has been introduced. Define x_k as the $(N \times 1)$ column k of X and rewrite the sample covariance matrix S as $S = \frac{1}{T} \sum_{k=1}^T x_k x_k^\top$. Since the matrices $x_k x_k^\top$ are iid across k , β^2 can be estimated by $\bar{b}^2 = 1/N^2 \sum_{i=1}^N \|x_i x_i^\top - S\|^2$, Ledoit and Wolf (2004). Further define $m = \text{tr}(S I_N)/N$, $d^2 = \|S - m I_N\|^2$, $\bar{b}^2 = \min(\bar{b}^2, d^2)$ and $k^2 = d^2 - \bar{b}^2$. A consistent estimator is then

$$\hat{\Sigma}^* = \frac{\bar{b}^2}{d^2} m I + \frac{k^2}{d^2} S.$$

References

- Borke, L., and W. K. Härdle. 2017. “GitHub API based QuantNet Mining infrastructure in R”. SFB 649 Discussion Paper.
- Borke, L., and W. K. Härdle. 2018. “Q3-D3-LSA”. *Handbook of Big Data Analytics*. Springer Verlag.
- Brière, M., K. Oosterlinck, and A. Szafarz. 2015. Virtual Currency, Tangible Return: Portfolio Diversification with Bitcoin. *Journal of Asset Management* 16: 365–373.
- Chen, C. Y., W. K. Härdle, A. J. Hou, and W. Wang. 2018. Pricing Cryptocurrency Options: The Case of CRIX and Bitcoin. *Journal of Financial Econometrics (Submitted)*.
- Chen, S., C. Y.-H. Chen, W. K. Härdle, T. M. Lee, and B. Ong. 2017. “CRIX Index: Analysis of a Cryptocurrency Index for Portfolio Investment.” In D. Lee Kuo Chuen and R. Deng. (eds), *Handbook of Digital Finance and Financial Inclusion: Cryptocurrency, FinTech, InsurTech, and Regulation*, Vol. 1. London, UK: Elsevier.
- Darolles, S., C. Gourieroux, and E. Jay. 2012. “Robust Portfolio Allocation with Systematic Risk Contribution Restrictions”. SSRN Scholarly Paper ID 2192399. Rochester, NY: Social Science Research Network.
- Eisl, A., S. M. Gasser, and K. Weinmayer. 2015. “Caveat Emptor: Does Bitcoin Improve Portfolio Diversification?”. SSRN Scholarly Paper ID 2408997. Rochester, NY: Social Science Research Network.
- Elendner, H., S. Trimborn, B. Ong, and T. M. Lee. 2018. “The Cross-Section of Crypto-Currencies as Financial Assets”. In D. Lee Kuo Chuen and R. Deng. (eds), *Handbook of Digital Finance and Financial Inclusion: Cryptocurrency, FinTech, InsurTech, and Regulation*. Vol. 1. London, UK: Elsevier.
- Fan, J., J. Zhang, and K. Yu. June. 2012. Vast Portfolio Selection with Gross-Exposure Constraints. *Journal of the American Statistical Association* 107: 592–606.

- Friedman, J., T. Hastie, and R. Tibshirani. 2008. Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics* 9: 432–441.
- Green, R. C. and B. Hollifield. 1992. When Will Mean-Variance Efficient Portfolios Be Well Diversified? *The Journal of Finance* 47: 1785–1809.
- Härdle, W. K., and L. Simar. 2015. “Applications in Finance”. *Applied Multivariate Statistical Analysis*. 4th ed. Heidelberg, Germany: Springer, pp. 487–499.
- Härdle, W. K., D. K. C. Lee, S. Nasekin, and A. Petukhina. 2018. Tail Event Driven ASset Allocation: Evidence from Equity and Mutual Funds’ Markets. *Journal of Asset Management* 19: 49–63.
- Jagannathan, R., and A. T. Ma. 2003. Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps. *The Journal of Finance* 58: 1651–1683.
- Kozak, S., S. Nagel, and S. Santosh. 2017. Interpreting Factor Models. *Journal of Finance* (Forthcoming).
- Krokhmal, P., J. Palmquist, and S. Uryasev. 2002. Portfolio Optimization with Conditional Value-at-Risk Objective and Constraints. *The Journal of Risk* 4: 43–68.
- Ledoit, O. and M. Wolf. 2004. A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices. *Journal of Multivariate Analysis* 88: 365–411.
- Lee Kuo Chuen, D., L. Guo, and Y. Wang. 2017. Cryptocurrency: A New Investment Opportunity? *The Journal of Alternative Investments* 20: 16–40.
- Markowitz, H. 1952. Portfolio Selection. *The Journal of Finance* 7: 77–91.
- Rockafellar, R. T., and S. Uryasev. 2000. Optimization of Conditional Value-at-Risk. *The Journal of Risk* 2: 21–42.
- Scaillet, O., A. Treccani, and C. Trevisan. 2018. High-Frequency Jump Analysis of the Bitcoin Market. *Journal of Financial Econometrics* (Forthcoming).
- Trimborn, S., and W. K. Härdle. 2018. CRIX an Index for Cryptocurrencies. *Journal of Empirical Finance* 49: 107–122.
- Wyss, R. v. 2004. “Measuring and Predicting Liquidity in the StockMarket”. Doctoral thesis.



Model-driven statistical arbitrage on ETF option markets

S. Nasekin & W. K. Härdle

To cite this article: S. Nasekin & W. K. Härdle (2019) Model-driven statistical arbitrage on ETF option markets, *Quantitative Finance*, 19:11, 1817-1837, DOI: [10.1080/14697688.2019.1605186](https://doi.org/10.1080/14697688.2019.1605186)

To link to this article: <https://doi.org/10.1080/14697688.2019.1605186>



Published online: 23 May 2019.



Submit your article to this journal [↗](#)



Article views: 207




View related articles [↗](#)



View Crossmark data [↗](#)

Model-driven statistical arbitrage on LETF option markets

S. NASEKIN *[†] and W. K. HÄRDLE[‡]

[†]Lancaster University Management School, Lancaster University, Lancaster LA14YX, UK
[‡]C.A.S.E.—Center for Applied Statistics & Economics, Humboldt-Universität zu Berlin, Spandauer Str. 1, Berlin 10178, Germany

(Received 9 September 2017; accepted 29 March 2019; published online 23 May 2019)

In this paper, we study the statistical properties of the moneyness scaling transformation, which adjusts the moneyness coordinate of the implied volatility smile in an attempt to remove the discrepancy between the IV smiles for levered and unlevered ETF options. We construct bootstrap uniform confidence bands which indicate that the implied volatility smiles are statistically different after moneyness scaling has been performed. An empirical application shows that there are trading opportunities possible on the LETF market. A statistical arbitrage type strategy based on a dynamic semiparametric factor model is presented. This strategy presents a statistical decision algorithm which generates trade recommendations based on comparison of model and observed LETF implied volatility surface. It is shown to generate positive returns with a high probability. Extensive econometric analysis of the LETF implied volatility process is performed including out-of-sample forecasting based on a semiparametric factor model and a uniform confidence bands' study. These provide new insights into the latent dynamics of the implied volatility surface. We also incorporate Heston stochastic volatility into the moneyness scaling method for better tractability of the model.

Keywords: Exchange-traded funds; Options; Implied volatilities; Moneyness scaling; Bootstrap; Dynamic factor models; Trading strategies

JEL Classification: C00, C14, C50

1. Introduction

Exchange-traded funds (ETFs) are financial products that track indices, commodities, bonds and baskets of assets. They have become increasingly popular due to diversification benefits as well as the investor's ability to perform short-selling, buying on margin and lower expense ratios than, for instance, those of mutual funds.

Leveraged ETFs (LETFs) are used to generate multiples or inverse multiples of returns on the underlying asset. For instance, the LETF ProShares Ultra S&P500 (SSO) with a leverage ratio $\beta = +2$ is supposed to grow 2% for every 1% daily gain in the price of the S&P500 index, minus an expense fee. An inverse leveraged ETF would invert the gain/loss of the underlying index and amplify it proportionally to the ratio: the ProShares UltraShort S&P500 (SDS) with leverage ratio $\beta = -2$ would generate a 2% gain for every 1% daily loss in the price of the underlying S&P500 index. Selected financial

information on LETFs relevant for this study is summarized in table 1.

Due to their growing popularity and the nature of ETF and LETF similar dynamics, recently there has been growing research on leveraged ETFs and their consistent pricing. Avellaneda and Zhang (2010) found that the terminal value of an option on a LETF depends on the integrated variance of the underlying LETF. Ahn *et al.* (2015) show that given Heston dynamics for the underlying ETF, the corresponding LETF also has Heston dynamics with different parameters. Leung and Santoli (2016) give a broad overview of LETFs, LETF options, their pricing and implied volatility.

Studies on LETF options' implied volatility should be mentioned as most relevant for this work. Figueroa-Lopez *et al.* (2016) derived asymptotic equivalence of options on ETFs and LETFs under restricted conditions. Lee and Wang (2015) studied a direct relationship between volatility skews of leveraged and unlevered products, providing asymptotic error estimates. Leung *et al.* (2017) studied the relationship between the ETF and LETF implied volatility

*Corresponding author. Email: sergey.nasekin@gmail.com

Table 1. Summary financial information on (leveraged) ETFs on S&P 500 underlying index.

(L)ETF	Ticker	Lev. ratio	Exp. ratio (%)	Div. yield (%)
SPDR S&P 500	SPY	+1	0.090	1.867
ProShares Ultra S&P500	SSO	+2	0.900	0.440
ProShares UltraPro S&P500	UPRO	+3	0.950	0.263
ProShares UltraShort S&P500	SDS	-2	0.890	0.000
ProShares UltraPro Short S&P500	SPXU	-3	0.900	0.000

surfaces when the underlying ETF is modeled by a general class of local-stochastic volatility models.

Various studies including Aït-Sahalia *et al.* (2001), Cont and da Fonseca (2002) and Fengler *et al.* (2003, 2007) apply non- and semiparametric approaches to model implied volatility surfaces (IVS). Some important issues of IVS estimation and modeling include choice between parametric and non-parametric methods, model selection and out-of-sample forecasting.

Leung and Sircar (2015) introduced the so-called ‘moneyness scaling’ technique which links implied volatilities (IV) between ETF and LETF in the way that the discrepancy between the implied volatility ‘smile’ pattern is removed. The question arises whether the moneyness scaling method indeed removes discrepancies consistently in time. To answer this question, we need to verify whether IV deviations are significant from the statistical point of view.

In this study, we use an econometric approach to study the issue of errors describing the difference between leveraged and unleveraged volatility smiles. Unlike Lee and Wang (2015), who derive an asymptotic error estimate directly for implied volatilities, we investigate the indirect approach of Leung and Sircar (2015) further, invoking a stochastic volatility framework. This approach allows one to apply the moneyness scaling technique under more general assumptions.

We consider the statistical properties of the moneyness scaling transformation which adjusts the moneyness coordinate of the implied volatility smile in an attempt to remove the discrepancy between the IV smiles for levered and unlevered ETF options. We construct bootstrap uniform confidence bands which allow for more flexible error analysis. The results indicate that the implied volatility smiles are statistically different, even after moneyness scaling has been performed.

Furthermore, we develop a trading strategy based on a dynamic semiparametric factor model. This strategy utilizes the dynamic structure of the implied volatility surface allowing out-of-sample forecasting and information on unleveraged ETF options to construct theoretical one-step-ahead implied volatility surfaces. This strategy exploits statistical discrepancies on (L)ETF markets and falls within the class of model-driven statistical arbitrage described in Avellaneda and Lee (2010).

2. Confidence analysis of moneyness scaling

2.1. Moneyness scaling

We begin by introducing basic results on (L)ETF options and moneyness scaling. The dynamics of the underlying asset is assumed to follow a stochastic process under a risk-neutral

measure \mathbb{Q} :

$$\frac{dS_t}{S_t} = (r - \delta) dt + \sigma_t dW_t^{\mathbb{Q}}, \quad (1)$$

where r is the risk-free interest rate, δ the dividend yield, $(\sigma_t)_{t \geq 0}$ is some stochastic volatility process.

The moneyness scaling technique proposed by Leung and Sircar (2015) proposes a coordinate transformation for the LETF option implied volatility and potentially reflects the increase of risk in the underlying index.

Figure 1 compares empirical implied volatilities for SSO, UPRO, SDS, SPXU before moneyness scaling has been applied and afterwards. In this example, the log-moneyness $LM \stackrel{\text{def}}{=} \log(K/L_t)$ is used, where K is the strike of the LETF option and L_t the LETF price at time t . After re-scaling, there are still visible discrepancies between the implied volatilities for the SPY ETF and its leveraged counterparts. The moneyness scaling procedure yields a more coherent picture when the LETF and ETF implied volatilities overlap visually better.

Based on the assumption that the distribution of the terminal price of the β -LETF depends on the leverage ratio β , the moneyness scaling formula includes an expectation of the β -LETF log-moneyness conditional on the terminal value of the unleveraged counterpart. For the LETF log-moneyness $LM^{(\beta)}$ (consider ETFs as LETFs with $\beta = 1$) the result linking the log-moneyness coordinates $LM^{(\beta)}$ and $LM^{(1)}$ of the leveraged and unleveraged ETF is written as follows:

$$LM^{(\beta)} = \beta LM^{(1)} - \{r(\beta - 1) + c^*\}T - \frac{\beta(\beta - 1)}{2} \mathbb{E}^{\mathbb{Q}} \left\{ \int_0^T \sigma_t^2 dt \mid \log \left(\frac{S_T}{S_0} \right) = LM^{(1)} \right\}, \quad (2)$$

where T is the time to maturity/expiration (TTM), $c^* = c + \delta$ is the LETF expense ratio c corrected for dividend yield δ . The expense ratio c is expressed in percent and approximates an annual fee charged by the ETF from the shareholders to cover the fund’s operating expenses.

More generally, for two LETFs with different leverage ratios β_1, β_2 expression (2) takes the form:

$$LM^{(\beta_1)} = \frac{\beta_1}{\beta_2} LM^{(\beta_2)} + \left[\left\{ \frac{\beta_1}{\beta_2} (\beta_2 - 1) - (\beta_1 - 1) \right\} r + \frac{\beta_1}{\beta_2} c_2^* - c_1^* \right] T + \frac{\beta_1(\beta_2 - 1) - \beta_1(\beta_1 - 1)}{2} \mathbb{E}^{\mathbb{Q}} \left\{ \int_0^T \sigma_t^2 dt \mid \log \left(\frac{S_T}{S_0} \right) = LM^{(1)} \right\}, \quad (3)$$

with $c_k^* = c_k + \delta_k$, $k = 1, 2$.

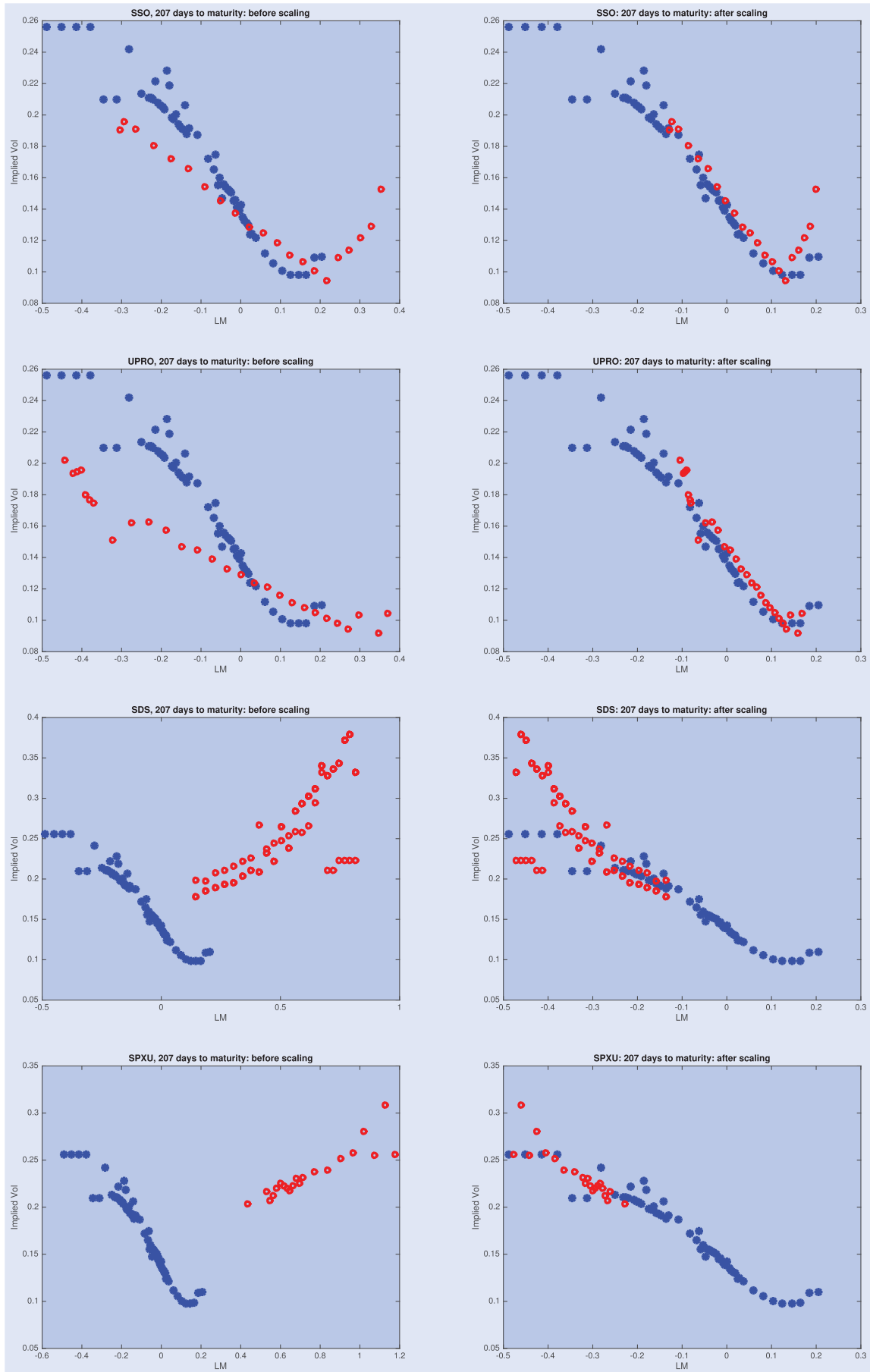


Figure 1. SPY (blue) and LETFs (red) implied volatilities before (left column) and after scaling (right column) on June 23, 2015 with 207 days to maturity, plotted against log-moneyness.

It is worth mentioning that declared leverage ratios are not always the same as the empirical ones. Leung and Santoli (2016) introduce a novel method to estimate the empirical leverage ratio realized by ETFs. Particularly for longer holding periods, return discrepancies between ETFs and LETFs increase. This can cause, e.g. for both long and short ETFs to have negative cumulative returns over a longer horizon. In this study we use declared leverage ratios. This is motivated by the use of short time horizons in the empirical study of LETF option portfolios' returns in section 5.

2.2. Confidence bands

Cont and da Fonseca (2002), Fengler et al. (2007) and Park et al. (2009) studied the implied volatility as a random process in time, so that the data generating process includes some non-parametric function m :

$$Y_t = m(X_t) + \varepsilon_t, \quad t = 1, \dots, T, \tag{4}$$

or can be driven by a latent factor process Z_t :

$$Y_t = Z_t^\top m(X_t) + \varepsilon_t, \quad t = 1, \dots, T, \tag{5}$$

where Y_t stands for an implied volatility process, the covariates X_t can be one- or multi-dimensional, including, for instance, moneyness and time-to-maturity.

The statistical properties of the estimators $\hat{m}(X_t)$ and $\hat{Z}_t^\top \hat{m}(X_t)$ for models (4) and (5) have been outlined, respectively, in, e.g. Härdle (1990), Ruppert and Wand (1994) and Park et al. (2009). To study the consistency of the implied volatility difference between the ETF and the moneyness-scaled LETF case, one needs to consider statistical differences of the corresponding estimators. Confidence band analysis may provide an insight into the matter. An important issue for smooth confidence bands for functions is the correct probability of covering the 'true' curve.

The approach of Härdle et al. (2015) proposes a uniform bootstrap bands construction for a wide class of non-parametric M and L -estimates. It is logical to use a robust M -type smoother for the estimation of (4) for implied volatility, as IV data often suffer from outliers. The procedure runs as follows: considering the sample $\{X_t, Y_t\}_{t=1}^T$, where Y_t denotes the IV process, X_t is taken to be one-dimensional and includes the log-moneyness covariate $LM^{(\beta)}$, do the following:

- (i) compute the estimate $\hat{m}_h(X_t)$ by a local linear M -smoothing procedure (see appendix A.1) with some kernel function and bandwidth h chosen by, e.g. cross-validation, and obtain residuals $\hat{\varepsilon}_t \stackrel{\text{def}}{=} Y_t - \hat{m}_h(X_t)$,
- (ii) do bootstrap resampling from $\hat{\varepsilon}_t$: for each $t = 1, \dots, T$, generate random variables $\varepsilon_{t,b}^* \sim \hat{F}_{\varepsilon|X_t}(z)$ for $b = 1, \dots, B$ according to the conditional edf

$$\hat{F}_{\varepsilon|x}(z) \stackrel{\text{def}}{=} \frac{\sum_{t=1}^T K_h(x - X_t) \mathbf{1}\{\hat{\varepsilon}_t \leq z\}}{\sum_{t=1}^T K_h(x - X_t)}, \tag{6}$$

which is further centered as shown in Härdle et al. (2015). Then construct the bootstrap sample $Y_{t,b}^*$

as follows:

$$Y_{t,b}^* = \hat{m}_g(X_t) + \varepsilon_{t,b}^*, \tag{7}$$

with an 'oversmoothing' bandwidth $g \gg h$ such as $g = \mathcal{O}(T^{-1/9})$ to allow for bias correction,

- (iii) for each bootstrap sample $\{X_t, Y_{t,b}^*\}_{t=1}^T$ compute $\hat{m}_{h,g}^*$ using the bandwidth h and construct the random variable

$$d_b \stackrel{\text{def}}{=} \sup_{x \in J} \left[\frac{|\hat{m}_{h,g}^*(x) - \hat{m}_g(x)| \sqrt{\hat{f}_X(x) \hat{f}_{\varepsilon|x}(\varepsilon_t^*)}}{\sqrt{\hat{E}_{Y|x}\{\psi^2(\varepsilon_t^*)\}}} \right], \tag{8}$$

where J is a finite compact support set of \hat{f}_X and $\psi(u) = \rho'(\cdot)$ as described in appendix A.1; the conditional expectation $\hat{E}_{Y|x}(\cdot)$ is defined with respect to the edf

$$\hat{F}_{Y|x}(z) \stackrel{\text{def}}{=} \frac{\sum_{t=1}^T K_h(x - X_t) \mathbf{1}\{Y_t \leq z\}}{\sum_{t=1}^T K_h(x - X_t)}, \tag{9}$$

where $\hat{f}_{\varepsilon|x}(\cdot)$ and $\hat{f}_X(x)$ are consistent estimators of conditional density corresponding to (6) and the density $f_X(x)$, respectively; for more details, see Härdle et al. (2015),

- (iv) calculate the $1 - \alpha$ quantile d_α^* of d_1, \dots, d_B ,
- (v) construct the bootstrap uniform confidence band centered around $\hat{m}_h(x)$:

$$\hat{m}_h(x) \pm \left[\frac{\sqrt{\hat{E}_{Y|x}\{\psi^2(\varepsilon_t^*)\}} d_\alpha^*}{\sqrt{\hat{f}_X(x) \hat{f}_{\varepsilon|x}(\varepsilon_t^*)}} \right]. \tag{10}$$

Such an approach utilizes bootstrap confidence bands while the distribution of the original data is 'mimicked' via a pre-specified random mechanism achieving both uniformity and better coverage. Additionally, it performs better than asymptotic confidence bands which generally tend to underestimate the true coverage probability, see Hall and Horowitz (2013). Compared to a Bonferroni approach, bootstrap uniform confidence bands would be less conservative and make use of the substantial positive correlation of the curve estimates at nearby points, see Härdle (1990).

3. Moneyness scaling under Heston stochastic volatility

3.1. An analytical approach

In the study of moneyness scaling, one needs to estimate the following conditional expectation:

$$\mathbb{E}^Q \left\{ \int_0^T \sigma_t^2 dt \mid \log \left(\frac{S_T}{S_0} \right) = LM^{(1)} \right\}. \tag{11}$$

Taking $\sigma_t = \sigma$ constant, one obtains $\sigma^2 T$. As empirical evidence shows, constant volatility is not a plausible assumption, therefore one needs to determine the measure Q for the case of random volatility under a model which allows random

dynamics of σ_t . Second, one needs to estimate the integrated variance

$$\int_0^T \sigma_t^2 dt. \quad (12)$$

Stochastic volatility presents a viable alternative to the constant case. One could choose among different specifications of stochastic volatility models. Popular special cases include specifications of Heston (1993), Hull and White (1987) and Schöbel and Zhu (1999). An example of a more general stochastic volatility system is given in Leung and Sircar (2015). Simpler models tend to generate semi-closed-form solutions for return distributions. For instance, a solution for the Heston model by Heston (1993) was proposed by Dragulescu and Yakovenko (2002).

We use the Heston model to compute the quantity in (11). As noticed in Leung and Santoli (2016), this approach allows for tractability and efficient numerical pricing of options on LETFs. Stochastic volatility framework also allows to better assess volatility decay, i.e. value erosion due to the increase of the realized variance with the holding horizon.

The Heston model with risk-neutral dynamics under a risk-neutral measure \mathbf{Q} and zero volatility risk premium is described by a two-dimensional system of stochastic differential equations

$$dS_t = (r - c - 0.5)V_t dt + \sqrt{V_t} S_t dW_{S,t}^{\mathbf{Q}}, \quad (13)$$

$$dV_t = \kappa(\theta - V_t) dt + \sigma dW_{V,t}^{\mathbf{Q}}, \quad (14)$$

where we have put $V_t = \sigma_t^2$; $r - c$ are costs of carry on S_t , θ is the long-run variance level, κ is the rate of reversion to θ , σ is the ‘volatility of the volatility’ parameter which determines the variance of V_t ; $W_{S,t}$, $W_{V,t}$ are correlated with parameter ρ . The tails of the Heston-implied densities for log-returns $x_t = \log(S_t/S_{t-1})$ are exponential and heavier than those of the normal distribution with the dispersion parameter equal to the long-term variance θ , see, i.e. Cizek *et al.* (2011).

An analytical solution to (11) requires knowledge of the conditional distribution of the integrated variance (12) given the logarithm of terminal stock price $\log(S_T)$. If we define:

$$\begin{aligned} \tilde{V} &\stackrel{\text{def}}{=} \int_0^T V_t dt, \\ X_T &\stackrel{\text{def}}{=} \log(S_T), \end{aligned}$$

then we can write

$$\begin{aligned} \mathbf{E}^{\mathbf{Q}} \left\{ \int_0^T V_t dt \mid \log \left(\frac{S_T}{S_0} \right) = LM^{(1)} \right\} \\ = \mathbf{E}^{\mathbf{Q}} \left\{ \tilde{V} \mid X_T = \widetilde{LM}^{(1)} \right\} \\ = \int_0^\infty f_{\tilde{V}|X_T}(\tilde{v} \mid X_T = \widetilde{LM}^{(1)}) d\tilde{v}, \end{aligned} \quad (15)$$

where $f_{\tilde{V}|X_T}(\tilde{v} \mid X_T)$ is the conditional density of \tilde{V} given X_T under the measure \mathbf{Q} and $\widetilde{LM}^{(1)} = \log(S_0) + LM^{(1)}$.

Unfortunately, $f_{\tilde{V}|X_T}(\tilde{v} \mid X_T)$ does not assume a simple form and is ultimately expressed in terms of Fourier transforms of

characteristic functions of these quantities. Technical details are given in appendix A.2. As follows from the details, four improper integrals have to be estimated. Numeric integration methods can be used to approximate (11).

Additional complexity arises from the necessity to evaluate a modified Bessel function of the first kind which takes a complex argument. Numerical approximation methods for such evaluations such as the trapezoidal rule are outlined in Broadie and Kaya (2006).

Considering the complexity of the density estimation, we consider a Monte-Carlo approach to evaluate (11). This method is feasible and straightforward from a practical point of view.

3.2. A Monte-Carlo approach

Alternatively, the conditional expectation in (11) can be computed using Monte-Carlo simulations. The simulations are performed using the Heston model and the calibrated parameters obtained minimizing the squared difference between theoretical Heston prices $C^\Theta(K, \tau)$ obtained from the model and observed market prices $C^M(K, \tau)$,

$$\min_{\Theta \in \mathbb{R}^5} \sum_{i=1}^N (C_i^\Theta(K_i, \tau_i) - C_i^M(K_i, \tau_i)), \quad (16)$$

where $\Theta \stackrel{\text{def}}{=} (\kappa, \theta, \sigma, \nu_0, \rho)$ Heston parameters, N number of options used for calibration, K strikes and τ times-to-maturity. Theoretical prices $C^\Theta(K, \tau)$ are obtained via numeric integration of the Heston characteristic function.

The Monte-Carlo algorithm is motivated by van der Stoep *et al.* (2014) and can be formulated as follows:

- (i) Generate N pairs of observations (s_i, v_i) , $i = 1, \dots, N$.
- (ii) Order the realizations s_i : $s_1 \leq s_2 \leq \dots \leq s_N$.
- (iii) Determine the boundaries of M bins $(l_k, l_{k+1}]$, $k = 1, \dots, M$ on an equidistant grid of values $S^* \stackrel{\text{def}}{=} S_0 e^{LM^{(1)}}$.
- (iv) For the k th bin approximate the conditional expectation (11) by

$$\mathbf{E}^{\mathbf{Q}} \left(\int_0^T \sigma_t^2 dt \mid S_T \in (l_k, l_{k+1}] \right) \approx \frac{h}{NQ(k)} \sum_{i=1}^H \sum_{j \in \mathcal{J}_k} V_{ij}, \quad (17)$$

where h is the discretization step for V_t , \mathcal{J}_k the set of numbers j , for which the observations S_T are in the k th bin and $Q(k)$ is the probability of S_T being in the k th bin.

The results of the simulation are presented in figure 2. Polynomial smoothing is applied to produce the smoothed version of SCO LETF realized variance. The generated expected realized variance has the form of a ‘smile’ which confirms the intuition behind using average square implied volatility in the case of constant-volatility moneyness scaling approach.

We use the Monte-Carlo approach given Euler discretization scheme for the empirical application in later sections given its tractability and theoretical justification. Both methods, analytical and Monte-Carlo, introduce errors into the

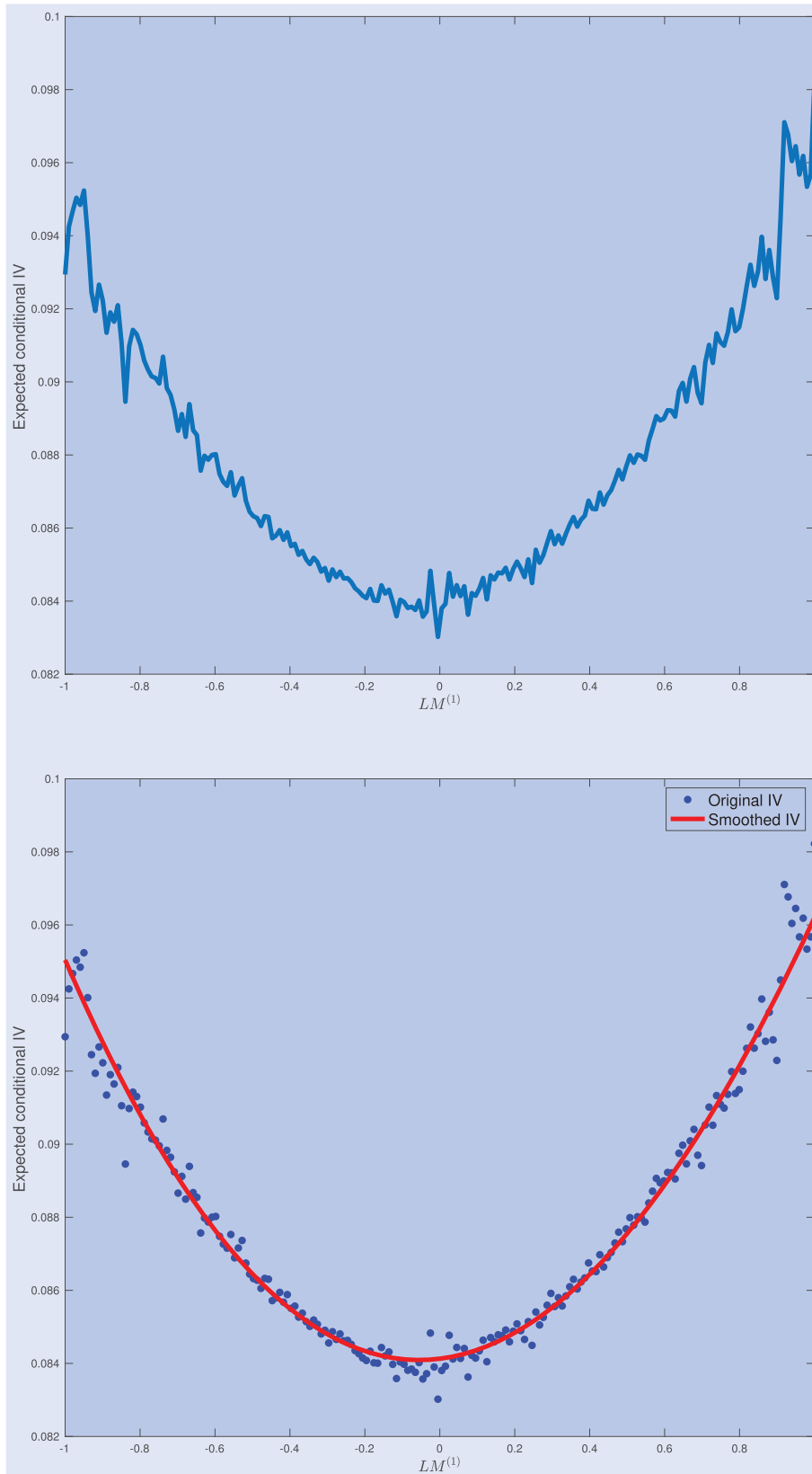


Figure 2. Upper panel: estimated value of $E^Q(\int_0^T \sigma_t^2 dt | \log(S_T/S_0) = LM^{(1)})$; lower panel: smoothed estimate.

calculation of (11). For the analytical method, discretization and truncation errors appear when the integral is estimated at discrete points and is truncated to be approximated as a finite

sum. If the trapezoidal rule is used to approximate the integrals in the analytical method, then the discretization error is of order $\mathcal{O}(M^{-2})$ where $M - 1$ is the number of discrete

intervals. However, as the integral dimensionality d increases, the discretization error order increases to $\mathcal{O}(M^{-2/d})$ ('curse of dimensionality').

For the current example, the discretization error order for (11) becomes $\mathcal{O}(M^{-1/2})$, which matches the convergence order of Monte-Carlo discretization bias. Additionally, analytical approximation of (11) effects a truncation error which is potentially significant due to the oscillatory nature of the integrand. Monte-Carlo approach inherently induces a statistical error of order $\mathcal{O}(N^{-1/2})$ which can be made sufficiently small by taking a large number N of samples. As noted in Higham and Mao (2005) and Lord *et al.* (2010), a simple discretization scheme such as the Euler scheme converges to the true process under certain conditions on the discretization size. This was shown to be true for the Heston model in particular by Higham and Mao (2005).

4. Dynamic semiparametric factor model

4.1. Model description

A generalized version of the model in (4) represented by (5) assumes the implied volatility Y_t to be a stochastic process driven by a latent stochastic factor process \mathcal{Z}_t contaminated by noise ε_t . To be more specific, define $\mathcal{J} \stackrel{\text{def}}{=} [\kappa_{\min}, \kappa_{\max}] \times [\tau_{\min}, \tau_{\max}]$, $Y_{t,j}$ implied volatility, $t = 1, \dots, T$ time index, $j = 1, \dots, J_t$ option intraday numbering on day t , $X_{t,j} \stackrel{\text{def}}{=} (\kappa_{t,j}, \tau_{t,j})^\top$, $\kappa_{t,j}$, $\tau_{t,j}$ are, respectively, a moneyness measure (log-, forward, etc.) and time-to-maturity at time point t for option j . Then the *dynamic semiparametric factor model* (DSFM) is defined as follows: assume

$$Y_{t,j} = \mathcal{Z}_t^\top m(X_{t,j}) + \varepsilon_{t,j}, \quad (18)$$

where $\mathcal{Z}_t = (1, Z_t^\top)$, $Z_t = (Z_{t,1}, \dots, Z_{t,L})^\top$ unobservable L -dimensional stochastic process, $m = (m_0, \dots, m_L)^\top$, real-valued functions; m_l , $l = 1, \dots, L + 1$ are defined on a subset of \mathbb{R}^d . One can estimate:

$$\hat{Y}_t = \hat{\mathcal{Z}}_t^\top \hat{m}(X_t) \quad (19)$$

$$= \hat{\mathcal{Z}}_t^\top \hat{\mathcal{A}} \psi(X_t), \quad (20)$$

with $\psi(X_t) \stackrel{\text{def}}{=} \{\psi_1(X_t), \dots, \psi_K(X_t)\}^\top$ being a space basis such as a tensor B-spline basis, \mathcal{A} is the $(L + 1) \times K$ coefficient matrix. In this case K denotes the number of tensor B-spline sites: let $(s_u)_{u=1}^U$, $(s_v)_{v=1}^V$ be the B-spline sites for moneyness and time-to-maturity coordinates, respectively, then $K = U \cdot V$. Given some spline orders n_κ and n_τ for both coordinates and sets of knots $(t_i^\kappa)_{i=1}^M$, $(t_j^\tau)_{j=1}^N$, one of the Schoenberg–Whitney conditions requires that $U = M - n_\kappa$, $V = N - n_\tau$, see deBoor (2001). The usage of the parameter K is roughly analogous to the bandwidth choice in Fengler *et al.* (2003) and Fengler *et al.* (2007); however the results of Park *et al.* (2009) demonstrate insensitivity of DSFM estimation results to the choice of K , n .

The estimates for the IV surfaces \hat{m}_l are re-calculated on a fine 2-dimensional grid of tensor B-spline sites: the estimated

coefficient matrix $\hat{\mathcal{A}}$ is reshaped into a $U \times V \times L + 1$ array of $L + 1$ matrices \hat{A} of dimension $U \times V$. Factor functions m_l can then be estimated as follows:

$$\hat{m}_{l,i,j} = \sum_i^U \sum_j^V \hat{A}_{l,i,j} \psi_{i,\kappa_\kappa}(\kappa_i) \psi_{j,\kappa_\tau}(\tau_j), \quad (21)$$

where κ_κ , κ_τ are knot sequences for the moneyness and time-to-maturity coordinates, respectively.

The estimated factor functions \hat{m}_l together with stochastic factor loadings $\hat{\mathcal{Z}}_t$ are combined into the dynamic estimator of the implied volatility surface:

$$\hat{IV}_{t,i,j} = \hat{m}_{0,i,j} + \sum_{l=1}^L \hat{\mathcal{Z}}_{l,t} \hat{m}_{l,i,j}, \quad (22)$$

where $\hat{\mathcal{Z}}_{l,t}$ can be modeled as a vector autoregressive process. It should be noted that \hat{m}_l and $\hat{\mathcal{Z}}_{l,t}$ are not uniquely defined, so an orthonormalization procedure must be applied.

An indication of possible mispricing of LETF options allows to test a trading strategy based on the comparison of the theoretical price obtained from the moneyness scaling correction as well as the application of the DSFM model and the market price. Such a strategy would mainly exploit the two essential elements of information from these two approaches. The first element is obtaining evidence of statistical discrepancies resulting from the mismatch between ETF and LETF IVs. The moneyness scaling approach allows to estimate LETF IV using richer unleveraged ETF data which also would make the DSFM IV estimator more consistent. The second element is implied volatility forecasting. The DSFM model allows to forecast a whole IV surface via the dynamics of stochastic factor loadings \mathcal{Z}_t .

4.2. Model estimation

The DSFM model is estimated numerically. The number of factors has to be chosen in advance. One should also notice that for m_l to be chosen as eigenfunctions of the covariance operator $K(u, v) \stackrel{\text{def}}{=} \text{Cov}\{Y(u), Y(v)\}$ in an L -dimensional approximating linear space, where Y is understood to be the random IV surface, they should be properly normalized, such that $\|m_l(\cdot)\| = 1$ and $\langle m_l, m_k \rangle = 0$ for $l \neq k$.

The choice of L can be based on the explained variance by factors:

$$EV(L) \stackrel{\text{def}}{=} 1 - \frac{\sum_{t=1}^T \sum_{j=1}^{J_t} \left\{ Y_{t,j} - \sum_{l=0}^L \hat{\mathcal{Z}}_{l,t} \hat{m}_l(X_{t,j}) \right\}^2}{\sum_{t=1}^T \sum_{j=1}^{J_t} (Y_{t,j} - \bar{Y})^2}. \quad (23)$$

The model's goodness-of-fit is evaluated by the root mean squared error (RMSE) criterion:

$$RMSE \stackrel{\text{def}}{=} \sqrt{\frac{1}{\sum_t J_t} \sum_{t=1}^T \sum_{j=1}^{J_t} \left\{ Y_{t,j} - \sum_{l=0}^L \hat{\mathcal{Z}}_{l,t} \hat{m}_l(X_{t,j}) \right\}^2}. \quad (24)$$

The prediction quality at time point $t + 1$ is measured by the root mean squared prediction error (RMSPE) given by

$$RMSPE \stackrel{\text{def}}{=} \sqrt{\frac{1}{J_{t+1}} \sum_{j=1}^{J_{t+1}} \left\{ Y_{t+1,j} - \sum_{l=0}^L \hat{Z}_{t+1,l} \hat{m}_l(X_{t+1,j}) \right\}^2}. \tag{25}$$

5. Empirical application

5.1. Data description

For the purpose of an empirical application, we use data on SPY, SSO, UPRO and SDS (L)ETF call options in the period November 2014–June 2015. The data summary statistics are outlined in table 2. The data were taken from the Datastream database by Thomson Reuters.

To give an impression of leveraged ETF option tradability, we give an illustration of the existing bid spreads and actual trades of the SSO LETF, as these data will be used for the trading strategy example below. Figures 3 and 4 show variation of existing trade prices and volumes for various option contracts based on exercise price and time to expiration. We can see that shorter-term contracts are traded more broadly. It has been also found that trades predominantly occur at or near mid-quotes. In figure 5 we show bid-ask spreads for the same range of option contracts, which tend to be quite high, but somewhat lower for longer-term contracts.

The option data we use for the empirical application are trade-based data, i.e. each observation corresponds to an actual trade, not price quotes or settlement data. Implied volatility and option prices are taken from the database and computed in accordance with standard conventions used by market participants using the midpoint of the best closing bid price and best closing offer price for the options, taking account of liquidity and dividends.

Additionally, we remove data which may contain noise, potential misprints and other errors. Such data include anomalous and outlier data resulting, e.g. from artificial extrapolation of implied volatilities for non-traded options or feature lower liquidity for the out-of-the-money or options which are deeply in-the-money.

5.2. Confidence bands

We use the data described above to construct bootstrap confidence bands for the M -smoother of implied volatility Y given log-moneyness X , according to methodology described in section 2.2. Accordingly, X is transformed using (3). The results are shown in figures 6 and 7 for time-to-maturity 0.5 and 0.6 years, respectively. In figure 8 combined bands are provided.

We can observe clear discrepancy between the implied volatilities of leveraged ETFs and their unleveraged counterpart SPY. For all LETFs, non-overlapping confidence bands imply that there is a statistically significant difference between IV functions at the significance level $\alpha = 0.05$. It is more pronounced for in- and out-of-the-money options. This phenomenon may occur due to lower liquidity of in- and out-of-the-money options compared to at-the-money options. On the other hand, as shown, e.g. in Etlting and Miller (2000), the relationship between option moneyness and liquidity is more complex than quadratic, maximized for at-the-money options. Therefore, liquidity need not be the only reason for this fact.

We can see from figure 8 that the bands for SSO demonstrate particularly strong deviation from those of SPY. This implies that discrepancies not removed by the moneyness scaling procedure are the largest for this LETF. Therefore we conclude this section with a trading strategy which is meant to exploit such statistical discrepancies on the market of SPY and SSO options.

5.3. DSFM estimation and forecasting

The EV , $RMSE$ and $RMSPE$ criteria are displayed in table 3. The model order $L = 3$ is chosen for estimation. The data for the SPY ETF option are used with parameters $n_\kappa, n_\tau = 3$; $M = 9, N = 7$, so that $U = 6, V = 4, K = 6 \times 4 = 24$.

Figure 9 shows the dynamics of \hat{Z}_t in time. Two largest ‘spikes’ in the value of the third stochastic loading in the beginning of the period correspond to the period of relatively large values of the CBOE volatility index (VIX). The second of the ‘spikes’ precedes in time an increase in the VIX value implying that the model has predictive value with respect to market instability dynamics. This shows that DSFM captures leading dynamic effects as well as can explain effects like skew or term structure changes.

Table 2. Summary statistics on (L)ETF options data (τ is time to maturity, LM log-moneyness, σ_I implied volatility).

		Min.	Max.	Mean	Stdd.	Skewn.	Kurt.
SPY	τ	0.258	2.364	1.202	0.515	0.421	2.316
	LM	-3.061	0.477	-0.381	0.574	-1.513	5.871
	σ_I	0.086	2.677	0.271	0.195	3.228	18.522
SSO	τ	0.208	2.236	1.239	0.585	-0.044	1.795
	LM	-1.704	0.558	-0.484	0.461	-0.089	2.264
	σ_I	0.154	1.340	0.363	0.091	1.774	12.224
UPRO	τ	0.208	2.236	1.205	0.585	0.043	1.795
	LM	-1.182	0.665	-0.168	0.331	-0.360	2.719
	σ_I	0.250	1.669	0.503	0.099	1.335	9.080
SDS	τ	0.208	2.236	1.146	0.581	0.196	1.852
	LM	-0.738	0.858	0.187	0.344	-0.276	2.226
	σ_I	0.107	1.262	0.424	0.129	0.792	4.830

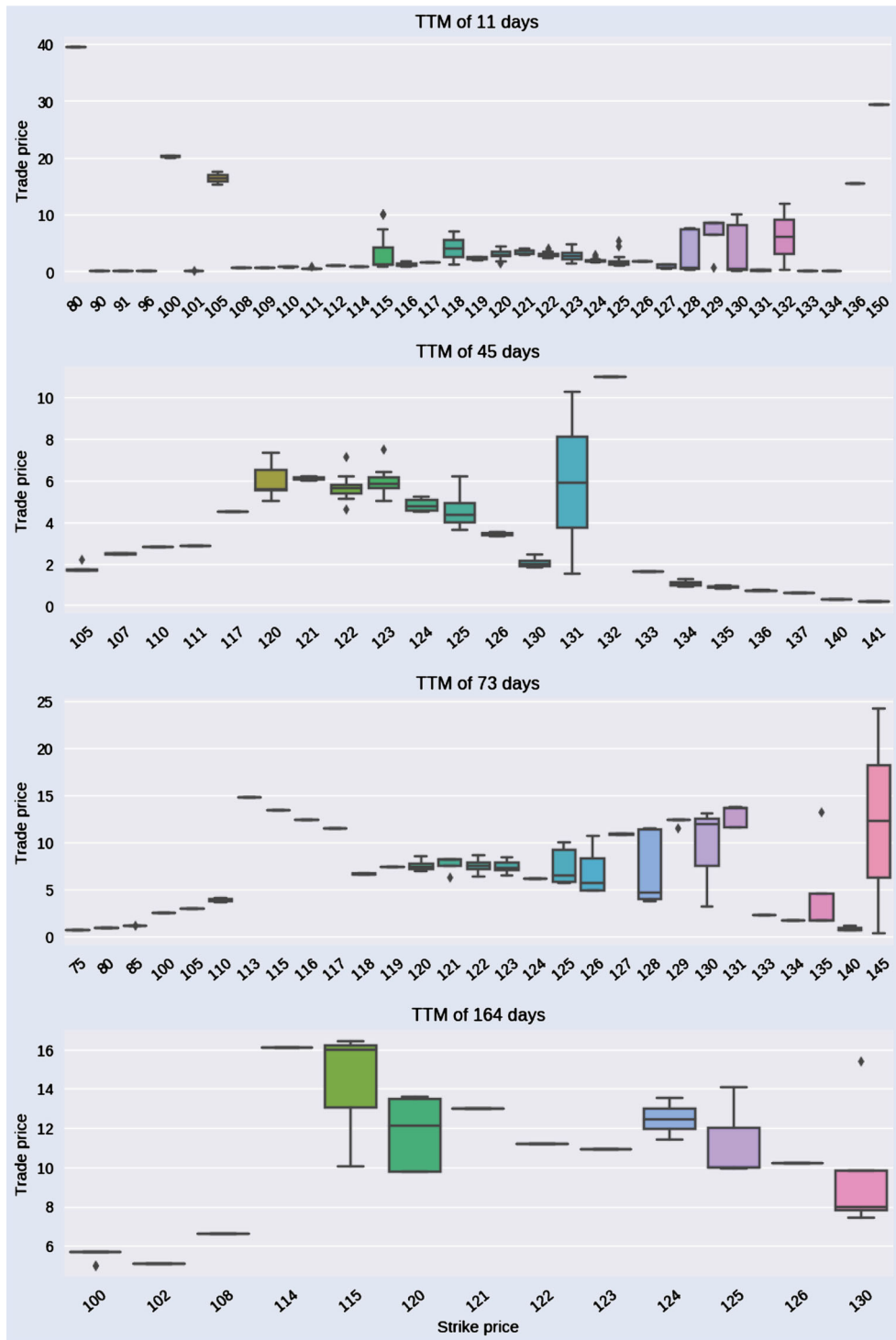


Figure 3. Intraday trade prices for various option contracts on SSO on 6 January, 2015; data source: Option Price Reporting Authority.

Theoretical and simulation results in Park *et al.* (2009) justify using vector autoregression (VAR) analysis to model \hat{Z}_t . To select a VAR model, we computed the Schwarz (SC), the Hannan-Quinn (HQ) and the Akaike (AIC) criteria, as shown in table 4. All three criteria select the VAR(1) model. Furthermore, the roots of the characteristic polynomial all lie inside the unit circle, which shows that the specified model is stationary. Portmanteau and Breusch-Godfrey LM test results with 12 lags for the autocorrelations of the error term fail to reject residual autocorrelation at 10% significance level.

The degenerate nature of implied volatility data is reflected by the fact that empirical observations do not cover estimation grids at given time points. This is due to the fact that contracts at certain maturities or strikes are not always traded. The DSFM fitting procedure introduces basis functions which approximate a high-dimensional space and depend on time. This allows to account for all information in the dataset simultaneously in one minimization procedure which runs over all \hat{m}_l and \hat{Z}_t and avoid bias problems which would inevitably occur if some kernel smoothing procedure such as

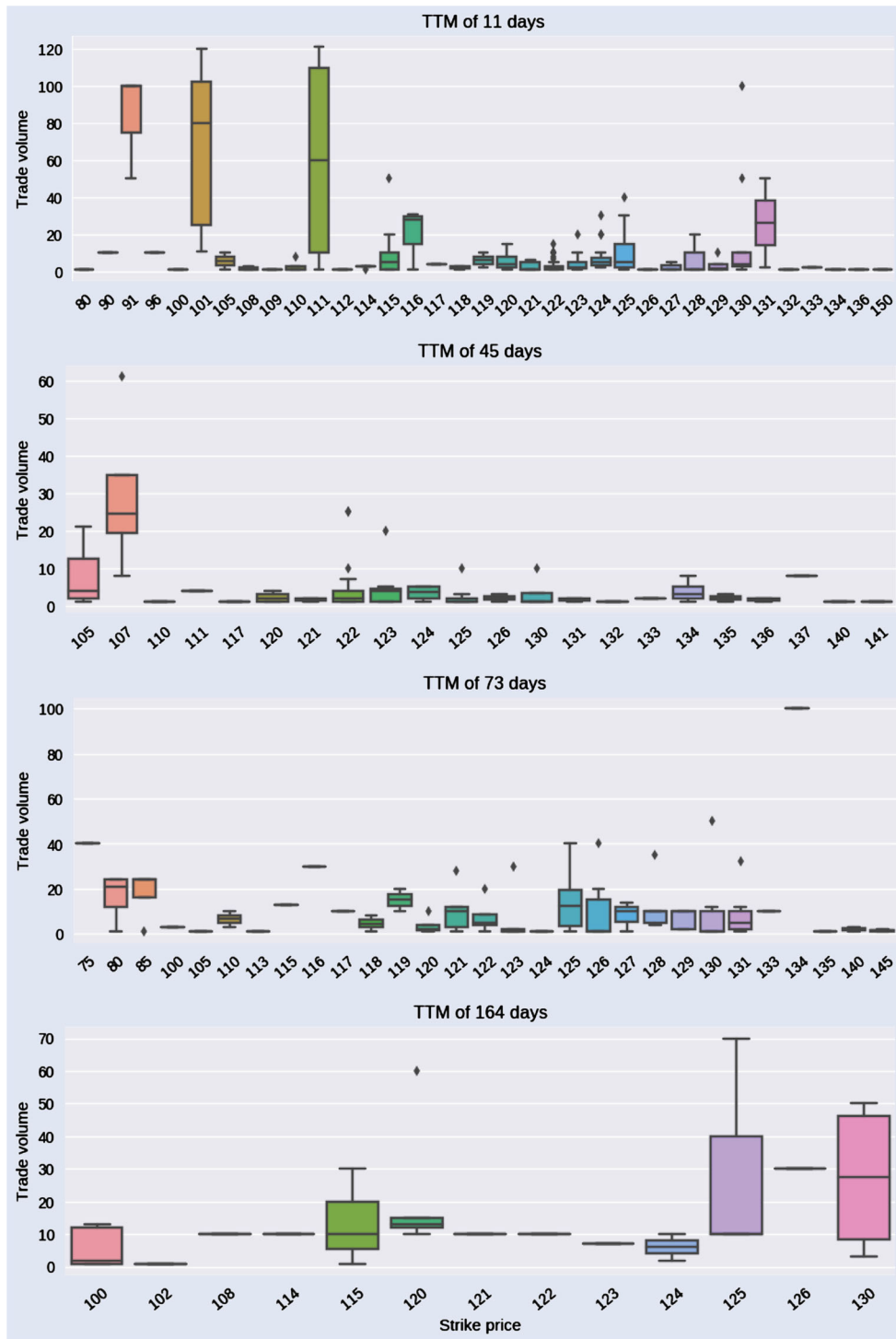


Figure 4. Intraday trade volumes for various option contracts on SSO on 6 January, 2015; data source: Option Price Reporting Authority.

Nadaraya–Watson were applied for this type of degenerate data.

5.4. Option trading strategy

5.4.1. Description. Ability to forecast the whole surface of implied volatility can be used in combination with the money-ness scaling technique to exploit potential discrepancies in ETF and LETF option prices or implied volatilities to build a trading strategy. A suitable strategy would be the so-called

‘trade-with-the-smile/skew’ strategy adapted for the special case of ETF-LETF option IV discrepancy. It would use the ETF option data to estimate the model (theoretical) smile of the leveraged counterpart and the information from the IV surface forecast to recognize the future (one-period-ahead) possible IV discrepancy.

Going back to the results in section 5.2, we see that the largest statistical discrepancy between leveraged and unleveraged ETF implied volatilities is the one between SPY and SSO, so we consider these two options in the strategy

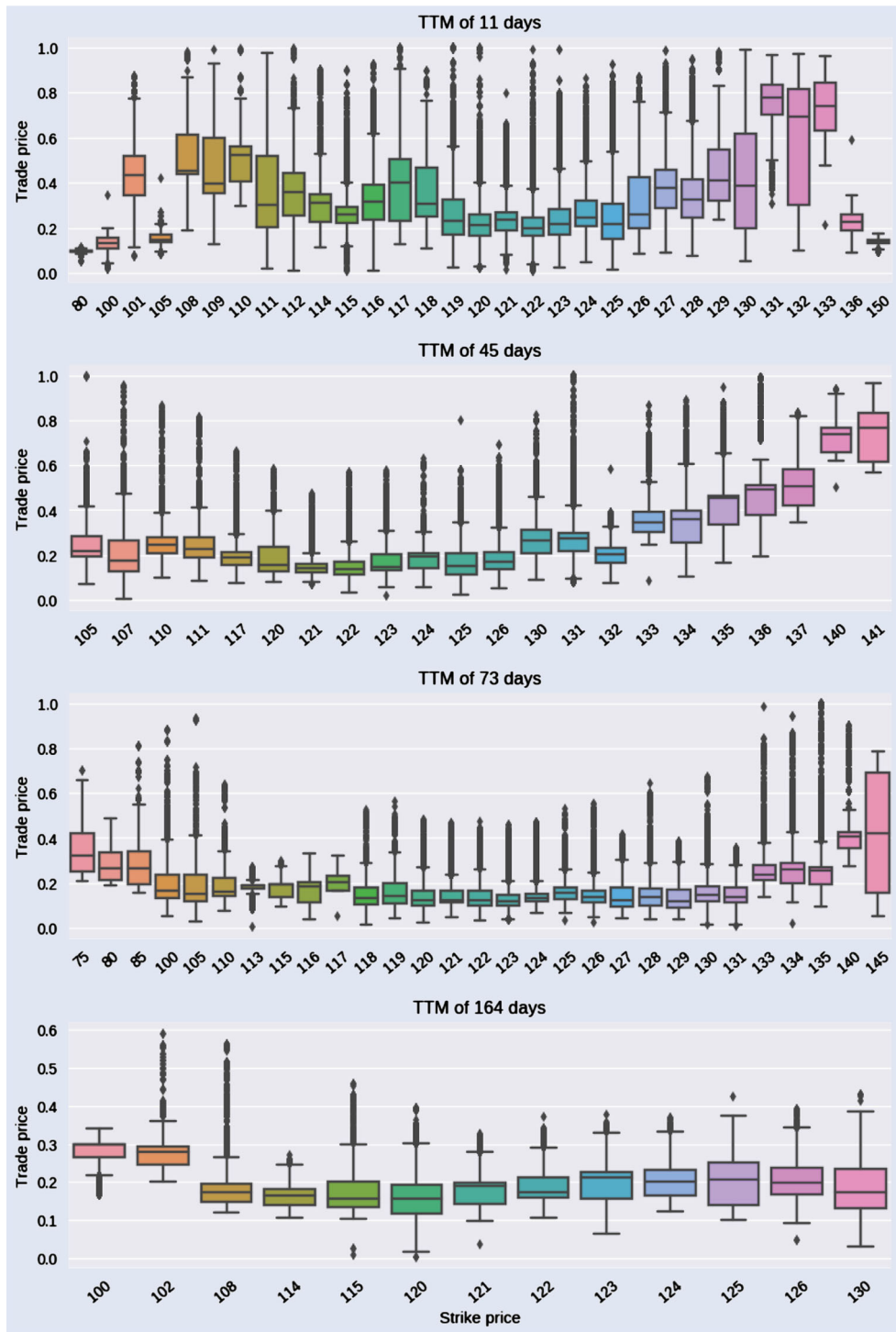


Figure 5. Intraday bid-ask spreads (percentage of the ask price) for various option contracts on SSO on January 6, 2015; data source: Option Price Reporting Authority.

setup. The strategy can be outlined as follows: choose a moving window width w ; then for each $t = w, \dots, T$ (T is the final time point in the sample) do the following:

- (i) given two leverage ratios $\beta_{SPY} = 1, \beta_{SSO} = 2$, re-scale the log-moneyness coordinate $LM^{(\beta_{SPY})}$ according to the moneyness scaling formula (3) to obtain $\widehat{LM}^{(\beta_{SSO})}$. This will be the ‘model’ moneyness coordinate for DSFM estimation,
- (ii) map the space $[\widehat{LM}_{\min}^{(\beta_{SSO})}, \widehat{LM}_{\max}^{(\beta_{SSO})}] \times [\tau_{\min}^{SPY}, \tau_{\max}^{SPY}]$ to $[0, 1] \times [0, 1]$ using marginal transformation,
- (iii) estimate the DSFM model (18) on $[0, 1] \times [0, 1]$. This will yield the IV surface estimates $\widehat{IV}_1^{SSO}, \dots, \widehat{IV}_t^{SSO}$,
- (iv) forecast the IV surface estimate \widehat{IV}_{t+1}^{SSO} using the VAR structure of the estimated stochastic loadings \hat{Z}_t and the factor functions \hat{m}_t ,
- (v) choose a time-to-maturity τ^* at time point t , take the corresponding real-world values of SSO log-moneyness $LM^{(\beta_{SSO})}$ and map them to $[0, 1]$ using the marginal distribution of $\widehat{LM}^{(\beta_{SSO})}$; denote the output as $LM_{\tau^*;M}^{(\beta_{SSO})}$,

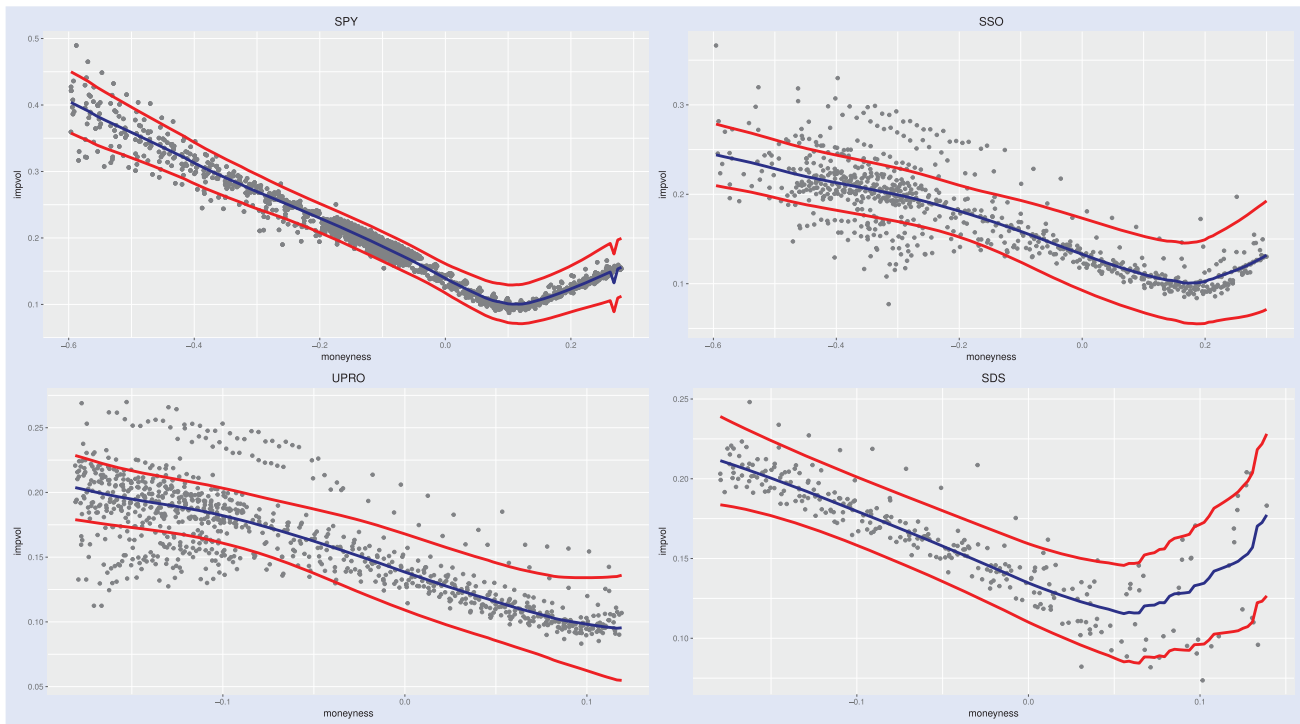


Figure 6. Fitted implied volatility and bootstrap uniform confidence bands for 4 (L)ETFs on S&P500; τ : 0.5 years.

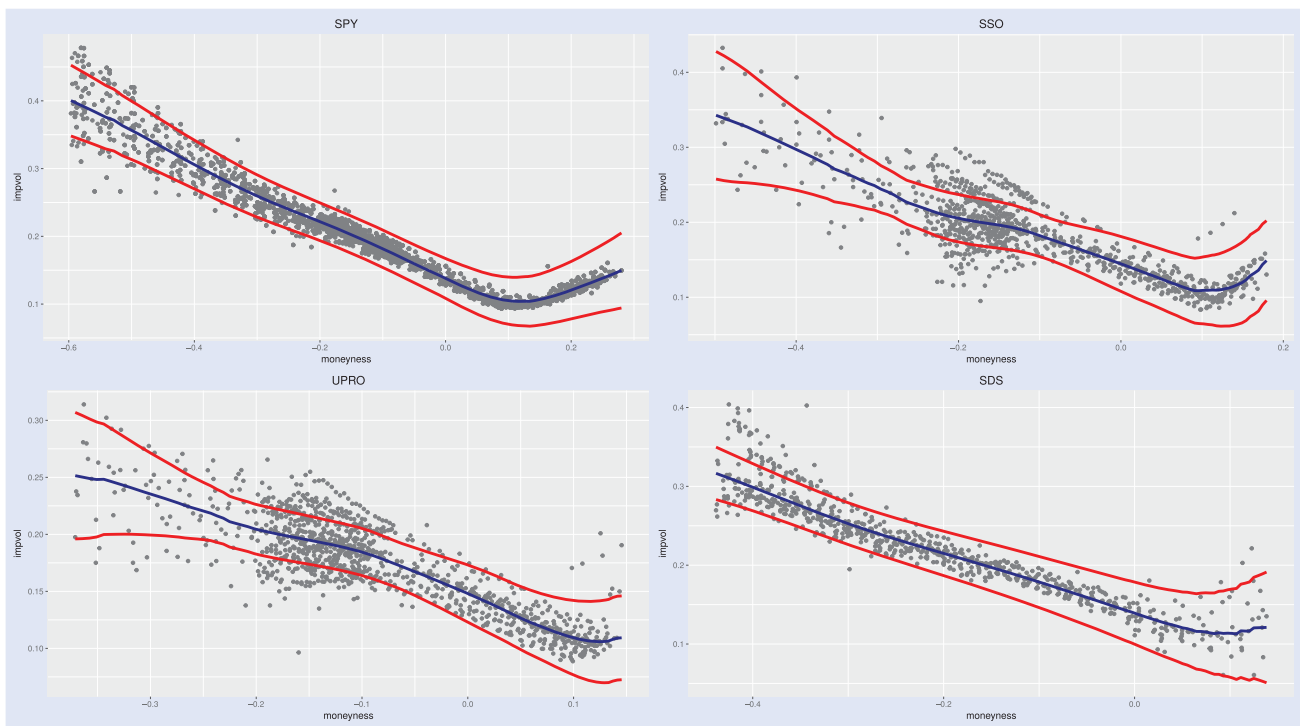


Figure 7. Fitted implied volatility and bootstrap uniform confidence bands for 4 (L)ETFs on S&P500; τ : 0.6 years.

- (vi) using the marginally re-scaled grid $[\widehat{LM}_{\min}^{(\beta_{SSO})}, \widehat{LM}_{\max}^{(\beta_{SSO})}] \times [\tau^*, \tau^*]$ and \widehat{IV}_{t+1}^{SSO} , obtain interpolated values $\widehat{IV}_{t+1;LM_{\tau^*;M}^{(\beta_{SSO})}, \tau^*}^{SSO}$ corresponding to $LM_{\tau^*;M}^{(\beta_{SSO})}, \tau^*$,
- (vii) compare the ‘theoretical’ values $\widehat{IV}_{t+1;LM_{\tau^*;M}^{(\beta_{SSO})}, \tau^*}^{SSO}$ with known real-world implied volatilities $IV_{t;LM_{\tau^*;M}^{(\beta_{SSO})}, \tau^*}^{SSO}$

corresponding to $LM_{\tau^*;M}^{(\beta_{SSO})}$, and construct a delta-hedged option portfolio:

- (a) if $\widehat{IV}_{t+1;LM_{\tau^*;M}^{(\beta_{SSO})}, \tau^*}^{SSO} > IV_{t;LM_{\tau^*;M}^{(\beta_{SSO})}, \tau^*}^{SSO}$ for all $LM_{\tau^*;M}^{(\beta_{SSO})}$, then buy (long) options corresponding to the largest difference $D_{long} \stackrel{\text{def}}{=} \widehat{IV}_{t+1;LM_{\tau^*;M}^{(\beta_{SSO})}, \tau^*}^{SSO} - IV_{t;LM_{\tau^*;M}^{(\beta_{SSO})}, \tau^*}^{SSO}$,

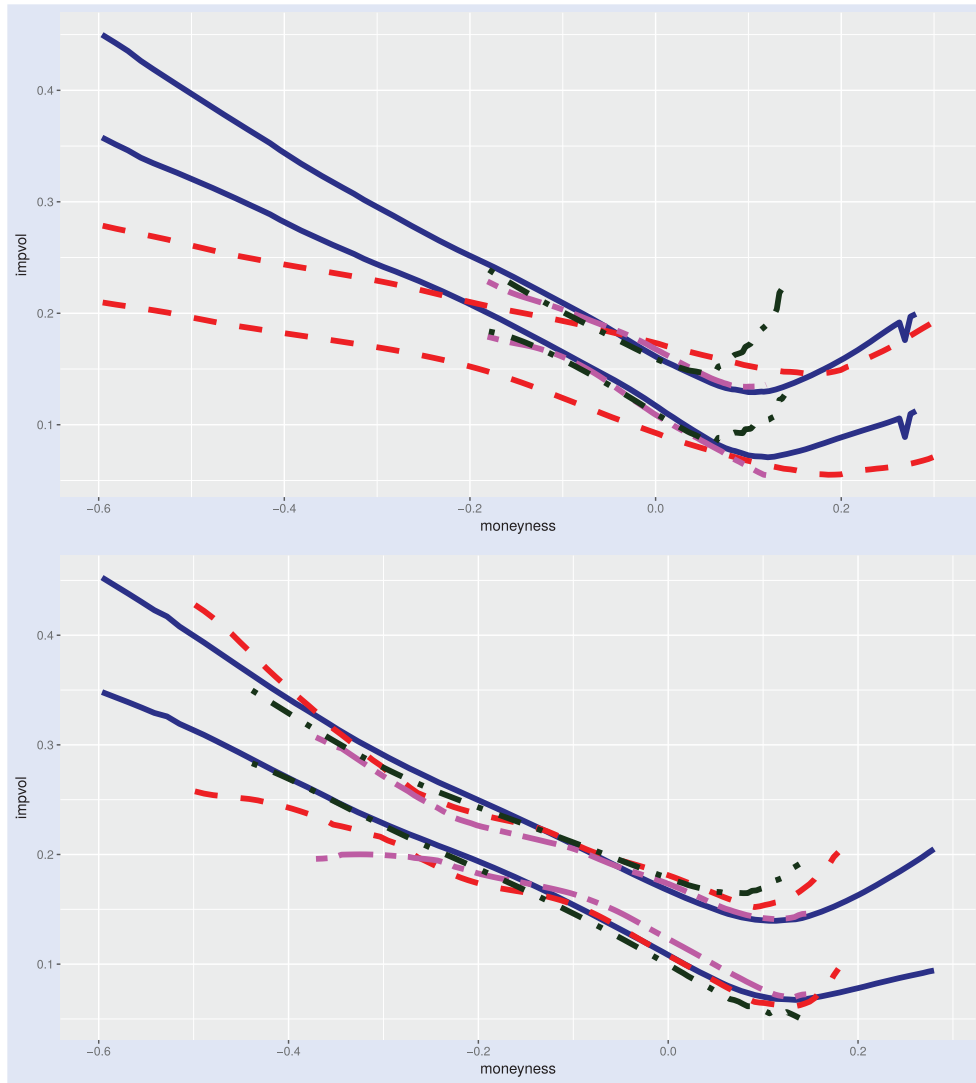


Figure 8. Combined uniform bootstrap confidence bands for SPY, SSO, UPRO and SDS after moneyness scaling ($\tau = 0.5$ (top) and $\tau = 0.6$ years (bottom), respectively).

Table 3. EV, RMSE and RMSPE criteria for different model order sizes.

Criterion	$L = 2$	$L = 3$	$L = 4$	$L = 5$
EV	0.915	0.921	0.925	0.930
RMSE	0.090	0.088	0.087	0.082
RMSPE	0.095	0.096	0.099	0.102

Table 4. The VAR model selection criteria.

Model order n	AIC(n)	HQ(n)	SC(n)
1	-4.20*	-4.10*	-3.96*
2	-4.13	-3.96	-3.72
3	-4.07	-3.83	-3.48
4	-4.03	-3.72	-3.27
5	-3.97	-3.59	-3.03

Note: The smallest value is marked by an asterisk.

(b) if $\widehat{IV}_{t+1;LM_{\tau^*;M}^{(\beta_{SSO}),\tau^*}}^{SSO} < IV_{t;LM_{\tau^*;M}^{(\beta_{SSO}),\tau^*}}^{SSO}$ for all $LM_{\tau^*;M}^{(\beta_{SSO})}$, then sell (short) options corresponding to the

largest difference $D_{short} \stackrel{\text{def}}{=} IV_{t;LM_{\tau^*;M}^{(\beta_{SSO}),\tau^*}}^{SSO} - \widehat{IV}_{t+1;LM_{\tau^*;M}^{(\beta_{SSO}),\tau^*}}^{SSO}$

(c) if it holds that both $\widehat{IV}_{t+1;LM_{\tau^*;M}^{(\beta_{SSO}),\tau^*}}^{SSO} > IV_{t;LM_{\tau^*;M}^{(\beta_{SSO}),\tau^*}}^{SSO}$ and $\widehat{IV}_{t+1;LM_{\tau^*;M}^{(\beta_{SSO}),\tau^*}}^{SSO} < IV_{t;LM_{\tau^*;M}^{(\beta_{SSO}),\tau^*}}^{SSO}$ for different $LM_{\tau^*;M}^{(\beta_{SSO})}$, then buy (long) options with the largest D_{long} and sell (short) options with the largest D_{short} . In all three cases use the underlying SSO LETF asset to make the portfolio delta-neutral,

(viii) at time point $t + 1$, terminate the portfolio via an offsetting sale/purchase, calculate profit/loss and repeat until time T .

The strategy described above aims to exploit the information from the statistical discrepancies between the forecast ‘theoretical’ (model) SSO LETF implied volatilities and the historical (‘true’) ones. It protects the portfolio against unfavorable moves in the underlying asset L_t through delta-hedging and aims to gain from forecast moves in another

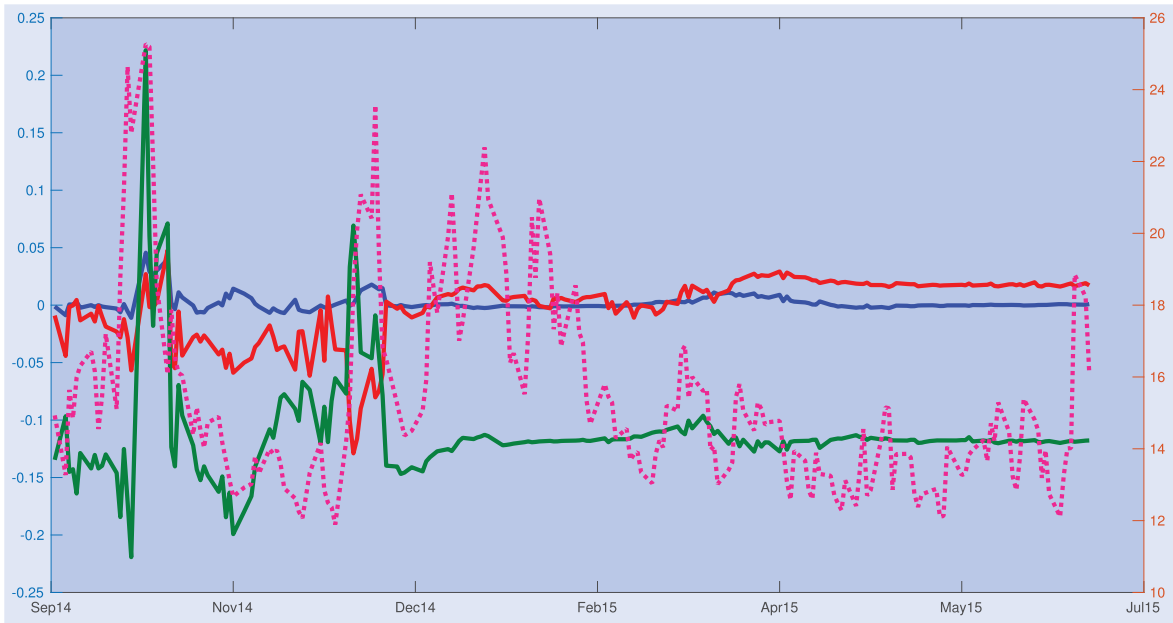


Figure 9. Time dynamics of $\hat{Z}_{t,1}$, $\hat{Z}_{t,2}$, $\hat{Z}_{t,3}$, VIX index.

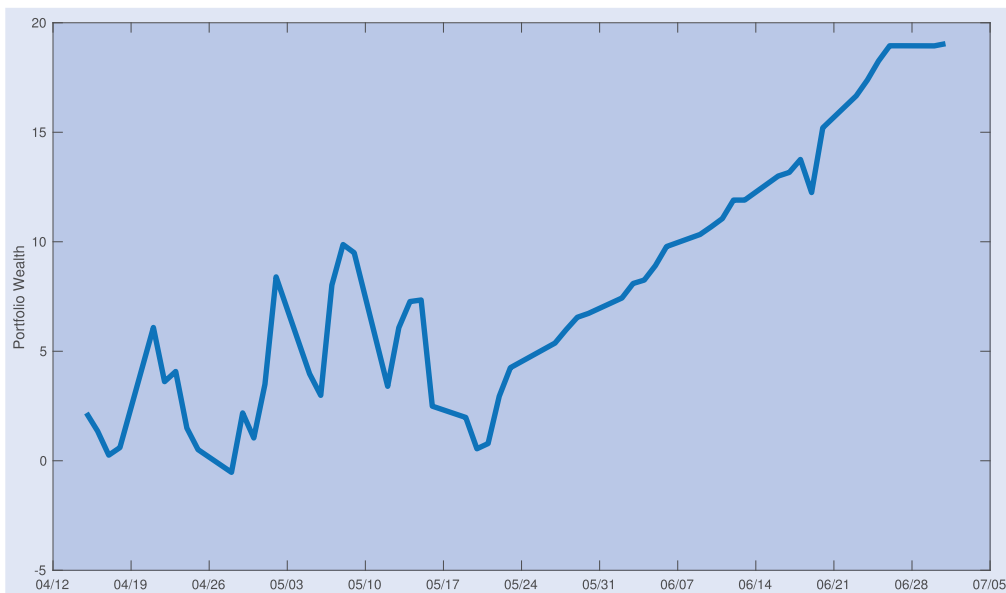


Figure 10. Cumulative performance of the trading strategy.

option risk factor, the implied volatility via its explicit estimation and forecasting.

The key transformation $\widehat{LM}^{(\beta_{SSO})}$ can be also perceived to drive a statistical equilibrium for IV_{t+1}^{SSO} through \widehat{IV}_{t+1}^{SSO} , deviations from which induce entry and exit points for trading. This reasoning is in line with that of Avellaneda and Lee (2010) who introduce a model-driven pairs-trading strategy in US equities.

The real-world SSO implied volatility $IV_{t;LM_{\tau^*;M}^{(\beta_{SSO})},\tau^*}^{SSO}$ at time step t is expected to converge to the forecast implied volatility $\widehat{IV}_{t+1;LM_{\tau^*;M}^{(\beta_{SSO})},\tau^*}^{SSO}$, which has been constructed using scaled moneyness and implied volatility input from the unleveraged LETF, i.e. SPY.

Such a strategy would fall into the class of model-driven statistical arbitrage in equities and equity options. It has the three characteristic features of statistical arbitrage described by Avellaneda and Lee (2010): (i) trading signals are systematic or rules-based, as opposed to driven by fundamentals, (ii) the trading portfolio is market-neutral, i.e. has zero sensitivity to the market, and (iii) the algorithm for generating excess returns is statistical. The market here is defined by the SSO LETF and delta neutrality implies market neutrality.

In the following section, we present the strategy's performance as well as a numerical example. It occurs that the existing statistical discrepancies between implied volatilities of leveraged and unlevered ETF options together with

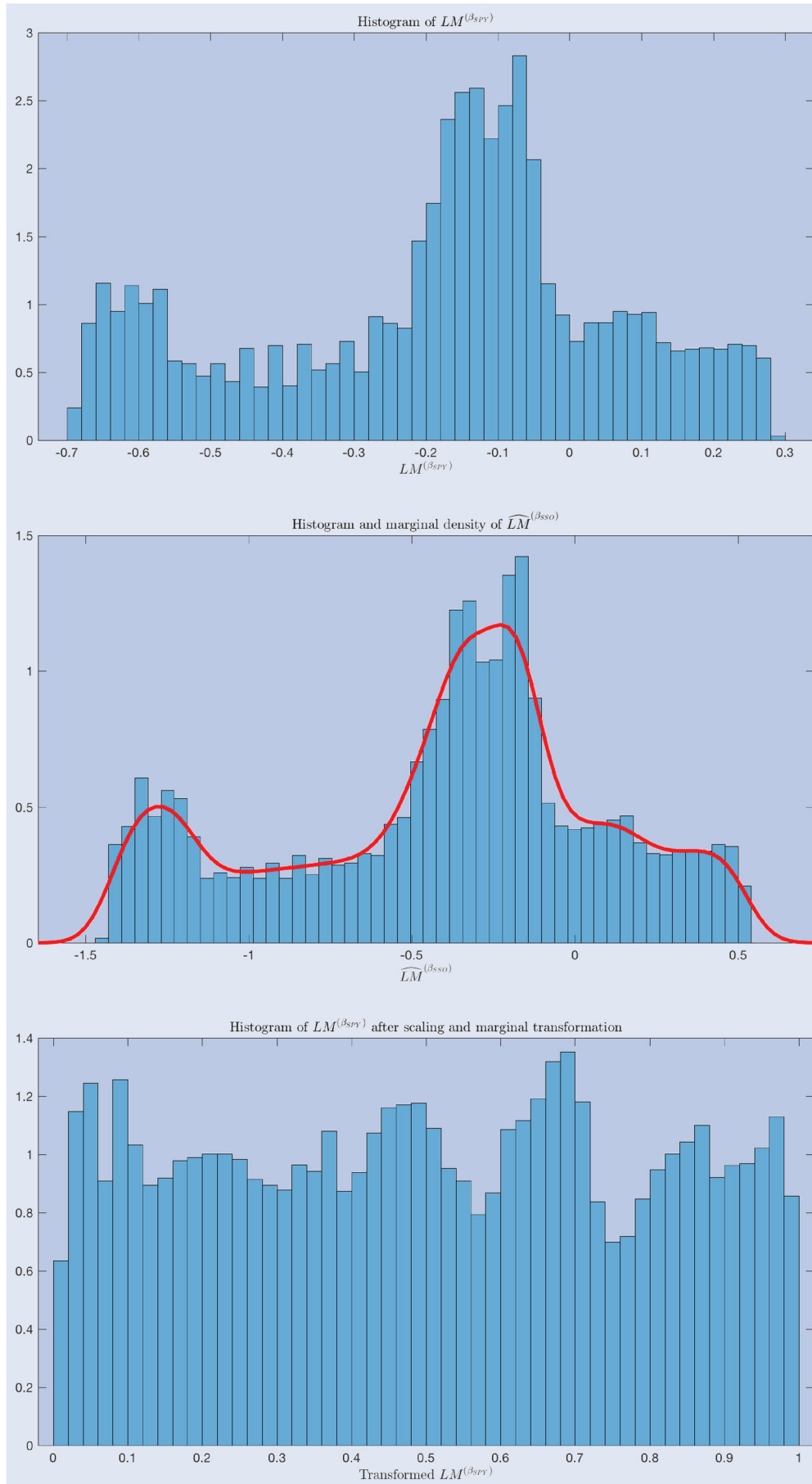


Figure 11. $LM^{(\beta_{SPV})}$ (top panel), $\widehat{LM}^{(\beta_{SSO})}$ (middle panel), $\widehat{LM}^{(\beta_{SSO})}$ after marginal transformation (bottom panel).

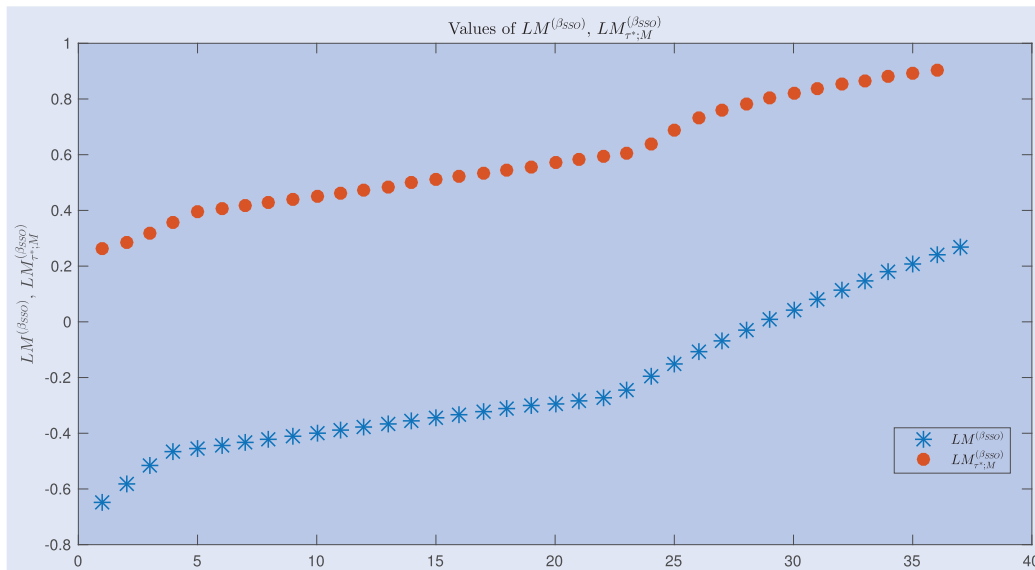


Figure 12. Values of $LM^{(\beta_{SSO})}$ and $LM_{\tau^*;M}^{(\beta_{SSO})}$.

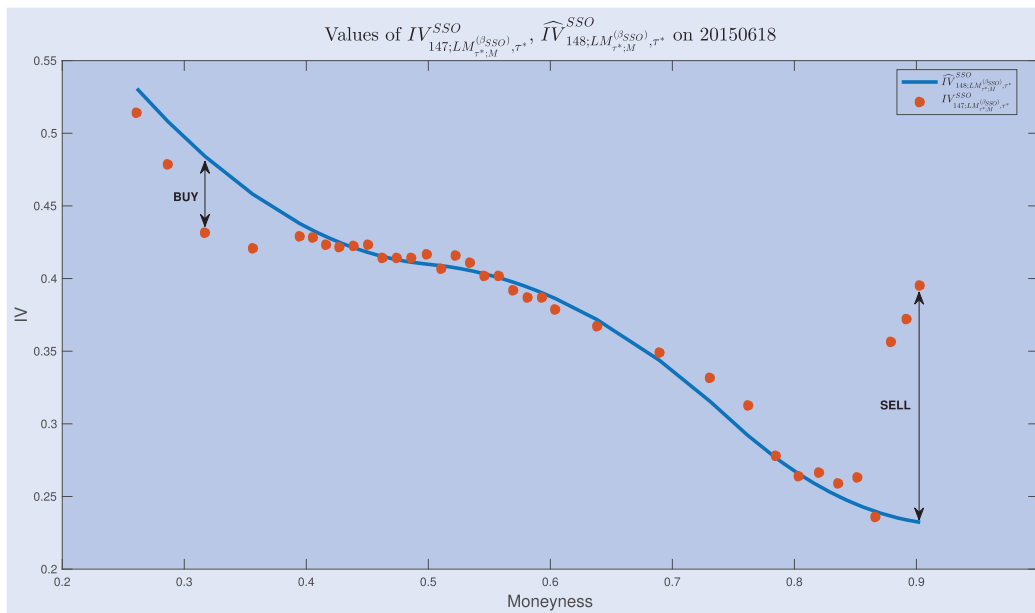


Figure 13. Values of real-world $IV_{147; LM_{\tau^*; M}^{(\beta_{SSO})}, \tau^*}^{SSO}$ and predicted $\widehat{IV}_{148; LM_{\tau^*; M}^{(\beta_{SSO})}, \tau^*}^{SSO}$; the IV points at which long and short trades are done, are indicated by arrows.

predictive capacity of the DSFM model can provide non-negative trading gains on the option market.

5.4.2. Numerical example. For the purpose of the estimation of the strategy from the previous section, the DSFM model parameters are taken to be the same as in section 5.3. The rolling window width is assumed to be $w = 100$ and the forecasting horizon is 1 day ahead.

The dynamic strategy performance in the period April 2015 – June 2015 is displayed in figure 10. Out of 55 investment periods, in 30 cases long-only portfolios were constructed, the remaining 25 cases short and long positions were taken; net portfolios were short portfolios in 42 cases, long portfolios in the remaining 13 cases.

For the sake of illustration, let us go through one step from the outlined strategy in a numerical example. Assume that we are at the step $t = 147$ of the sample, which corresponds to June 18, 2015. At this point, we have a training sample of 100 days for DSFM estimation, encompassing 14,859 observations of the option data for contracts with various strike prices and time-to-maturity. The histogram and density estimates for SPY log-moneyness $LM^{(\beta_{SPY})}$, ‘theoretical’ SSO log-moneyness $\widehat{LM}^{(\beta_{SSO})}$ (that is, rescaled $LM^{(\beta_{SPY})}$) and its marginally transformed version are given in figure 11. Further we proceed as proposed in the strategy above:

- (i) estimate (18) and perform a forecast to obtain \widehat{IV}_{148}^{SSO} on June 19, 2015 (day 148 in the sample),

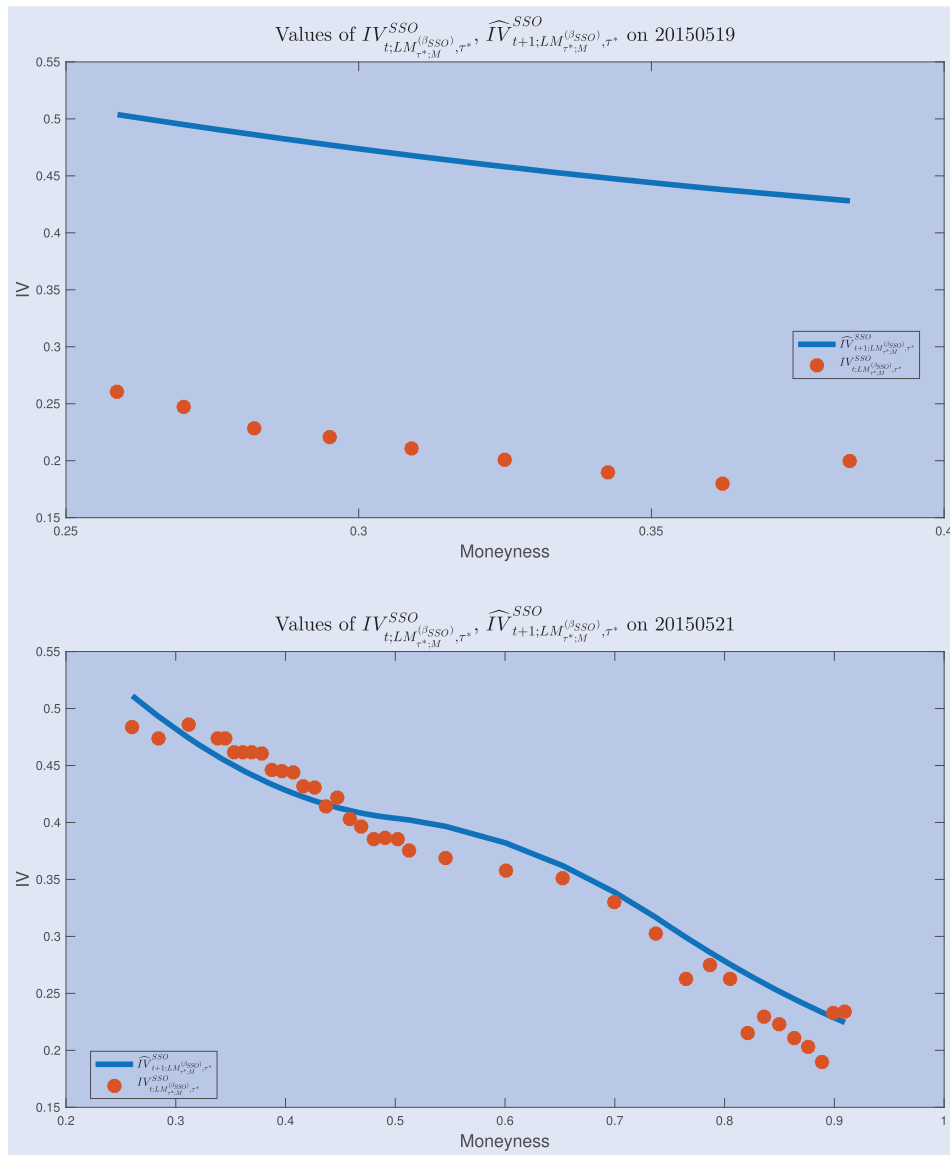


Figure 14. Model and real-world IVs before (top panel) and after (bottom panel) the split on May 20, 2015.

- (ii) choose $\tau^* = 0.6$; we have 37 values of $LM^{(\beta_{SSO})}$ for $\tau^* = 0.6$. We use the marginal distribution of $\widehat{LM}^{(\beta_{SSO})}$ shown in figure 11 to calculate the corresponding ‘theoretical’ values $LM_{\tau^*;M}^{(\beta_{SSO})}$ implied by SPY data and the moneyness scaling procedure, both shown in figure 12,
- (iii) using the forecast IVS \widehat{IV}_{148}^{SSO} , we can determine ‘theoretical’ IV values corresponding to $LM_{\tau^*;M}^{(\beta_{SSO})} \in [0, 1]$, $\widehat{IV}_{148;LM_{\tau^*;M}^{(\beta_{SSO})},\tau^*}^{SSO}$ and the real-world IV values $IV_{147;LM_{\tau^*;M}^{(\beta_{SSO})},\tau^*}^{SSO}$ corresponding to the same $LM_{\tau^*;M}^{(\beta_{SSO})}$ through the mapping of $LM^{(\beta_{SSO})}$, described above,
- (iv) the resulting $\widehat{IV}_{148;LM_{\tau^*;M}^{(\beta_{SSO})},\tau^*}^{SSO}$ and $IV_{147;LM_{\tau^*;M}^{(\beta_{SSO})},\tau^*}^{SSO}$ are demonstrated in figure 13. We buy an option corresponding to the largest $D_{long} = 0.052$ at the closing price $C_{long;147} = \$27.375$ and sell an option corresponding to the largest $D_{short} = 0.163$ at the closing price $C_{short;147} = \$2.740$,
- (v) additionally, we delta-hedge the resulting option portfolio using portfolio delta $\Delta_{Port f;147} = \Delta_{long;147} -$

$\Delta_{short;147} = 0.859 - 0.461 = 0.398$ and the underlying SSO which has a closing price of $L = \$66.960$; the total portfolio value on June 18, 2015 is: $C_{long;147} - C_{short;147} - \Delta_{Port f;147} \times L_{147} = \$27.375 - \$2.740 - 0.398 \times \$66.960 = -2.015$,

(vi) at the start of the next day, i.e. June 19, 2015, we check the prices of the options $C_{long;148}$ and $C_{short;148}$ as well as the price of the underlying L_{148} . We find that on June 19, 2015, $C_{long;148} = \$28.450$ and $C_{short;148} = \$0.320$, $L_{148} = \$68.300$. With the share of L_{148} still equal to the previous-day portfolio delta $\Delta_{Port f;147}$, the portfolio is now worth $C_{long;148} - C_{short;148} - \Delta_{Port f;147} \times L_{148} = \$28.450 - \$0.320 - 0.398 \times \$68.300 = 0.947$. We sell that portfolio in an offsetting trade and have secured a gain of $\$2.962$. As expected, C_{long} gained in value while C_{short} went down in value. The coupled gain was larger than an offsetting loss of $\$0.533$ from the delta hedge which resulted in the total gain.

In the end, the cumulative gain of the strategy applied daily as shown in the illustrating example above, is 19.043

Table 5. Stationarity tests' statistics for SPY, SSO price series.

Lags	SPY			SSO		
	PP	ADF	KPSS	PP	ADF	KPSS
1	-3.621**	-3.505**	0.216**	-3.685**	-3.578**	0.200**
2	-3.693**	-3.593**	0.157**	-3.784**	-3.779**	0.145*
3	-3.752**	-3.725**	0.127*	-3.855**	-3.940**	0.118
4	-3.730**	-3.402*	0.110	-3.837**	-3.601**	0.102
5	-3.716**	-3.400*	0.098	-3.824**	-3.622**	0.092
6	-3.664**	-3.133	0.091	-3.770**	-3.360*	0.085
7	-3.610**	-2.970	0.085	-3.706**	-3.152*	0.081
8	-3.575**	-2.908	0.081	-3.660**	-3.094	0.077
9	-3.537**	-2.897	0.078	-3.613**	-3.103	0.074
10	-3.513**	-3.003	0.075	-3.572**	-3.125	0.072

Notes: PP: Phillips–Perron test; ADF: augmented Dickey–Fuller test; KPSS: KPSS test for trend stationarity. * * *, **, *: significant on 1%, 5%, 10% level, respectively

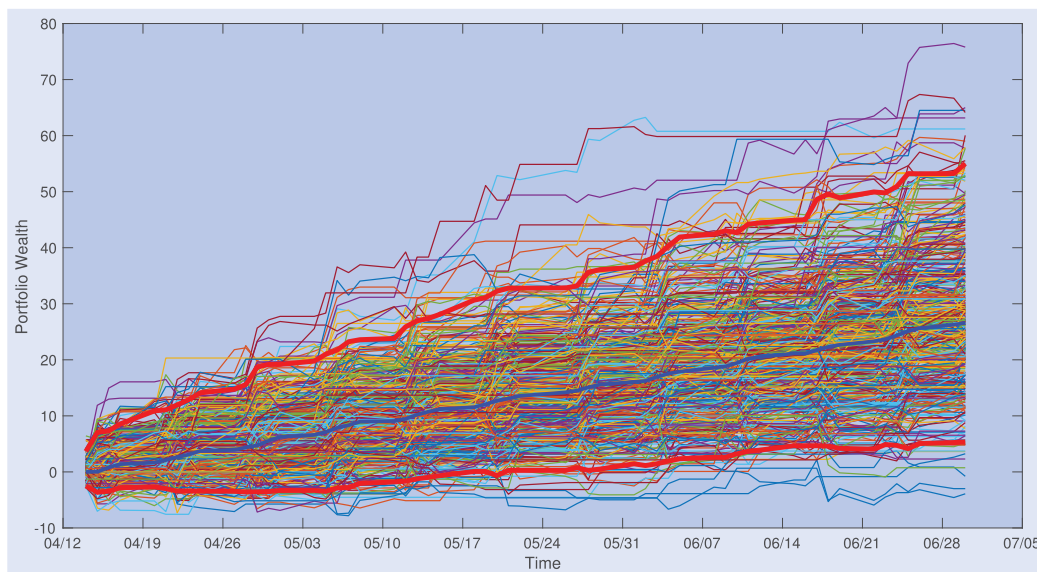


Figure 15. Bootstrapped performance of the strategy; in red: 2.5% and 97.5% empirical percentiles of the cumulative performance; in blue: median of the cumulative performance.

after 55 investment periods. It occurs that 39 out of 55 investment decisions correctly determined the direction of one-step-ahead implied volatility smile change. In the strategy, this smile change is anticipated according the relation between the ‘model’ IVS computed using the moneyness scaling approach and the real-world IVS of a LETF option. There is a high positive chance of generating positive cumulative returns exploiting statistical deviations of leveraged and unlevered implied volatility smiles in the ETF option market.

It should be mentioned that the sample period includes the day of an underlying SSO stock 2-for-1 split which took place on May 20, 2015. The split was implemented to attract a wider range of buyers at the resulting lower price per share. It has been shown by many researchers, such as Ohlson and Penman (1985), Sheikh (1989) and Desai *et al.* (1998) that stock splits result in post-split increases of implied stock volatilities. For instance, Ohlson and Penman (1985) show that stock splits cause short-term increases in volatility upon announcement and long-term increases in volatility after the date the split is effective.

In figure 14, we show real-world and DSFM-forecast IVS on two different dates: before (19 June, 2015) and after the split (21 June, 2015). It can be seen that the model anticipates a significant increase in implied volatilities after the split which indeed takes place.

5.4.3. Robustness check. The performance of the option trading strategy obtained in section 5.4.2 above may seem to have occurred purely by chance. Therefore some sort of a robustness check is necessary.

We perform a bootstrap resampling exercise on the time series of the underlying prices of SPY and SSO (L)ETFs and re-run the strategy on the resampled data. Overlapping block bootstrap approach proposed by Künsch (1989) is applied. It works as follows: given the data observations $\{X_i : i = 1, \dots, T\}$, a block size b is specified. With overlapping blocks of length b , block 1 is then observations $\{X_j : j = 1, \dots, b\}$, block 2 is $\{X_{j+1} : j = 1, \dots, b\}$ and so on. Random sampling is then performed on the level of blocks.

Overlapping block bootstrap assumes the data to be stationary. However, in the current case we do not do inference on the resampled data, so this stringent assumption is less relevant here. Nevertheless, we run standard stationarity tests on SPY and SSO price series in the period from November 2014 to June 30, 2015 such as Phillips–Perron, augmented Dickey–Fuller and KPSS tests. The first two tests have presence of unit root in the series as a null hypothesis, while the KPSS test tests trend stationarity as the null.

We perform a series of tests for each approach using the number of lags from 1 to 10 in the Newey–West estimator of the long-run variance. In table 5 the results of stationarity tests are demonstrated. Most of this evidence does not reject the hypothesis of possible trend stationarity of price series of the (L)ETFs. Therefore we proceed with the bootstrap.

We run 500 bootstrap iterations on 2-dimensional series of SPY and SSO prices in the period from November 2014 to June 30, 2015 and take the block size equal to 5. In figure 15 cumulative performance of the strategy on bootstrapped time series is shown. At each of the 155 time steps the values of 2.5% and 97.5% empirical percentiles are found. We can see that at the end of the period, positive performance occurs with more than a 95% probability. Positive performance with this probability occurs from period 23 onwards until the end of the test sample, which yields 32 periods out of 55 in total.

6. Conclusion

In this paper, we provide statistical and econometric analysis of the moneyness scaling transformation for leveraged and unlevered exchange-traded funds' options' implied volatility smiles. This transformation adjusts the moneyness coordinate of the smile in an attempt to remove the discrepancy between the levered and unlevered counterparts.

We incorporate stochastic volatility into the moneyness scaling method by explicit estimation of the conditional expectation of the realized variance. We present two approaches to implement this estimate: via an analytical approach and using a Monte-Carlo method.

We construct bootstrap uniform confidence bands which reveal a statistically significant discrepancy between the implied volatility smiles, even after moneyness scaling has been performed. We find that this discrepancy is stronger for in- and out-of-the-money options which, however, is unlikely to be explained by liquidity issues alone.

This discrepancy allows one to define a theoretical statistical equilibrium value of LETF moneyness. Based on deviations from this equilibrium, possible trading gain opportunities on the (L)ETF market which can be exploited. We construct a trading strategy based on a dynamic semiparametric factor model. This model-based statistical arbitrage strategy utilizes the dynamic structure of implied volatility surface allowing out-of-sample forecasting and information on unlevered ETF options to construct theoretical one-step-ahead implied volatility surfaces.

The proposed strategy has the potential to generate trading gains due to simultaneous use of the information from the discrepancies between the forecast 'theoretical' (model) SSO LETF implied volatilities and the historical ('true') ones. It protects the portfolio against unfavorable moves in the underlying asset through delta-hedging and aims to gain from forecast moves in volatility. The strategy is shown via bootstrap technique to generate positive returns with a high probability.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was supported by the Deutsche Forschungsgemeinschaft under Grant 'IRTG 1792'.

ORCID

S. Nasekin  <http://orcid.org/0000-0002-5236-7029>

References

- Ahn, A., Haugh, M. and Jain, A., Consistent pricing of options on leveraged ETFs. *SIAM J. Financ. Math.*, 2015, **6**(1), 559–593.
- Ait-Sahalia, Y., Bickel, P.J. and Stoker, T.M., Goodness-of-fit tests for kernel regression with an application to option implied volatilities. *J. Econom.*, 2001, **105**(2), 363–412.
- Avellaneda, M. and Lee, J.-H., Statistical arbitrage in the US equities market. *Quant. Finance*, 2010, **10**(7), 761–782.
- Avellaneda, M. and Zhang, S., Path-dependence of leveraged ETF returns. *SIAM J. Financ. Math.*, 2010, **1**(1), 586–603.
- Broadie, M. and Kaya, Ö., Exact simulation of stochastic volatility and other affine jump diffusion processes. *Oper. Res.*, 2006, **54**(2), 217–231.
- Cizek, P., Härdle, W. and Weron, R., *Statistical Tools for Finance and Insurance*, 2nd ed., 2011 (Springer-Verlag: Heidelberg).
- Cont, R. and da Fonseca, J., Dynamics of implied volatility surfaces. *Quant. Finance*, 2002, **2**(1), 45–60.
- Cox, J.C., Ingersoll, J.E. and Ross, S.A., A theory of the term structure of interest rates. *Econometrica*, 1985, **53**(2), 385–407.
- de Boor, C., *A Practical Guide to Splines*, Revised ed., 2001 (Springer-Verlag: New York).
- Desai, A.S., Nimalendran, M. and Venkataraman, S., Changes in trading activity following stock splits and their effect on volatility and the adverse-information component of the bid-ask spread. *J. Financ. Res.*, 1998, **21**(2), 159–183.
- Dragulescu, A. and Yakovenko, V., Probability distribution of returns in the Heston model with stochastic volatility. *Quant. Finance*, 2002, **2**, 443–453.
- Etling, C. and Miller, T.W., The relationship between index option moneyness and relative liquidity. *J. Futures Mark.*, 2000, **20**(10), 971–987.
- Fengler, M.R., Härdle, W.K. and Mammen, E., A semiparametric factor model for implied volatility surface dynamics. *J. Financ. Econom.*, 2007, **5**(2), 189–218.

Fengler, M.R., Härdle, W.K. and Villa, C., The dynamics of implied volatilities: A common principal components approach. *Rev. Deriv. Res.*, 2003, **6**, 179–202.

Figuerola-Lopez, J.E., Gong, R. and Lorig, M., Short-time expansions for call options on leveraged ETFs under exponential Lévy models with local volatility. Preprint, 2016.

Glasserman, P. and Kim, K.-K., Gamma expansion of the Heston stochastic volatility model. *Finance Stoch.*, 2011, **15**(2), 267–296.

Hall, P. and Horowitz, J., A simple bootstrap method for constructing nonparametric confidence bands for functions. *Ann. Statist.*, 2013, **41**(4), 1892–1921.

Härdle, W., Asymptotic maximal deviation of M -smoothers. *J. Multivar. Anal.*, 1989, **29**, 163–179.

Härdle, W., *Applied Nonparametric Regression*, 1990 (Cambridge University Press: Cambridge).

Härdle, W., Ritov, Y. and Wang, W., Tie the straps: Uniform bootstrap confidence bands for semiparametric additive models. *J. Multivar. Anal.*, 2015, **134**, 129–145.

Heston, S.L., A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Rev. Financ. Stud.*, 1993, **6**, 327–343.

Higham, D.J. and Mao, X., Convergence of Monte Carlo simulations involving the mean-reverting square root process. *J. Comput. Finance*, 2005, **8**(3), 35–61.

Huber, P.J., Robust estimation of a location parameter. *Ann. Math. Statist.*, 1964, **35**(1), 73–101.

Hull, J.C. and White, A., The pricing of options on assets with stochastic volatilities. *J. Finance*, 1987, **42**, 281–300.

Künsch, H.R., The jackknife and the bootstrap for general stationary observations. *Ann. Statist.*, 1989, **17**(3), 1217–1241.

Lee, R. and Wang, R., How leverage shifts and scales a volatility skew: Asymptotics for continuous and jump dynamics. Preprint, 2015.

Leung, T., Lorig, M. and Pascucci, A., Leveraged ETF implied volatilities from ETF dynamics. *Math. Finance*, 2017, **27**(4), 1035–1068.

Leung, T. and Santoli, M., *Leveraged Exchange-Traded Funds: Price Dynamics and Options Valuation*. SpringerBriefs in Quantitative Finance, 2016 (Springer-Verlag: Heidelberg).

Leung, T. and Sircar, R., Implied volatility of leveraged ETF options. *Appl. Math. Finance*, 2015, **22**(2), 162–188.

Lord, R., Koekkoek, R. and Van Dijk, D., A comparison of biased simulation schemes for stochastic volatility models. *Quant. Finance*, 2010, **10**(2), 177–194.

Ohlson, J.A. and Penman, S.H., Volatility increases subsequent to stock splits: An empirical aberration. *J. Financ. Econ.*, 1985, **14**(2), 251–266.

Park, B.U., Mammen, E., Härdle, W. and Borak, S., Time series modelling with semiparametric factor dynamics. *J. Amer. Statist. Assoc.*, 2009, **104**(485), 284–298.

Rouah, F., *The Heston Model and its Extensions in Matlab and C#*, 2013 (Wiley Finance: Hoboken).

Ruppert, D. and Wand, M.P., Multivariate locally weighted least squares regression. *Ann. Statist.*, 1994, **22**(3), 1346–1370.

Schöbel, R. and Zhu, J., Stochastic volatility with an Ornstein-Uhlenbeck process: An extension. *Eur. Finance Rev.*, 1999, **3**, 24–46.

Sheikh, A., Stock splits, volatility increases, and implied volatilities. *J. Finance*, 1989, **44**(5), 1361–1372.

van der Stoep, A.W., Grzelak, L.A. and Oosterlee, C.W., The Heston stochastic-local volatility model: Efficient Monte-Carlo simulation. *Int. J. Theoret. Appl. Finance*, 2014, **17**(7), 1–30.

Appendix

A.1. The local linear M -smoothing estimator

M -type smoothers apply a non-quadratic loss function $\rho(\cdot)$ to make estimation more robust. Given the model

$$Y_i = m(X_i) + \varepsilon_i, \tag{A1}$$

where $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^d$, $\varepsilon_i \stackrel{\text{def}}{=} \sigma(X_i)u_i$, $u_i \sim (0, 1)$, i.i.d, $\mathcal{X} \stackrel{\text{def}}{=} \{(X_i, Y_i); 1 \leq i \leq n\}$, the local linear M -smoothing estimator is obtained from:

$$\min_{\alpha \in \mathbb{R}, \beta \in \mathbb{R}^p} \sum_{i=1}^n \rho \left\{ Y_i - \alpha - \beta^\top (X_i - x) \right\} W_{ih}(x), \tag{A2}$$

where

$$W_{hi}(x) \stackrel{\text{def}}{=} \frac{h^{-2} K' \{ (x - X_i) / h \}}{\hat{f}_h(x)} - \frac{K_h(x - X_i) \hat{f}'_h(x)}{\hat{f}_h^2(x)} \tag{A3}$$

is a kernel weight sequence with $\hat{f}'_h(x) \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^n K'_h(x - X_i)$, h is the bandwidth, K is a kernel function; $\int K(u)du = 1$, $K_h(\cdot) \stackrel{\text{def}}{=} h^{-1}K(\cdot/h)$. The function $\rho(\cdot)$ is designed to provide more robustness than the quadratic loss. An example of such a function is given by Huber (1964), see also Härdle (1989):

$$\rho(u) = \begin{cases} 0.5u^2, & \text{if } |u| \leq c; \\ c|u| - 0.5c^2 & \text{if } |u| > c. \end{cases} \tag{A4}$$

with the constant c regulating the degree of resistance.

A.2. Derivation of the conditional density of Heston integrated variance given terminal log-price

As pointed out in (15), we require the conditional density $f_{\tilde{V} | X_T}(\tilde{v} | x_T)$, under the measure \mathbb{Q} , to analytically determine the expression (11). As noted by Broadie and Kaya (2006) and Glasserman and Kim (2011), the conditional distribution of X_T is conditionally normal given \tilde{V} and $\int_0^T \sqrt{V_t} dW_{V,t}^{\mathbb{Q}}$:

$$\begin{aligned} X_T &\sim \mathbb{N} \left((r - c - 0.5)T - 0.5\tilde{V} + \rho \int_0^T \sqrt{V_t} dW_{V,t}^{\mathbb{Q}}, \sqrt{1 - \rho^2} \tilde{V} \right), \\ &\sim \mathbb{N} \left((r - c - 0.5)T - 0.5\tilde{V} + \frac{\rho}{\sigma} (V_T - V_0 - \kappa\theta T + \kappa\tilde{V}), \right. \\ &\quad \left. \sqrt{1 - \rho^2} \tilde{V} \right), \end{aligned} \tag{A5}$$

where (A5) follows from

$$\int_0^T \sqrt{V_t} dW_{V,t}^{\mathbb{Q}} = \sigma^{-1} (V_T - V_0 - \kappa\theta T + \kappa\tilde{V}),$$

which, in turn, follows from (14).

Expression (A5) yields the conditional density $f_{X_T | \tilde{V}, V_T}$ given Heston parameters $\kappa, \theta, V_0, \sigma$ and ρ . The joint density $f_{\tilde{V}, X_T, V_T}$ can then be obtained by simply multiplying $f_{X_T | \tilde{V}, V_T}$ by $f_{\tilde{V}, V_T}$, the joint density of \tilde{V} and V_T given some starting Heston variance level V_0 .

Broadie and Kaya (2006) have derived the characteristic function of the distribution of \tilde{V} given variance endpoint values V_T and V_0 . This function is quite complex and involves modified Bessel functions of a complex variable of the first kind:

$$\begin{aligned} \varphi_{\tilde{V} | V_T}(\omega) &= \frac{\gamma(\omega) e^{-0.5(\gamma(\omega) - \kappa)T} (1 - e^{-\kappa T})}{\kappa (1 - e^{-\gamma(\omega)T})} \exp \\ &\quad \left\{ \frac{V_T + V_0}{\sigma^2} \frac{\kappa (1 + e^{-\kappa T})}{1 - e^{-\kappa T}} - \frac{\gamma(\omega) (1 + e^{-\gamma(\omega)T})}{1 - e^{-\gamma(\omega)T}} \right\} \\ &\quad \times \frac{B_{2\kappa\theta\sigma^{-2}-1} \left(\sqrt{V_T V_0} \frac{4\gamma(\omega) e^{-0.5\gamma(\omega)T}}{\sigma^2 (1 - e^{-\gamma(\omega)T})} \right)}{B_{2\kappa\theta\sigma^{-2}-1} \left(\sqrt{V_T V_0} \frac{4\kappa e^{-0.5\kappa T}}{\sigma^2 (1 - e^{-\kappa T})} \right)}, \end{aligned} \tag{A6}$$

where $\gamma(\omega) = \sqrt{\kappa^2 - 2\sigma^2 i\omega}$, $i = \sqrt{-1}$ and $B_\nu(z)$ is the modified Bessel function of the first kind given by

$$B_\nu(z) \stackrel{\text{def}}{=} (z/2)^\nu \sum_{j=0}^{\infty} \frac{(z^2/4)^j}{j! \Gamma(\nu + j + 1)},$$

where $\Gamma(x) \stackrel{\text{def}}{=} \int_0^\infty t^{x-1} e^{-t} dt$ is the gamma function.

Using the inversion formula for characteristic functions, we can compute the density $f_{\tilde{V}|V_T, V_0}$ as follows:

$$f_{\tilde{V}|V_T}(\tilde{v}|v_T) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-i\tilde{v}\omega} \varphi_{\tilde{V}|V_T}(\omega) d\omega.$$

To find the joint density $f_{\tilde{V}, X_T, V_T}$, the transitional density $f_{V_T|V_0}$ is required. As noted by Cox *et al.* (1985) in the context of short interest rate process, V_T given V_0 follows a scaled non-central chi-squared distribution:

$$V_T = \frac{\sigma^2(1 - e^{-\kappa T})}{4\kappa} \tilde{\chi}_d^2 \left(\frac{4\kappa e^{-\kappa T} V_0}{\sigma^2(1 - e^{-\kappa T})} \right), \quad (A7)$$

where $\tilde{\chi}_d^2(\lambda)$ stands for the non-central chi-squared random variable with d degrees of freedom and non-centrality parameter λ . The probability density function of $\tilde{\chi}_d^2(\lambda)$ is defined using $B_\nu(z)$:

$$f_{\tilde{\chi}_d^2(\lambda)}(x) = 0.5 e^{-0.5(x+\lambda)} (x\lambda^{-1})^{0.25d-0.5} B_{0.5d-1}(\sqrt{\lambda x}).$$

Using a change-of-variables technique, it is straightforward to show that the density $f_{V_T|V_0}$ takes the form:

$$\begin{aligned} f_{V_T|V_0}(v_T|V_0) &= \frac{2\kappa}{\sigma^2(1 - e^{-\kappa T})} \exp \left\{ \frac{\kappa^2 \theta T}{\sigma^2} - 0.5\kappa T - \frac{2\kappa(v_T + e^{-\kappa T} V_0)}{\sigma^2(1 - e^{-\kappa T})} \right\} \\ &\times \left(\frac{v_T}{V_0} \right)^{\kappa\theta\sigma^{-2}-0.5} B_{2\kappa\theta\sigma^{-2}-1} \left(\frac{4\kappa e^{-0.5\kappa T}}{\sigma^2(1 - e^{-\kappa T})} \sqrt{V_0 v_T} \right). \end{aligned}$$

Using known rules for computing joint densities via conditional and marginal densities, it follows that

$$\begin{aligned} f_{\tilde{V}, X_T, V_T}(\tilde{v}, x_T, v_T) &= f_{X_T|\tilde{V}, V_T}(x_T|\tilde{v}, v_T) f_{\tilde{V}|V_T}(\tilde{v}|v_T) f_{V_T|V_0}(v_T|V_0), \\ f_{\tilde{V}, X_T}(\tilde{v}, x_T) &= \int_0^\infty f_{\tilde{V}, X_T, V_T}(\tilde{v}, x_T, v_T) dv_T. \end{aligned}$$

Therefore we have for $f_{\tilde{V}, X_T}$ estimated at \tilde{v}, x_T , given the Heston parameters:

$$\begin{aligned} f_{\tilde{V}, X_T}(\tilde{v}, x_T) &= \frac{2\kappa}{(2\pi)^{3/2} \sqrt{(1 - \rho^2)\tilde{v}\sigma^2(1 - e^{-\kappa T})}} \int_0^\infty \exp \\ &\left\{ \frac{\kappa^2 \theta T}{\sigma^2} - 0.5\kappa T - \frac{2\kappa(v_T + e^{-\kappa T} V_0)}{\sigma^2(1 - e^{-\kappa T})} \right\} \end{aligned}$$

$$\begin{aligned} &+ \frac{(v_T + V_0)\kappa(1 + e^{-\kappa T})}{\sigma^2(1 - e^{-\kappa T})} \\ &- (x_T - \log(S_0) - (r - c - 0.5)T \\ &+ 0.5\tilde{v} - \frac{\rho}{\sigma}(v_T - V_0 - \kappa\theta T + \kappa\tilde{v}))^2 \Bigg\} \\ &\times \left(\frac{v_T}{V_0} \right)^{\kappa\theta\sigma^{-2}-0.5} \int_{-\infty}^{+\infty} e^{-i\tilde{v}\omega} \\ &\times \left[\frac{\gamma(\omega) e^{-0.5(\gamma(\omega)-\kappa)T} (1 - e^{-\kappa T})}{\kappa(1 - e^{-\gamma(\omega)T})} \exp \right. \\ &\left. \left\{ - \frac{\gamma(\omega)(1 + e^{-\gamma(\omega)T})(v_T + V_0)}{\sigma^2(1 - e^{-\gamma(\omega)T})} \right\} \right. \\ &\left. \times B_{2\kappa\theta\sigma^{-2}-1} \left(\sqrt{v_T V_0} \frac{4\gamma(\omega) e^{-0.5\gamma(\omega)T}}{\sigma^2(1 - e^{-\gamma(\omega)T})} \right) \right] d\omega dv_T. \end{aligned}$$

Finally, the density $f_{\tilde{V}|X_T}(\tilde{v}|x_T)$ is found as

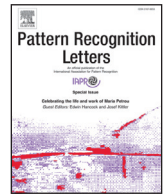
$$f_{\tilde{V}|X_T}(\tilde{v}|x_T) = \frac{f_{\tilde{V}, X_T}(\tilde{v}, x_T)}{f_{X_T}(x_T)},$$

where $f_{X_T}(x_T)$ is the probability density of X_T estimated at x_T . This marginal density under the risk-neutral measure \mathbf{Q} is again found via inversion of the characteristic function, see Rouah (2013):

$$\varphi(\omega)_{X_T} = \exp \{C(\omega) + D(\omega)V_0 + i\omega \log(S_0)\},$$

where

$$\begin{aligned} C(\omega) &= r i \omega T + \frac{\kappa \theta}{\sigma^2} \{(\kappa - \rho \sigma i \omega + d(\omega))T \\ &- 2 \log \left(\frac{1 - g(\omega) e^{-d(\omega)T}}{1 - g(\omega)} \right)\}, \\ D(\omega) &= \frac{(\kappa - \rho \sigma i \omega + d(\omega))}{\sigma^2} \left(\frac{1 - e^{-d(\omega)T}}{1 - g(\omega) e^{-d(\omega)T}} \right), \\ g(\omega) &= \frac{\kappa - \rho \sigma i \omega + d(\omega)}{\kappa - \rho \sigma i \omega - d(\omega)}, \\ d(\omega) &= \sqrt{(\rho \sigma i \omega - \kappa)^2 + \sigma^2(i\omega - \omega^2)}. \end{aligned}$$



Towards the interpretation of time-varying regularization parameters in streaming penalized regression models



Lenka Zboňáková^a, Ricardo Pio Monti^{b,c,*}, Wolfgang Karl Härdle^{a,d,e}

^aC.A.S.E. - Center for Applied Statistics & Economics, Humboldt-Universität zu Berlin, Spandauer Str. 1, Berlin 10178, Germany

^bGatsby Computational Neuroscience Unit, UCL, 25 Howland Street, London W1T 4JG, UK

^cAdvanced Telecommunications Research Institute International (ATRI), Kyoto, Japan

^dSim Kee Boon Institute for Financial Economics, Singapore Management University, 50 Stamford Road 178899, Singapore

^eThe Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen 361005, China

ARTICLE INFO

Article history:

Received 29 October 2018

Available online 20 June 2019

Keywords:

Penalized regression

Non-stationarity

Regularization

Streaming data

ABSTRACT

High-dimensional, streaming datasets are ubiquitous in modern applications. Examples range from finance and e-commerce to the study of biomedical and neuroimaging data. As a result, many novel algorithms have been proposed to address challenges posed by such datasets. In this work, we focus on the use of ℓ_1 regularized linear models in the context of (possibly non-stationary) streaming data. Recently, it has been noted that the choice of the regularization parameter is fundamental in such models and several methods have been proposed which iteratively tune such a parameter in a time-varying manner; thereby allowing the underlying sparsity of estimated models to vary. Moreover, in many applications, inference on the regularization parameter may itself be of interest, as such a parameter is related to the underlying sparsity of the model. However, in this work, we highlight and provide extensive empirical evidence regarding how various (often unrelated) statistical properties in the data can lead to changes in the regularization parameter. In particular, through various synthetic experiments, we demonstrate that changes in the regularization parameter may be driven by changes in the true underlying sparsity, signal-to-noise ratio or even model misspecification. The purpose of this letter is, therefore, to highlight and catalog various statistical properties which induce changes in the associated regularization parameter. We conclude by presenting two applications: one relating to financial data and another to neuroimaging data, where the aforementioned discussion is relevant.

© 2019 Elsevier B.V. All rights reserved.

1. Introduction

High-dimensional, streaming datasets pose a unique challenge to modern statisticians. To date, the challenges associated with high-dimensional and streaming data have been extensively studied independently. In the case of the former, a popular avenue of research is the use of regularization methods such as the Lasso [9]. Such methods effectively address issues raised by high-dimensional data by assuming the underlying model is sparse, thereby having only a small number of non-zero coefficients. Sparse models are often easier to both estimate and interpret. Concurrently, many methods have been developed to handle streaming datasets; popular examples include sliding window methods and their generalizations to weighted moving averages [10].

Recently, the intersection of these two avenues of research has begun to receive increasing attention as large-scale, streaming datasets become commonplace. Prominent examples include [5] and [8] who propose methods through which to efficiently estimate ℓ_1 penalized models in a streaming data context. However, an important aspect, which has been largely overlooked, corresponds to the optimal choice of the regularization parameter. While it is possible to employ a fixed regularization parameter, it may be the case that the statistical properties of the data vary over time, suggesting that the optimal choice of the regularization parameter may itself also vary over time. Throughout this work, we consider the optimal choice of regularization parameter both in terms of negative log-likelihood on unseen data as well as in terms of correctly recovering the true underlying model (i.e., sparse support recovery). We choose to consider two distinct criteria for the choice of the regularization parameter in order to systematically study how changes in the statistical distribution of observations affect the choice of regularization parameter. Examples of large-scale, non-stationary datasets, where the choice of the regularization

* Corresponding author.

E-mail address: r.monti@ucl.ac.uk (R.P. Monti).

parameter has been reported to be time-varying, include finance [22] and neuroscience [16].

We note that many methods have been proposed for selecting the regularization parameter in the context of non-streaming data, the standard approach being to employ some variant of cross-validation or bootstrapping, e.g. in Hastie et al. [9] or Chernozhukov et al. [7]. However, such methods are infeasible in the domain of streaming datasets due to limited computational resources. More importantly, the statistical properties of a data stream may vary, further complicating the use of sub-sampling methods. Recently, methods to handle time-varying regularization parameters have been proposed. Monti et al. [14] propose a novel framework through which to iteratively infer a time-varying regularization parameter via the use of adaptive filtering. The proposed framework is developed for penalized linear regression (i.e., the Lasso) and subsequently extended to penalized generalized linear models. Zbonakova et al. [23] study the dynamics of the regularization parameter, focusing particularly on quantile regression in the context of financial data. Using the sliding windows method, they demonstrate that the choice of time-varying regularization parameter based on the adjusted Bayesian information criterion (BIC) is closely correlated with the financial volatility. The BIC criterion was employed, as such a choice of the parameter has been proved to consistently select a true model [21].

While the aforementioned methods correspond to valuable contributions, the purpose of this paper is to highlight potential shortcomings when interpreting time-varying regularization parameters. In particular, through a series of simulation studies, we enumerate several (often unrelated) statistical properties of the underlying data which may lead to changes in the optimal choice of the regularization parameter. This paper, therefore, serves to highlight important issues associated with the interpretation of time-varying regularization parameters as well as the associated model parameters. Given the increased importance of both regularization-based methods as well as streaming algorithms within modern applications of statistics, this work serves as an important precursor.

The remainder of this paper is organized as follows. We formally outline the challenge of tuning time-varying regularization parameters as well as related work in Section 2. In Section 3, we present extensive empirical results, highlighting how various aspects of the underlying data may result in changes in the estimated regularization parameter. Computations included in this work were performed with the help of the R software environment [19] and we provide code to reproduce all experiments on the Quantlet platform.¹

2. Preliminaries and related work

In this work, we focus on streaming linear regression problems. Formally, it is assumed that we observe a sequence of pairs (X_t, y_t) , where $X_t \in \mathbb{R}^p$ corresponds to a p -dimensional vector of predictor variables and $y_t \in \mathbb{R}$ is a univariate response. We write subscript $t \in \mathbb{N}$ to denote the time index of observations. The objective of penalized streaming linear regression problems consists in accurately predicting future responses, y_{t+1} , from predictors X_{t+1} via a linear model. Following the work of [20], an ℓ_1 penalty, parameterized by $\lambda \in \mathbb{R}_+$, is subsequently introduced in order to encourage sparse solutions as well as ensure the associated optimization problem is well-posed. For a pre-specified choice of a fixed regularization parameter, λ , time-varying regression coefficients can be estimated by minimizing the following convex objective:

$$L_t(\beta, \lambda) = \sum_{i=1}^t w_i (y_i - X_i^\top \beta)^2 + \lambda \|\beta\|_1, \quad (1)$$

where $w_i > 0$ are weights indicating the importance given to past observations [1]. For example, it is natural to allow w_i to decay monotonically in a manner which is proportional to the chronological proximity of the i th observation. While the weights w_i may be tuned using a fixed forgetting factor, throughout this work we opt for the use of a sliding window due to the simplicity of the latter method. We note that in the case of sliding windows the weights, w_i , are either set to 0 or 1 depending on whether the corresponding observation falls within the specified window.

As noted in Section 1, in the context of non-stationary data the optimal estimates of regression coefficients, $\hat{\beta}_t$, may vary over time and several methods have been proposed in order to address this issue [5,8]. In this work, we note that the same argument may be posed in terms of the associated regularization parameter, λ . The choice of such a parameter dictates the severity of the associated ℓ_1 penalty, implying that different choices of λ will result in vastly different estimated models. However, the issue of iteratively tuning the regularization parameter has, until recently, been largely overlooked.

Formally, [17] clearly outline the relationship between the Lasso parameter, λ , and the data. They note that the regularization parameter may be interpreted as the Lagrange multiplier associated with a constraint on the ℓ_1 norm of the regression coefficients. As such, considering the dual formulation yields:

$$\lambda = \frac{\{\mathbf{Y} - \mathbf{X}\hat{\beta}(\lambda)\}^\top \mathbf{X}\hat{\beta}(\lambda)}{\|\hat{\beta}(\lambda)\|_1}, \quad (2)$$

where we have omitted the weights, w_i , for simplicity and use bold notation to denote vectors and matrices respectively. Note that in Eq. (2) we clearly denote the dependence of the estimated regression coefficients on λ . As a result, we observe three main effects driving the optimal choice of the regularization parameter:

1. Variance or magnitude of the residuals: $\mathbf{Y} - \mathbf{X}\hat{\beta}(\lambda)$. As the variance of residuals increases (decreases) so does the associated regularization parameter, leading to an increase (decrease) in sparsity of $\hat{\beta}(\lambda)$. This is natural as an increase in the variance of residuals is indicative of a drop in the signal-to-noise ratio of the data.
2. The ℓ_1 norm of the model coefficients: $\|\hat{\beta}(\lambda)\|_1$. As this term appears in the denominator of Eq. (2), it is inversely correlated with the regularization parameter. This is to be expected as we require a small regularization parameter in order to accurately recover regression coefficients with large ℓ_1 norm. We further note that the ℓ_1 norm will often be closely associated with the ℓ_0 norm, indeed the ℓ_1 is often employed as the convex approximation to the ℓ_0 norm. As such, we further expect there to be an inverse correlation between the ℓ_0 norm and the regularization parameter.
3. Covariance structure of the design matrix: \mathbf{X} . The term related to the covariance structure of the design matrix, $\mathbf{X}^\top \mathbf{X}$, can be extracted from the elements in the numerator of Eq. (2). This suggests that the covariance matrix of the predictors will have a significant impact on the value of the regularization parameter, λ . We note that this effect will also affect the ℓ_1 and ℓ_0 norms of the model coefficients, resulting in a complicated relationship with the regularization parameter. In Section 3.1.3 we demonstrate the non-linear nature of this relationship.

As such, it follows that multiple aspects of the data may influence the choice of the associated regularization parameter. Crucially, whilst such a parameter is often interpreted as being indicative of the *sparsity* of the underlying model, Eq. (2) together with the aforementioned discussion demonstrates that this is not necessarily the case. In the remainder of this work, we provide extensive empirical evidence to validate these claims.

¹ <http://quantlet.de/>.

3. Experimental results

In this section, we provide an extensive simulation study to demonstrate the effects of the three aforementioned model properties on the choice of the optimal regularization parameter. Based on the observations from Section 2, we designed a series of experiments where one property of the data was allowed to vary whilst the remaining two were left unchanged. A further concern is to show that if two or more of the properties of the data should simultaneously change it can result in cancelling out their effects on the regularization parameter. Further experiments were designed to study those scenarios. The purpose of the experimental results presented in this section is two-fold. First, we identify the various statistical properties which cause the optimal choice of regularization parameter to vary. Second, we also highlight how changes of such properties interact with each other and catalog their joint effects on the choice of the regularization parameter.

3.1. Synthetic data generation

We focus exclusively on a linear model of the form:

$$y_t = X_t \beta_t + \varepsilon_t.$$

We define the number of observations as n , the number of non-zero parameters as $q = \|\beta\|_0 \leq p$ and an iid error term $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top$, such that $\varepsilon_t \sim N(0, \sigma_t^2)$. The p -dimensional vector of predictor variables X_t was generated from the normal distribution $N_p(0, \Sigma)$, where the elements of the $(p \times p)$ covariance matrix $\Sigma = (\sigma_{ij})_{i,j=1}^p$ were set to be $\sigma_{ij} = \rho^{|i-j|}$, for $i, j = 1, \dots, p$, with a correlation parameter ρ . The choice of such a model for the covariance structure of predictor variables, X_t , was motivated by three considerations. First, it is a widely studied class of covariance model [4]. Second, it yields a family of covariance matrices which are indexed by a single parameter, ρ . This is important in the context of visualizing and interpreting the results of our experiments. Finally, such a structure is often encountered in real-data applications. For example, in the context of functional MRI data, which we consider in Section 3.2, such covariance structure is often reported due to the spatial auto-correlation present in the data [18]. The same is true in financial applications where one can use this matrix form to approximate an empirical correlation matrix of asset returns [2].

We generate synthetic data where one of the following properties varies over time (thereby resulting in non-stationarity):

1. Time-varying variance of residuals: σ_t^2 varies over time.
2. Time-varying ℓ_1 or ℓ_0 norm of regression coefficients: q varies over time.
3. Time-varying correlation within design matrix: ρ varies over time.

For each experiment, the total number of observations was set to $n = 400$ with a dimensionality of $p = 20$. In order to study the effects of the aforementioned properties on the regularization parameter, we consider the use of three distinct methods. First, we consider the use of the sliding windows of length of 50 observations in combination with both the Bayesian information criterion (BIC) and the generalized cross-validation (GCV) to select the associated regularization parameter. Finally, the gradient method proposed by [14], named Real-time Adaptive Penalization (RAP), is also applied. The use of BIC and GCV based methods is motivated by the fact that such methods are frequently employed to select regularization parameters [9], while the RAP method is a recently proposed streaming algorithm motivated by adaptive filtering [10]. The fundamental difference between the sliding window methods and RAP is that the RAP algorithm directly computes the derivative of the negative log-likelihood (in the case of the Lasso this is

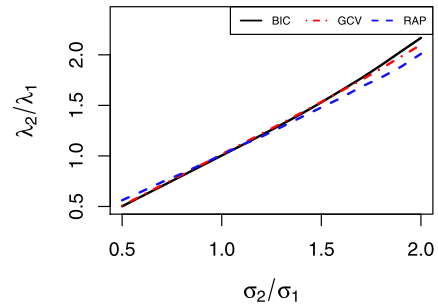


Fig. 1. Relative changes of λ in as a function of the on relative changes of the standard deviation σ .

the mean squared error) of unseen data with respect to λ . This gradient information is then used to tune the regularization parameter in a streaming fashion. In contrast, sliding window based approaches perform a grid search at each iteration, optimizing either the BIC of GCV scores.

For each of the three methods, a burn-in period of 50 observations was employed to obtain an initial estimate for the regression coefficients as well as λ . Each experiment was repeated 100 times and the mean value of the regularization parameter was studied.

3.1.1. Change of the variance of residuals

We begin by studying the effect of residual variance on the choice of the regularization parameter, λ . The regression coefficients were set to $\beta_t = (1, 1, 1, 1, 1, 0, \dots, 0)^\top$, yielding $q = 5$ and the covariance parameter was set to be $\rho = 0.5$. The vector of residuals was simulated according to a piece-wise stationary distribution as follows:

$$\varepsilon_t \sim \begin{cases} N(0, \sigma_1^2), & \text{for } t < 200; \\ N(0, \sigma_2^2), & \text{for } t \geq 200, \end{cases} \quad (3)$$

resulting in a change in the variance of residuals at the 200th observation. Throughout these experiments, we set $\sigma_1 = 1$ and allowed σ_2 to vary from $\sigma_2 \in \{0.5, \dots, 1.5\}$. In order to study the effects of changes in the variance of residuals, we consider the change in the estimated regularization parameter defined by the ratio of the values of λ after (λ_2) and before (λ_1) the change point as a function of the ratio σ_2/σ_1 . Following the discussion from Section 2, we would expect increasing values of the ratio to result in increasing values of the regularization parameter and vice versa.

Fig. 1 plots the effect of the changes in the standard deviation of residuals on the Lasso parameter λ . As expected when looking at the formula (2), there is a linear dependency visible. In the case of all the methods employed (BIC, GCV, RAP), the lines are almost identical and the dependence of $\frac{\lambda_2}{\lambda_1}$ on $\frac{\sigma_2}{\sigma_1}$ can clearly be seen.

In order to illustrate how the series of values of the Lasso parameter changes over time and how long it takes to adjust for the new settings of the model, we depict the average λ over the 100 scenarios in Fig. 2, where $\sigma_1 = 1$ and $\sigma_2 = 1.5$. We normalize the BIC, GCV and RAP values of λ to fit into the interval $[0, 1]$.

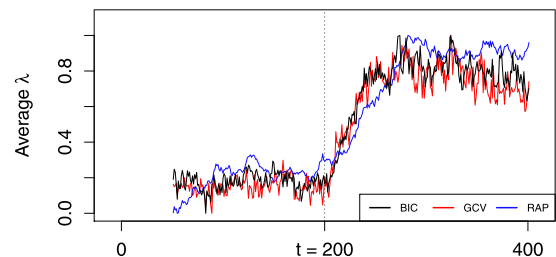


Fig. 2. Standardized series of average λ over 100 scenarios with a change point at $t = 200$ and $\sigma_1 = 1$ and $\sigma_2 = 1.5$.

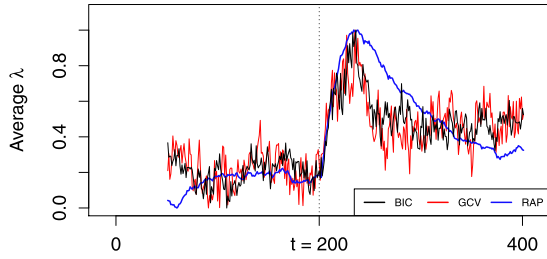


Fig. 3. Standardized series of average λ over 100 scenarios with a change point at $t = 200$ and regression coefficients β_t defined by (4).

From Fig. 2 it is clear that the values of λ adjust for the new model settings for the whole length of the moving window (50 in this case) if this approach is implemented. We note there is a clear change in the regularization parameter following $t = 200$, indicating the need to adaptively estimate the regularization parameter and demonstrating the drawback of using a fixed and pre-specified value of λ .

3.1.2. Change of the ℓ_1 and ℓ_0 norm of β

It follows that the choice of the regularization parameter is closely related to the true underlying ℓ_1 and ℓ_0 norm of the regression coefficients; the relation to the latter is because the Lasso constraint is introduced as a convex relaxation of the ℓ_0 norm.

In this set of simulations, we therefore quantify the effects of changes in both the ℓ_0 and ℓ_1 norms on the optimal choice of the regularization parameter. In particular, we set $\sigma_1 = \sigma_2 = 1$ and $\rho = 0.5$. As a first example, we consider the following changes in the ℓ_1 norm:

$$\beta_t = \begin{cases} (1, 1, 1, 1, 1, 0, \dots, 0)^\top, & \text{for } t < 200; \\ (1, 0.8, 0.6, 0.4, 0.2, 0, \dots, 0)^\top, & \text{for } t \geq 200. \end{cases} \quad (4)$$

We note that the changes induced by Eq. (4) correspond to a drop in the signal-to-noise ratio at time $t = 200$. This is a phenomenon that is often encountered in neuroimaging data analysis and is frequently attributed to the subjects losing interest in the task at hand [11,18]. The time series of estimated λ values is presented in Fig. 3.

We note that the change in the ℓ_1 norm of the model coefficients β results in an upward trend in λ for the BIC and GCV parameter choice which is visible in the long run. For a short period after the change, exactly the period of 50 observations from the moving window, the misspecification of the model drives the size of residuals and with them, the values of λ higher and lower again in a “bump”-shaped line. The same holds for the RAP algorithm, however, because of the fixed forgetting factor, the values of λ are adjusting to the new model settings more slowly.

In order to study the effect of changes in the ℓ_0 norm (i.e., the size of the active set) we generated synthetic data whereby:

$$\|\beta_t\|_0 = \begin{cases} q_1, & \text{for } t < 200; \\ q_2, & \text{for } t \geq 200, \end{cases} \quad (5)$$

with $q_1 = 10$ and $q_2 \in \{5, \dots, 15\}$.

Fig. 4 visualizes the relative changes of λ as a function of the relative changes in the size of the active set, defined as q_2/q_1 . We note there is a clear decay of values of λ_2/λ_1 as q_2/q_1 increases. This is to be expected, as an increase in the latter ratio implies a larger active set in the second part of the time series and thus causes a decrease in λ_2 . Hence, this figure provides empirical validation of the inverse correlation between the magnitude of the active set and the estimated regularization parameter.

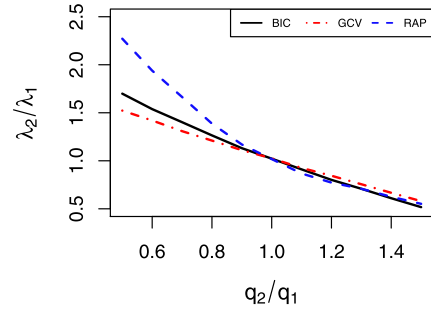


Fig. 4. Relative changes of λ in dependence on relative changes of the size of the active set q .

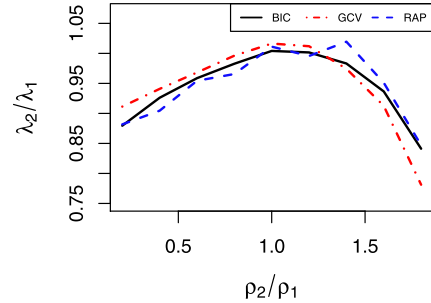


Fig. 5. Relative changes of λ in dependence on relative changes of the correlation parameter ρ .

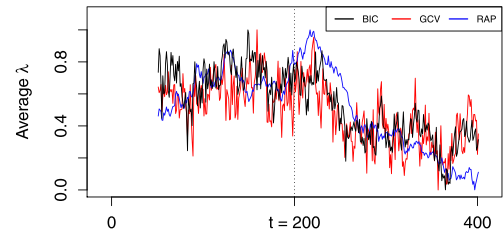


Fig. 6. Standardized series of average λ over 100 scenarios with a change point at $t = 200$ and $\rho_1 = 0.5$ and $\rho_2 = 0.1$.

3.1.3. Change of covariance parameter ρ

Finally, we study the effect of changes in the covariance structure of features, X_t , on the regularization parameter. We note that whilst it is possible to vary the covariance structure in many ways, we consider a simple model whereby $\Sigma = (\sigma_{ij})_{i,j=1}^p$ and set $\sigma_{ij} = \rho^{|i-j|}$. The benefit of such a model is that it only depends on a single parameter, ρ , simplifying the interpretation and visualization of results. As such, we investigate changes in the covariance parameter ρ , while fixing $\sigma = 1$ and $q = 5$. Formally, piece-wise stationary data was generated such that:

$$\rho_t = \begin{cases} \rho_1, & \text{for } t < 200; \\ \rho_2, & \text{for } t \geq 200, \end{cases} \quad (6)$$

where $\rho_1 = 0.5$ and $\rho_2 \in \{0.1, \dots, 0.9\}$.

As in the previous experiments, we visualize the relative changes of λ with respect to the relative changes of ρ in Fig. 5. The time series of the estimated values of λ over the whole sample size are depicted in Fig. 6.

From Fig. 5 it is important to note that the changes of λ no longer demonstrate a linear dependency with the statistical property of interest. The values of λ tend to decay with both decreasing and rising covariance of the predictors. A potential explanation for the non-linear nature of the relationship demonstrated in Fig. 5 is due to the selection properties of the Lasso. It is widely acknowledged that in the presence of strongly correlated variables

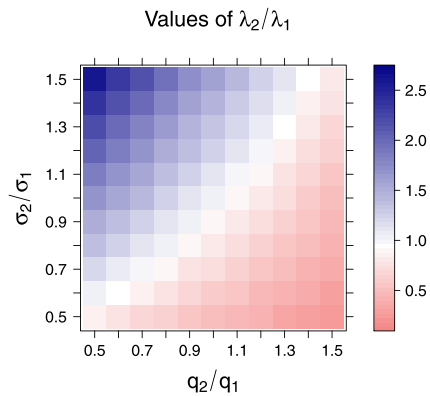


Fig. 7. Relative changes of λ corresponding to the combination of relative changes of q and σ .

(corresponding to large ρ values) the Lasso tends to choose only a single variable from the group of strongly correlated covariates (indeed this phenomenon is the inspiration for the elastic net [24]). As such, as ρ increases, the term $\mathbf{X}^T \mathbf{X}$ from the numerator of λ drives its values higher. If the ρ value is too high, we speak of multicollinearity, where the denominator of λ is affected and becomes larger, which consequently causes the λ values to drop.

In Fig. 6 the change from $\rho_1 = 0.5$ to $\rho_2 = 0.1$ is depicted. We note there is a change in λ despite the fact that the ℓ_1 and ℓ_0 norms remain unchanged.

3.1.4. Simultaneous changes of model specifications

While the previous experiments have examined the effects of changing a single property of the data, we now consider combinations of specific changes. In particular, the purpose of the remaining experiments is to highlight how simultaneous changes to two properties of the data result in a *cancellation* of their effects on the regularization parameter. The purpose of this section is, therefore, to highlight the fact that it is possible to have non-stationary data where the three properties discussed in Section 2 vary, and yet the optimal choice of the sparsity parameter is itself constant.

We begin by studying simultaneous changes in the ℓ_0 or ℓ_1 norm of the regression parameters, β_t , together with changes in the variance of residuals, σ^2 . Recall that the optimal choice of regularization parameter was positively correlated with the magnitude of residuals (see Fig. 1) whilst being negative correlated with q (see Fig. 4). Fig. 7 shows the relative change of λ as a function of both q_2/q_1 and σ_2/σ_1 . It is important to note the diagonal trend, which indicates that for any increase in q , a proportional increase in σ directly cancels out the change in the estimated regularization parameter. This is a natural result as the changes in σ influence the numerator, whilst changes in the ℓ_0 or ℓ_1 norm affect the denominator in (2).

Next, we consider the combination of varying the covariance parameter, ρ , and the variance of the residuals, parameterized by σ . Recall from the previous discussion that the covariance parameter, ρ , did not have a linear relationship with the regularization parameter, λ . Such a non-linear relationship can be clearly seen again in Fig. 8. Furthermore, we note that changes in σ tend to dominate the changes in ρ , as is illustrated by the rising values of λ in the upper part of the plot.

Finally, we also studied the combination of changes in the ℓ_0 norm, denoted by q , together with changes in the covariance parameter, ρ . Note that changes in these parameters are strongly coupled due to the effect of multicollinearity induced by simultaneously increasing the number of non-zero regression coefficients together with their correlations. The results, provided in Fig. 9, highlight these dependencies. The left-hand side of the heatmap

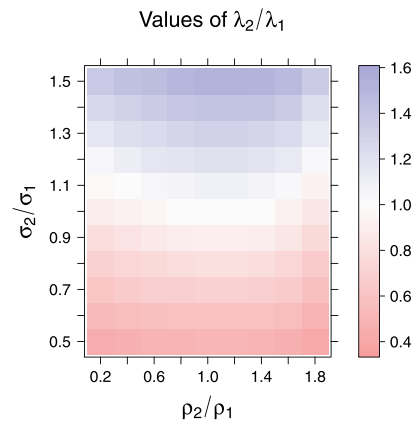


Fig. 8. Relative changes of λ corresponding to the combination of relative changes of ρ and σ .

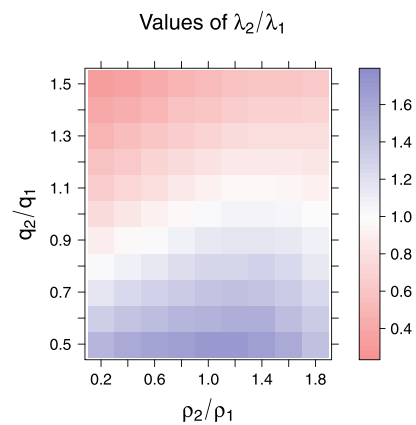


Fig. 9. Relative changes of λ corresponding to the combination of relative changes of q and ρ .

shows a diagonal pattern for changes in ρ and q where the number of non-zero coefficients tends to have a larger influence on the values of λ . However, as the values of ρ_2 get greater and the effect of multicollinearity becomes relevant, the diagonal pattern disappears and values of λ do not decrease as significantly as for the smaller ρ_2 's.

3.2. Application to financial and neuroimaging data

Until now we have provided empirical evidence based on a variety of simulations, each varying one or more of the statistical properties of the data. In this section, we conclude by presenting two distinct real-world datasets where we observe significant variability in the time-varying regularization parameter. The two datasets presented in this section serve as two examples from diverse applications where non-stationary data is encountered.

We study two high-dimensional real-world datasets from distinct applications: the first consists of stock returns and the second corresponds to functional MRI (fMRI) dataset taken from an emotion task. The stock return data consists of daily stock returns of 100 largest financial companies over a period of 11 years from 2007 to 2018. The companies listed on NASDAQ are ordered by the market capitalization and downloaded from Yahoo Finance. This data is particularly interesting as it covers the financial crisis of 2008–2009.

The second dataset we consider corresponds to fMRI data collected as a part of the Human Connectome Project (HCP). This dataset consists of measurements of 15 distinct brain regions taken

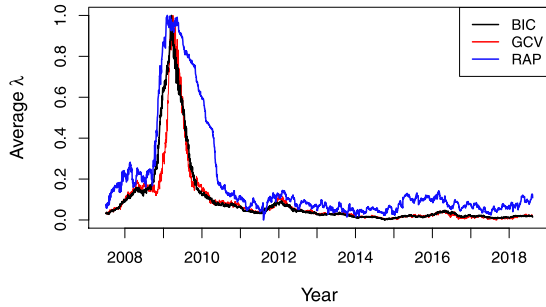


Fig. 10. Standardized series of average λ in the US stock returns data, daily observations from January 3, 2007 to August 10, 2018.

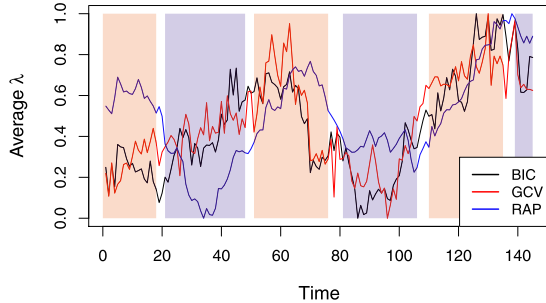


Fig. 11. Standardized series of average λ in the fMRI dataset. Distinct tasks are indicated by the background color (red indicates a neutral task, blue indicates an emotion task and white denotes the resting period). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

during an emotion task, as described in [3]. Data was analyzed over a subset of 50 subjects. While traditional neuroimaging studies were premised on the assumption of stationarity, an exciting avenue of neuroscientific research corresponds to understanding the non-stationary properties of the data and how these may potentially correspond to changes induced by distinct tasks [16,26] or changes across subjects [13]. We note that within the context of neuroimaging, understanding variability induced by distinct tasks or across a population of subjects is an important challenge [27–30].

The modelling procedure employed for both of the datasets consisted of regressing each of the components of the multivariate time series on the rest. Length of the sliding windows corresponded to 126 observations (approximately 6 months) in case of the stock returns data and 30 observations in case of the fMRI data. This way we got either 100 or 15 sequences of the Lasso parameter values, for financial and neuroimaging data respectively, which were then averaged and normalized to the $[0, 1]$ interval as before. The resulting time series for the US stock market data are depicted in Fig. 10 and for the fMRI data the graphical output can be seen in Fig. 11.

From Fig. 10 it is visible that the values of λ react to the situation on the market in both of the algorithms, the standard one with the BIC and GCV as selecting rules and the RAP. Especially pronounced is the change of the values during the financial crisis of 2008–2009 where the volatility observable on the market was elevated and thus results in increased values of the Lasso parameter, too. Interestingly, both of the considered methods react instantly if some change occurs, but take a different amount of observations to adjust back to the standard situation.

Fig. 11 shows the time series of the average regularization parameter over eight distinct subjects performing an emotion related task. The task required participants to perform a series of trails presented in blocks. The trails either required them to de-

cide which of the two faces presented on the bottom of the screen match the face at the top of the screen, or which of the two shapes presented at the bottom of the screen match the shape at the top of the screen. The former was considered the emotion task (and denoted in blue in Fig. 11) and the latter the neutral task (denoted in red in Fig. 11). From Fig. 11 we see clear changes in the estimated regularization parameter induced by changes in the underlying cognitive task, and thus, changes in the connectedness of the brain regions. This finding is in line with the current trend in the study of the fMRI data, which is interested in quantifying and understanding the non-stationarity properties of such datasets and how these relate to changes in cognitive state [6,15,31].

4. Discussion

In this work, we have highlighted and provided empirical evidence for the various statistical properties which affect the optimal choice of a regularization parameter in a penalized linear regression model. Based on the theory of the Lasso, we specifically consider three distinct properties: the variance of residuals, the ℓ_0 and ℓ_1 norms of the regression coefficients and the covariance structure of the design matrix. Throughout a series of experiments, we confirm the manner in which each of these properties affects the optimal choice of the regularization parameter. We relate the dependencies between each of the aforementioned statistical properties and estimated regularization parameter to the theoretical properties presented in [17]. In particular, we conclude that:

- There is a (positive) linear relationship between changes in the variance of residuals, σ^2 , and the estimated regularization parameter, as clearly demonstrated in Fig. 1.
- There is a (negative) linear relationship between changes in the size of the active set (either ℓ_0 or ℓ_1 norm) and the estimated regularization parameter, as shown in Fig. 4.
- There is a non-linear relationship between changes in the correlation structure in the design matrix and the estimated regularization parameter, as visualized in Fig. 5.

We further provide a series of experiments where two of the statistical properties jointly varied in order to demonstrate the possibility of having non-stationary time-series data where the optimal regularization parameter does not alter. This is most clearly seen in the case of changes in the active set, q , together with changes in the residual variance, σ^2 , shown in Fig. 7. It should be pointed out that throughout the various simulation studies we have focused primarily on ℓ_1 penalized linear regression models. As such, our conclusions cannot be directly translated to the domain of alternative regularization schemes (e.g., ℓ_2 regularization as in ridge regression) or alternative regression methods (e.g., generalized linear models). However, we believe the use of ℓ_1 penalized linear regression serves as a useful case study to highlight the many aforementioned issues associated with the interpretation of the regularization parameter. Moreover, due to the many similarities between the Lasso and more general penalized regression methods (many of which can often be directly written in terms of a Lasso problem e.g., [12,13,24,25]), we expect the aforementioned issues to be relevant beyond the Lasso.

It is also important to note that the choice of the regularization parameter will be affected by the choice of the window length employed. This follows intuitively from Eq. (1), where the different choices of the window length will give the observations different weightings. This can be seen by considering various extreme choices for the window length. If we set the window length to be as large as possible, effectively performing offline linear regression, the corresponding regularization parameter will be static. This, however, comes at the cost of potentially misspecifying the model and ignoring any non-stationarity present. Conversely, if the

window length is set to be too small, then relevant observations may be unnecessarily discarded, leading to increased variability in the estimated regularization parameter.

Finally, we conclude by presenting two case studies involving high-dimensional time-series data in the context of finance and neuroimaging. Both datasets demonstrate significant temporal variability in the estimated regularization parameter, thereby validating the need for the methods through which to iteratively tune such a parameter.

In conclusion, the purpose of this letter is to highlight and rigorously catalog the various statistical properties which may lead to changes in the choice of the regularization parameters in ℓ_1 penalized models. Such models are widely employed, indicating that an appreciation of the relationships between the various statistical properties of the data and the choice of the regularization parameter is important.

Declaration of Competing Interest

All authors confirm that there are no conflict of interest.

Acknowledgements


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.

References

- [1] C.C. Aggarwal, *Data Streams: Models and Algorithms*, Springer, 2007.
- [2] A.N. Akansu, M.U. Torun, Toeplitz approximation to empirical correlation matrix of asset returns: a signal processing perspective, *IEEE J. Sel. Top. Signal Process.* 6 (4) (2012) 319–326.
- [3] D.M. Barch, G.C. Burgess, M.P. Harms, S.E. Petersen, B.L. Schlaggar, M. Corbetta, M.F. Glasser, S. Curtiss, S. Dixit, C. Feldt, et al., Function in the human connectome: task-fMRI and individual differences in behavior, *NeuroImage* 80 (2013) 169–189.
- [4] P.J. Bickel, E. Levina, et al., Regularized estimation of large covariance matrices, *Ann. Stat.* 36 (1) (2008) 199–227.
- [5] L. Bottou, Large-scale machine learning with stochastic gradient descent, in: *Proc. COMPSTAT2010*, 2010, pp. 177–186.
- [6] V.D. Calhoun, R. Miller, G. Pearson, T. Adali, The chroconnectome: time-varying connectivity networks as the next frontier in fmri data discovery, *Neuron* 84 (2) (2014) 262–274.
- [7] V. Chernozhukov, W.K. Härdle, C. Huang, W. Wang, Lasso-driven inference in time and space, [arXiv:1806.05081](#) (2018).
- [8] J. Duchi, E. Hazan, Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization, *J. Mach. Learn. Res.* 12 (2011) 2121–2159.
- [9] T. Hastie, R. Tibshirani, M.W. Hastie, *Statistical Learning with Sparsity: The Lasso and Generalizations*, Chapman&Hall/CRC, 2015.
- [10] S. Haykin, *Adaptive Filter Theory*, Pearson Education, 2008.
- [11] R. Lorenz, I.R. Violante, R.P. Monti, G. Montana, A. Hampshire, R. Leech, Dissociating frontoparietal brain networks with neuroadaptive Bayesian optimization, *Nat. Commun.* 9 (1) (2018) 1227.
- [12] N. Meinshausen, P. Bühlmann, et al., High-dimensional graphs and variable selection with the lasso, *Ann. Stat.* 34 (3) (2006) 1436–1462.
- [13] R.P. Monti, C. Anagnostopoulos, G. Montana, Learning population and subject-specific brain connectivity networks via mixed neighborhood selection, *Ann. Appl. Stat.* 11 (4) (2017) 2142–2164.
- [14] R.P. Monti, C. Anagnostopoulos, G. Montana, Adaptive regularization for lasso models in the context of nonstationary data streams, *Stat. Anal. Data Min.* (2018), doi:10.1002/sam.11390.
- [15] R.P. Monti, P. Hellyer, D. Sharp, R. Leech, C. Anagnostopoulos, G. Montana, Estimating time-varying brain connectivity networks from functional MRI time series, *NeuroImage* 103 (2014) 427–443.
- [16] R.P. Monti, R. Lorenz, R.M. Braga, C. Anagnostopoulos, R. Leech, G. Montana, Real-Time estimation of dynamic functional connectivity networks, *Hum Brain Mapp* 38 (1) (2017) 202–220.
- [17] M.R. Osborne, B. Presnell, B.A. Turlach, On the lasso and its dual, *J. Comput. Graph. Stat.* 9 (2) (2000) 319–337.
- [18] R.A. Poldrack, J.A. Mumford, T.E. Nichols, *Handbook of Functional MRI Data Analysis*, Cambridge University Press, 2011.
- [19] RCore Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, 2014. (Available at <http://www.R-project.org/>)
- [20] R. Tibshirani, Regression shrinkage and selection via the lasso, *J. R. Stat. Soc. Ser. B* (1996) 267–288.
- [21] H. Wang, C. Leng, Unified LASSO estimation via least squares approximation, *J. Am. Stat. Assoc.* 102 (479) (2007) 1039–1048.
- [22] L. Yu, W.K. Härdle, L. Borke, T. Benschop, FRM : a financial risk meter based on penalizing tail events occurrence, 2017, [SFB 649 Discussion Paper 2017-003](#)
- [23] L. Zboňáková, W.K. Härdle, W. Wang, Time varying quantile lasso, in: *Applied Quantitative Finance*, 3rd ed., 2017, pp. 331–353, doi:10.1007/978-3-662-54486-0.
- [24] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, *J. R. Stat. Soc.* 67 (2) (2005) 301–320.
- [25] H. Zou, T. Hastie, R. Tibshirani, Sparse principal component analysis, *J. Comput. Graph. Stat.* 15 (2) (2006) 265–286.
- [26] R.P. Monti, et al., Decoding time-varying functional connectivity networks via linear graph embedding methods, *Front. Comput. Neurosci.* 11 (2017) 14.
- [27] R. Lorenz, I.R. Violante, R.P. Monti, G. Montana, A. Hampshire, R. Leech, Dissociating frontoparietal brain networks with neuroadaptive Bayesian optimization, *Nature communications* 9 (1), 1227.
- [28] R. Lorenz, R.P. Monti, A. Hampshire, Y. Koush, C. Anagnostopoulos, Towards tailoring non-invasive brain stimulation using real-time fMRI and Bayesian optimization, in: *International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, 2016, pp. 1–4.
- [29] R. Lorenz, R.P. Monti, I.R. Violante, C. Anagnostopoulos, A.A. Faisal, The automatic neuroscientist: a framework for optimizing experimental design with closed-loop real-time fMRI, *NeuroImage* 129, 320–334.
- [30] R.P. Monti, A. Hyvärinen, A unified probabilistic model for learning latent factors and their connectivities from high-dimensional data, in: *34th International Conference on Uncertainty in Artificial Intelligence*, 2018.
- [31] R. Monti, et al., Graph embeddings of dynamic functional connectivity reveal discriminative patterns of task engagement in HCP data, in: *2015 International Workshop on Pattern Recognition in Neuroimaging*, IEEE, 2015.

Article

Forecasting in Blockchain-Based Local Energy Markets

Michael Kostmann ^{1,*}  and Wolfgang K. Härdle ^{2,3,4}

¹ School of Business and Economics, Humboldt-Universität zu Berlin, Spandauer Str. 1, 10178 Berlin, Germany

² Ladislaus von Bortkiewicz Chair of Statistics, School of Business and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

³ Wang Yanan Institute for Studies in Economics, Xiamen University, 422 Siming Road, Xiamen 361005, China

⁴ Department of Mathematics and Physics, Charles University Prague, Ke Karlovu 2027/3, 12116 Praha 2, Czech

* Correspondence: michael.kostmann@hu-berlin.de

Received: 2 June 2019; Accepted: 9 July 2019; Published: 16 July 2019



Abstract: Increasingly volatile and distributed energy production challenges traditional mechanisms to manage grid loads and price energy. Local energy markets (LEMs) may be a response to those challenges as they can balance energy production and consumption locally and may lower energy costs for consumers. Blockchain-based LEMs provide a decentralized market to local energy consumer and prosumers. They implement a market mechanism in the form of a smart contract without the need for a central authority coordinating the market. Recently proposed blockchain-based LEMs use auction designs to match future demand and supply. Thus, such blockchain-based LEMs rely on accurate short-term forecasts of individual households' energy consumption and production. Often, such accurate forecasts are simply assumed to be given. The present research tested this assumption by first evaluating the forecast accuracy achievable with state-of-the-art energy forecasting techniques for individual households and then, assessing the effect of prediction errors on market outcomes in three different supply scenarios. The evaluation showed that, although a LASSO regression model is capable of achieving reasonably low forecasting errors, the costly settlement of prediction errors can offset and even surpass the savings brought to consumers by a blockchain-based LEM. This shows that, due to prediction errors, participation in LEMs may be uneconomical for consumers, and thus, has to be taken into consideration for pricing mechanisms in blockchain-based LEMs.

Keywords: blockchain; local energy market; smart contract; smart meter; short-term energy forecasting; machine learning; least absolute shrinkage and selection operator (LASSO); long short-term memory (LSTM); prediction errors; market mechanism; market simulation

JEL Classification: Q47; D44; D47; C53

1. Introduction

The “Energiewende”, or energy transition, is a radical transformation of Germany's energy sector towards carbon free energy production. This energy revolution led in recent years to widespread installation of renewable energy generators [1,2]. In 2017, more than 1.6 million photovoltaic micro-generation units were already installed in Germany [3]. Although this is a substantial step towards carbon free energy production, there is a downside: The increasing amount of distributed and volatile renewable energy resources, possibly combined with volatile energy consumption, presents a serious challenge for grid operators. As energy production and consumption have to be balanced in electricity grids at all times [4], modern technological solutions to manage grid loads and price renewable energy are needed. One possibility to increase the level of energy distribution efficiency

on low aggregation levels is the implementation of local energy markets (LEMs) in a decentralized approach, an example being the Brooklyn Microgrid [5].

LEMs enable interconnected energy consumers, producers, and prosumers to trade energy in near real-time on a market platform with a specific pricing mechanism [6]. A common pricing mechanism used for this purpose are discrete double auctions [7–9]. Blockchain-based LEMs utilize a blockchain as underlying information and communication technology and a smart contract to match future supply and demand and to settle transactions [10]. As a consequence, a central authority that coordinates the market is obsolete in a blockchain-based LEM. Major advantages of such LEMs are the balancing of energy production and consumption in local grids [11], lower energy costs for consumers [12], more customer choice (empowerment) [13], and less power line loss due to shorter transmission distances [14].

In the currently existing energy ecosystem, the only agents involved in electricity markets are utilities and large-scale energy producers and consumers. Household-level consumers and prosumers do not actively trade in electricity markets. Instead, they pay for their energy consumption or they are reimbursed for their infeed of energy into the grid according to fixed tariffs. In LEMs, on the contrary, households are the participating market agents that typically submit offers in an auction [7,15]. This market design requires the participating households to estimate their future energy demand and/or supply, to be able to submit a buy or sell offer to the market [16]. Therefore, accurate forecasts of household energy consumption/production are a necessity for such LEM designs. This is due to the market mechanism employed and does not depend on whether an LEM is implemented on a blockchain or not. However, research on blockchain-based LEM mostly employ market mechanisms that require accurate forecasts of household energy consumption/production making the aspect of forecasting especially relevant here. Despite this, it is frequently assumed in existing research on (blockchain-based) LEMs that such accurate forecasts are readily available (see, e.g., [6–8,16,17]). However, forecasting the consumption/production of single households is difficult due to the inherently high degree of uncertainty, which cannot be reduced by the aggregation of households [18]. Hence, the assumption that accurate forecasts are available cannot be taken in practice to be correct. Additionally, given the substantial uncertainty in individual households' energy consumption or production, prediction errors may have a significant impact on market outcomes.

This is where we focused our research: We evaluated the possibility of providing accurate short-term household-level energy forecasts with existing methods and currently available smart meter data. Moreover, our study aimed to quantify the effect of prediction errors on market outcomes in blockchain-based LEMs. For the future advancement of the field, it seemed imperative that the precondition of accurate forecasts of individual households' energy consumption and production for LEMs is assessed. Because, if the assumption cannot be met, the proposed blockchain-based LEMs may not be a sensible solution to support the transformation of our energy landscape. This, however, is urgently necessary to limit CO₂ emissions and the substantial risks of climate change.

1.1. Related Research

Although LEMs started to attract interest in academia already in the early 2000s, it is still an emerging field [11]. Mainly driven by the widespread adoption of smart meters and Internet-connected home appliances, recent work on LEMs focuses on use cases in developed and highly technologized energy grid systems [19]. While substantial work regarding LEMs in general has been done (e.g., [7,8,15]), there are only few examples of blockchain-based LEM designs in the existing literature. Mengelkamp et al. [10] derived seven principles for microgrid energy markets and evaluated the Brooklyn Microgrid according to those principles. With a more practical focus, Mengelkamp et al. [6] implemented and simulated a local energy market on a private Ethereum-blockchain that enables participants to trade local energy production on a decentralized market platform with no need for a central authority. Münsing et al. [20] similarly elaborate a peer-to-peer energy market concept on a blockchain but focus on operational grid constraints and a fair payment rendering. Additionally,

there are several industry undertakings to put blockchain-based energy trading into practice, such as Grid Singularity (gridsingularity.com) in Austria, Powerpeers (powerpeers.nl) in the Netherlands, Power Ledger (powerledger.io) in Australia, and LO3 Energy (lo3energy.com) in the United States.

Interestingly, none of the above cited works, that employ market mechanisms requiring household energy forecasts for bidding, check whether the assumed availability of such forecasts is given. However, without this assumption, trading through an auction design, as described by, e.g., Block et al. [9] or Buchmann et al. [8], and implemented in a smart contract by Mengelkamp et al. [6] is not possible. Unfortunately, this forecasting task is not trivial due to the extremely high volatility of individual households' energy patterns [18]. However, research by Arora and Taylor [21], Kong et al. [22], Shi et al. [23], and Li et al. [24] shows that advances in the energy forecasting field also extend to household-level energy forecasting problems and serve as a promising basis for the present study.

1.2. Present Research

We investigated the prerequisites necessary to implement blockchain-based distributed local energy markets. In particular this means:

- (a) forecasting net energy consumption and production of private consumers and prosumers one time-step ahead;
- (b) evaluating and quantifying the effects of forecasting errors; and
- (c) evaluating the implications of low forecasting quality for a market mechanism.

The prediction task was fitted to the setup of a blockchain-based LEM. Thereby, the present research distinguishes itself notably from previous studies that solely try to forecast smart meter time series in general. The evaluation of forecasting errors and their implications was based on the commonly used market mechanism for discrete interval, double-sided auctions, while the forecasting error settlement structure was based on the work of Mengelkamp et al. [6]. The following research questions were examined:

1. Which prediction technique yields the best 15-min ahead forecast for smart meter time series measured in 3-min intervals using only input features generated from the historical values of the time series and calendar-based features?
2. Assuming a forecasting error settlement structure, what is the quantified loss of households participating in the LEM due to forecasting errors by the prediction technique identified in Question (a)?
3. Depending on Question (b), what implications and potential adjustments for an LEM market mechanism can be identified?

The present research found that regressing with a least absolute shrinkage and selection operator (LASSO) on one week of historical consumption data is the most suitable approach to household-level energy forecasting. However, this method's forecasting errors still substantially diminish the economical benefit of a blockchain-based LEM. Thus, we conclude that changes to the market designs are the most promising way to still employ blockchain-based LEMs as means to meet some of the challenges generated by Germany's current energy transition.

The remainder of the paper is structured as follows: Section 2 presents the forecasting models and error measures used to evaluate the prediction accuracy. Moreover, it introduces the market mechanism and simulation used to evaluate the effect of prediction errors in LEMs. Section 3 describes the data used. Section 4 presents the prediction results of the forecasting models, evaluates their performance relative to a baseline model and assesses the effect of prediction errors on market outcomes. The insights gained from this are then used to identify potential adjustments for future market mechanisms. Finally, Section 5 concludes with a summary, limitations, and an outlook on further research questions that emerge from the findings of the present research.

All code and data used in the present research are available through the Quantnet website (www.quantlet.de). They can be easily found by entering BLEM (Blockchain-based Local Energy

Markets) into the search bar. As part of the Collaborative Research Center, the Center for Applied Statistics and Economics and the International Research Training Group (IRTG) 1792 at the Humboldt-University Berlin, Quantnet contributes to the goal of strengthening and improving empirical economic research in Germany.

2. Method

To select the forecasting technique, we applied the following criteria:

1. The forecasting technique has to produce deterministic (i.e., point) forecasts.
2. The forecasting technique had—for comparison—to be used in previous studies.
3. The previous study or studies using the forecasting technique had to use comparable data, i.e., recorded by smart meters in 60-min intervals or higher resolution, recorded in multiple households, and not recorded in small and medium enterprises (SMEs) or other business or public buildings.
4. The forecasting task had to be comparable to the forecasting task of the present research, i.e., single consumer household (in contrast to the prediction of aggregated energy time series) and very short forecasting horizon (≤ 24 h).
5. The forecasting technique had to take historical and calendar features only as input for the prediction.
6. The forecasting technique had to produce absolutely and relative to other studies promisingly accurate predictions.

Based on these criteria, two forecasting techniques were selected for the prediction task at hand. As short-term energy forecasting techniques are commonly categorized into statistical and machine learning (or artificial intelligence) methods [25–27], one method of each category was chosen: Long short-term memory recurrent neural network (LSTM RNN) adapted from the procedure outlined by Shi et al. [23] and autoregressive LASSO as implemented by Li et al. [24]. Instead of LSTM RNN, gated recurrent unit (GRU) neural networks could have been used as well. However, despite needing fewer computational resources, their representational power may be lower compared to LSTM RNNs [28] and their successful applicability in household-level energy forecasting has not been proven in previous studies. The forecasting techniques used data from 1 January 2017 to 30 September 2017 as training input and the forecast was evaluated on data from 1 October 2017 to 31 December 2017. This means that no data from autumn were included in the training data. However, this seems unlikely to influence the forecasting performance as the German climate in the months from February to April (which are included in the training data) is comparable to the climate in the months from October to December; the forecasting horizon is very short-term; and the input for the forecasting techniques is too short to reflect any seasonal changes in temperature or sunshine hours.

2.1. Baseline Model

A frequent baseline model used for deterministic forecasts is the simple persistence model [29]. This model assumes that the conditions at time t persist at least up to the period of forecasting interest at time $t + h$. The persistence model is defined as

$$\hat{x}_{t+1} = x_t. \quad (1)$$

There are several other baseline models commonly used in energy load forecasting. Most of them are, in contrast to the persistence model, more sophisticated benchmarks. However, as the forecasting task at hand serves the specific use case of being an input for the bidding process in a blockchain-based LEM, the superiority of the forecasting model over a benchmark model is of secondary importance. Hence, in the present research, only the persistence model served as a baseline for the forecasting techniques presented in Sections 2.2 and 2.3.

2.2. Machine Learning-Based Forecasting Approach

The first sophisticated forecasting technique that was employed in the present research to produce as accurate as possible predictions for the blockchain-based LEM is a machine learning algorithm. Long short-term memory (LSTM) recurrent neural networks (RNN) have been introduced only very recently in load forecasting studies (e.g., [22,23,27,30]).

Neural networks do not need any strong assumptions about their functional form, such as traditional time series models (e.g., autoregressive moving average, ARMA). However, they are universal approximators for finite input [31] and, therefore, are especially well suited for the prediction of volatile time series such as energy consumption or production. The most basic building blocks of any neural network are three types of layers: an input layer, one or more hidden layer(s), and an output layer. Each layer consists of one or more units (sometimes called neurons). Each unit in a layer takes in an input, applies a transformation to this input, and outputs it to the next layer. Formally, this can be written as

$$\begin{aligned} h_{1,i} &= \phi_1 (W_1 x_i + b_1) \\ h_{2,i} &= \phi_2 (W_2 h_{1,i} + b_2) \\ &\vdots \\ o_i &= \phi_n (W_n h_{(n-1),i} + b_n) = \hat{y}_i, \end{aligned} \quad (2)$$

where n denotes a layer, ϕ_n is the activation function, W_n is the weight matrix, and b_n is the bias vector in layer n . x_i is the i th input vector and o_i is the output value of the output layer, which is the estimation of the true value y_i . The weight matrices and bias vectors in each layer are parameters that are adjusted during the training of the model.

However, such a simple neural network is not particularly well-suited for time series learning [28]. This is because simple neural networks, such as the one described above, do not have an internal state that could retain a memory of previously processed input. That is, to learn a sequence or time series, the described neural network would always need the complete time series as a single input. It cannot retain a memory of something learned in a previous chunk of the time series to apply it to the next chunk that is fed into the model. This problem is tackled by recurrent neural networks.

RNNs still consist of the basic building blocks of units and layers. However, the units not only feed forward the transformed input as output but also have a recurrent connection that feeds an internal state back into the unit as input. Thereby, a RNN unit loops over individual elements of an input sequence, instead of processing the whole sequence in a single step. This means that the RNN unit applies the transformation to the first element of the input sequence and combines it with its internal state. This introduces the notion of time into neural networks. Formally, this can be written as

$$\begin{aligned} h_{1,t} &= \phi_1 (W_1^{(i)} x_t + W_1^{(r)} h_{1,(t-1)} + b_1) \\ h_{2,t} &= \phi_2 (W_2^{(i)} h_{1,t} + W_2^{(r)} h_{2,(t-1)} + b_2) \\ &\vdots \\ o_t &= \phi_n (W_n^{(i)} h_{(n-1),t} + b_n) = \hat{y}_t, \end{aligned} \quad (3)$$

where n denotes a layer, ϕ_n is the activation function, $W_n^{(i)}$ is the weight matrix for the input, $W_n^{(r)}$ is the weight matrix for the recurrent input (i.e., the output of layer n in the previous time step), and b_n is the bias vector in layer n . x_t is the input vector at time t and o_t is the output value of the output layer which is the estimation of the true value y_t . Note that the output layer has no recurrent units but is the same as in a simple feed forward network.

The cyclical structure of an RNN unit can be unrolled across time (see Figure 1). This illustrates that a RNN is basically a simple neural network that has one layer for each time step that has to be processed per input. Theoretically, this feedback structure enables RNNs to retain information about sequence elements that have been processed many steps before the current step and use it for the prediction of the current step. However, in practice, the vanishing gradient problem occurs (for more details on the vanishing gradient problem, see, e.g., [32]). This problem makes RNNs basically untrainable for very long sequences.

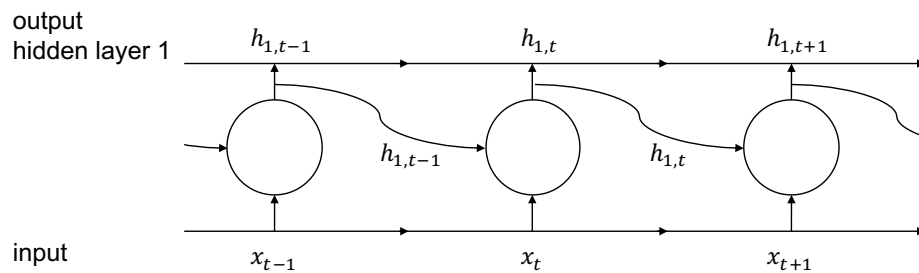


Figure 1. Schematic representation of an unfolded RNN unit. Adapted from [28].

To overcome the vanishing gradient problem, Hochreiter and Schmidhuber [33] developed LSTM units. LSTM RNN is an advanced architecture of RNN that is particularly well suited to learn long sequences or time series due to its ability to retain information over many time steps [28]. LSTM units extend RNN units by an additional state. This state can retain information for as long as needed. In which step this additional state is updated and in which state the information it retains is used in the transformation of the input is controlled by three so-called gates [34]. These three gates have the form of a simple RNN cell. Formally, by slightly adapting the notation of Lipton et al. [35]—who used h_{t-1} instead of s_{t-1} , whereas the notation used here (s_{t-1}) accounts for the modern LSTM architecture with peephole connections—the gates can be written as

$$\begin{aligned} i_t &= \sigma \left(W^{(ix)} x_t + W^{(is)} s_{t-1} + b_i \right) \\ f_t &= \sigma \left(W^{(fx)} x_t + W^{(fs)} s_{t-1} + b_f \right) \\ o_t &= \sigma \left(W^{(ox)} x_t + W^{(os)} s_{t-1} + b_o \right), \end{aligned} \quad (4)$$

where σ is the sigmoid activation function $\sigma(z) = \frac{1}{1+e^{-z}}$, W denotes the weight matrices that are intuitively labeled (ix for the weight matrix of gate i_t multiplied with the input x_t etc.), and b denotes the bias vectors. Again, following the notation of Lipton et al. [35], the full algorithm of a LSTM unit is given by the three gates specified above, the input node,

$$g_t = \sigma \left(W^{(gx)} x_t + W^{(gh)} h_{t-1} + b_g \right), \quad (5)$$

the internal state of the LSTM unit at time step t ,

$$s_t = g_t \odot i_t + s_{t-1} \odot f_t, \quad (6)$$

where \odot is pointwise multiplication, and the output at time step t ,

$$h_t = \phi(s_t) \odot o_t. \quad (7)$$

The internal structure of a LSTM cell is further clarified in Figure 2. For an intuitive but more detailed explanation of LSTM neural networks, see [28] (Ch. 6.2).

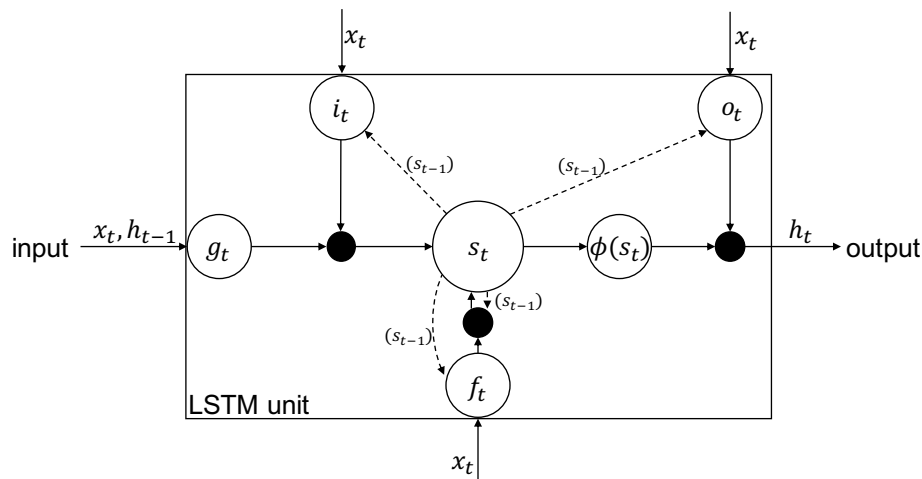


Figure 2. Schematic representation of an LSTM unit. Adapted from [36]. The filled in circles represent the pointwise multiplication operation denoted by \odot in Equations (6) and (7).

In summary, LSTM RNNs are capable of learning highly complex, non-linear relationships in time series data, which makes them a promising forecasting technique to predict households' very short-term energy consumption and production.

The specific LSTM RNN approach adopted in the present research was based on the procedure employed by Shi et al. [23] to forecast individual households' energy consumption. According to the relevant use case in the present research, LSTM RNNs were trained for each household individually using only the household's historic consumption patterns and calendar features. Specifically, seven days of past consumption, an indicator for weekends, and an indicator for Germany-wide holidays were used as input for the neural network in the present research. This follows the one-hot encoding used by Chen et al. [30]. Seven days of lagged data were used as input because preliminary results indicated that the autocorrelation in the time series becomes very weak in lags beyond one week. Moreover, using the previous week as input data still preserves the weekly seasonality and represents a reasonable compromise between as much input as possible and the computational resources needed to process the input in the training process of the LSTM neural network. The target values in the model training were single consumption values in 15-min aggregation. The following example serves as illustration: Assume the consumption values in 3-min intervals from 13 November 2017 13:00 to 20 November 2017 13:00 and zero/one-indicators for weekends and holidays (i.e., 3×3360 data points) are fed into the neural network. The model then produces a single output value that estimates the household's energy consumption in kWh from 20 November 2017 13:00 to 20 November 2017 13:15.

A neural network is steered by several hyperparameters: the number and type of layers, the number of hidden units within each layer, the activation functions used within each unit, dropout rates for the recurrent transformation, and dropout rates for the transformation of the input. To identify a well working combination of hyperparameter values, tuning is necessary which is unfortunately computationally very resource intensive. Table 1 presents the hyperparameters that were tuned and their respective value ranges. The tuning was done individually for each network layer. Optimally, the hyperparameters of all layers should be tuned simultaneously. However, due to computational constraints, that was not possible here and, thus, the described, second-best option was chosen. As the hyperparameter values specified in Table 1 for layer 1 alone result in 81 possible hyperparameter combinations, only random samples of these combinations were taken, the resulting models trained on a randomly chosen dataset and compared. In total, 16 models with one layer, 13 models with two layers and 13 models with three layers were tuned. The model tuning was conducted on four Tesla P100 graphical processing units (GPUs) through the Machine Learning (ML) Engine of the Google Cloud Platform. The job was submitted to the Google Cloud ML Engine via Google Cloud SDK and the R package cloudml. Although neural networks can be trained much faster on GPUs than


on conventional central processing units (CPUs) [28], usage of GPUs through the Google Cloud ML Engine incurs substantial monetary cost. Thus, they were only used for the model tuning in this study.

Table 1. The hyperparameters that were tuned for an optimal LSTM RNN model specification.

	Hyperparameter	Possible Values	Possible Combinations	Sampling Rate	# of Assessed Combinations
layer 1	batch size	{128, 64, 32}	81	0.2	16
	hidden units	{128, 64, 32}			
	recurrent dropout	{0, 0.2, 0.4}			
	dropout	{0, 0.2, 0.4}			
layer 2	hidden units	{128, 64, 32}	26	0.5	13
	recurrent dropout	{0, 0.2, 0.4}			
	dropout	{0, 0.2, 0.4}			
layer 3	hidden units	{128, 64, 32}	26	0.5	13
	recurrent dropout	{0, 0.2, 0.4}			
	dropout	{0, 0.2, 0.4}			

Based on the hyperparameter tuning results, a model with the specification shown in Table 2 was used for the prediction of a single energy consumption value for the next 15 min.

The total length of data points covered in the training process equals the batch size times the input data points times the number of data points that are aggregated for each prediction (i.e., 5 data points): $700 \times 32 \times 5 = 112,000$ data points. This is equivalent to the time period from 1 January 2017 00:00 to 22 August 2017 09:03. The tuning process and results can be replicated by following the Quantlet link in the caption of Table 2.

Table 2. Tuned hyperparameters for LSTRM RNN prediction model.  [BLEMtuneLSTM \(github.com/QuantLet/BLEM/tree/master/BLEMtuneLSTM\)](https://github.com/QuantLet/BLEM/tree/master/BLEMtuneLSTM)

Hyperparameter	Tuned Value
layers	1
hidden units	32
dropout rate	0
recurrent dropout rate	0
batch size	32
number of input data points	3360
number of training samples	700
number of validation samples	96

The general procedure of model training, model assessment and prediction generation is shown in Procedure 1. The parameter tuple was set globally for all household datasets based on the hyperparameter tuning. Thereafter, the same procedure was repeated for each dataset: First, the consumption data time series was loaded, target values were generated, and the input data were transformed. The transformation consisted of normalizing the log-values of the consumption per 3-min interval between 0 and 1. This ensured fast convergence of the model training process. The data batches for the model training and the cross-validation were served to the training algorithm by so-called generator functions. Second, the LSTM RNN was compiled and trained with Keras, which is a neural network application programming interface (API) written in Python. The Keras R package (v2.2.0.9), which was used with RStudio v1.1.453 and TensorFlow 1.11.0 as back-end, is a wrapper of the Python library and is maintained by Chollet et al. [37]. The model training and prediction for each household was performed on a Windows Server 2012 with 12 cores and 24 logical processors of Intel Xeon 3.4 GHz CPUs. The model training was done in a differing number of epochs as early stopping was employed to prevent overfitting: Once the mean absolute error on the validation data did not decrease by more than 0.001 in three consecutive epochs, the training process was stopped. Third, the trained model was used to generate predictions on the test set that comprised data from 1 October 2017 00:00 to 1 January 2018 00:00 (i.e., 44,180 data points). As the prediction was made

in 15-min intervals, in total, 8836 data points were predicted. Using the error measures described in Section 2.4, the model performance was assessed. Finally, the predictions for all datasets were saved for the evaluation in the LEM market mechanism.

Procedure 1 Supervised training of and prediction with LSTM RNN.

- 1: Set parameter tuple $\langle l, u, b, d \rangle$: number of layers $l \subseteq L$, number of hidden LSTM-units $u \subseteq U$, batch size $b \subseteq B$, and dropout rate $d \subseteq D$.
 - 2: Initiate prediction matrix P and list for error measures Θ .
 - 3: **for** Household i in dataset pool I **do**
 - 4: Load dataset Ψ_i .
 - 5: Generate target values y by aggregating data to 15-min intervals.
 - 6: Transform time series in dataset Ψ_i and add calendar features.
 - 7: Set up training and validation data generators according to parameter tuple $\langle b, d \rangle$.
 - 8: Split dataset Ψ_i into training dataset $\Psi_{i,tr}$ and testing dataset $\Psi_{i,ts}$.
 - 9: Build LSTM RNN ζ_i on Tensorflow with network size (l, h) .
 - 10: **repeat**
 - 11: **At** k th epoch **do**:
 - 12: Train LSTM RNN ζ_i with data batches $\varphi_{train} \subseteq \Psi_{i,tr}$ supplied by training data generator.
 - 13: Evaluate performance with mean absolute error Λ_k on cross-validation data batches $\varphi_{val} \subseteq \Psi_{i,tr}$ supplied by validation data generator.
 - 14: **until** $\Lambda_{k-1} - \Lambda_k < 0.001$ for the last 3 epochs.
 - 15: Save trained LSTM RNN ζ_i .
 - 16: Set up testing data generator according to tuple $\langle b, d \rangle$.
 - 17: Generate predictions \hat{y}_i with batches $\varphi_{ts} \subseteq \Psi_{i,ts}$ fed by testing data generator into LSTM RNN ζ_i .
 - 18: Calculate error measures Θ_i to assess performance of X_i .
 - 19: Write prediction vector \hat{y}_i into column i of matrix P .
 - 20: **end for**.
 - 21: Save matrix P .
 - 22: **End**.
-

2.3. Statistical Method-Based Forecasting Approach

To complement the machine learning approach of a LSTM RNN with a statistical approach, a second, regression-based method was used. For this purpose, the autoregressive LASSO approach proposed by Li et al. [24] seemed most suitable. Statistical methods have the advantage of much lower model complexity compared to neural networks which makes them computationally much less resource intensive.

Li et al. [24] used LASSO [38] to find a sparse autoregressive model that generalizes better to new data. Formally, the LASSO estimator can be written as

$$\hat{\beta}_{\text{LASSO}} = \arg \min_{\beta} \frac{1}{2} \|(y - X\beta)\|_2^2 + \lambda \|\beta\|_1, \quad (8)$$

where X is a matrix with row t being $[1 \ x_t^T]$ (the length of x_t^T is the number of lag-orders n available), and λ is a parameter that controls the level of sparsity in the model, i.e., which of the n available lag-orders are included to predict y_{t+1} . This model specification selects the best recurrent pattern in the energy time series by shrinking coefficients of irrelevant lag-orders to zero and, thereby, improves the generalizability of the prediction model. In the present research, the sparse autoregressive LASSO approach was implemented using the R package `glmnet` [39]. As for the LSTM RNN approach, model training and prediction were performed for every household individually. Following the procedure of Li et al. [24], only historical consumption values were used as predictors. Specifically, for comparability to the LSTM approach, seven days of lagged consumption values served as input to the LASSO model. The response vector consisted of single consumption values in 15-min aggregation. The same example as above serves as illustration: Assume the consumption values in 3-min intervals from 13 November 2017 13:00 to 20 November 2017 13:00 (i.e., 3360 data points) are available to the model for prediction. Based on the training data, the model chooses the lagged values with the highest predictive power

and makes a linear estimation of a single value for the household's energy consumption in kWh from 20 November 2017 13:00 to 20 November 2017 13:15.

The detailed description of the model estimation and prediction is presented in Procedure 2. As the LASSO model requires a predictor matrix, the time series of each household was split in sequences of length $n = 3360$ with five data points skipped in between. The skip accounted for the fact that the response vector was comprised of 15-min interval consumption values (i.e., five aggregated 3-min values). After generating the predictor matrix for the model estimation, the optimal λ was found in a K -fold cross-validation. Here, K was set to 10. The sequence of λ -values that was tested via cross-validation was of length $L = 100$ and was constructed by calculating the minimum λ -value as a fraction of the maximum λ -value ($\lambda_{min} = \varepsilon\lambda_{max}$, where λ_{max} was such that all β -coefficients were set equal to zero) and moving along the log-scale from λ_{max} to λ_{min} in L steps. However, the glmnet algorithm used early-stopping to reduce computing times if the percent of null deviance explained by the model with a certain λ did not change sufficiently from one to the next λ -value. The cross-validation procedure identified the biggest λ that is still within one standard deviation of the λ with the lowest mean absolute error. The final coefficients for each household were then computed by solving Equation (8) for the complete predictor matrix. Thereafter, the predictions were made on the testing data. Again, the time series was sliced according to the sliding window of length $n = 3360$ skipping five data points and written into a predictor matrix. This matrix comprised data from 1 October 2017 00:00 to 1 January 2018 00:00 (i.e., 8836 cases of 3360 lagged values), resulting again in 8836 predicted values as in the case of the LSTM approach. The predictions on all datasets were assessed using the error measures described in Section 2.4 and saved for the evaluation of the prediction in the context of the LEM market mechanism.

Procedure 2 Cross-validated selection of λ for LASSO and prediction.

- 1: Initiate prediction matrix P and list for error measures Θ .
 - 2: **for** Household i in dataset pool I **do**
 - 3: Load dataset Ψ_i .
 - 4: Generate target values \mathbf{y} by aggregating data to 15-min intervals.
 - 5: Split dataset Ψ_i into training dataset $\Psi_{i,tr}$ and testing dataset $\Psi_{i,ts}$.
 - 6: Generate predictor matrix M_{tr} by slicing time series $\Psi_{i,tr}$ with sliding window.
 - 7: Generate sequence of λ -values $\{l_s\}_{s=1}^L$.
 - 8: Set number of cross-validation (CV) folds K .
 - 9: Split predictor matrix M_{tr} into K folds.
 - 10: **for** k in K **do**
 - 11: Select fold k as CV testing set and folds $j \neq k$ as CV training set.
 - 12: **for** each l_s in $\{l_s\}_{s=1}^L$ **do**
 - 13: Compute vector $\hat{\beta}_{k,l_s}$ on CV training set.
 - 14: Compute mean absolute error Λ_{k,l_s} on CV testing set.
 - 15: **end for.**
 - 16: **end for.**
 - 17: For each $\hat{\beta}_{k,l_s}$ calculate average mean absolute error $\bar{\Lambda}_s$ across the K folds.
 - 18: Select cross-validated λ -value l_s^{CV} with the highest regularization (min no. of non-zero β -coeff.) within one SD of the minimum $\bar{\Lambda}_s$.
 - 19: Compute $\hat{\beta}_{l_s^{CV}}$ on complete predictor matrix M_{tr} .
 - 20: Generate predictor matrix M_{ts} by slicing time series $\Psi_{i,ts}$ with sliding window.
 - 21: Generate predictions $\hat{\mathbf{y}}_i$ from predictor matrix M_{ts} and coefficients $\hat{\beta}_{l_s^{CV}}$.
 - 22: Calculate error measures Θ_i to assess performance.
 - 23: Write prediction vector $\hat{\mathbf{y}}_i$ into column i of matrix P .
 - 24: **end for.**
 - 25: Save matrix P .
 - 26: **End.**
-

2.4. Error Measures

Forecasting impreciseness is measured by a variety of norms. The L_1 -type mean absolute error (MAE) is defined as the average of the absolute differences between the predicted and true values [40]:

$$\text{MAE} = \frac{1}{N} \sum_{t=1}^N |\hat{x}_t - x_t|, \quad (9)$$

where N is the length of the forecasted time series, \hat{x}_t is the forecasted value and x_t is the observed value. As MAE is only a valid error measure if one can assume that for the forecasted distribution the mean is equal to the median (which might be too restrictive), an alternative is the root mean square error (RMSE), i.e., the square root of the average squared differences [29,41]:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{t=1}^N (\hat{x}_t - x_t)^2}. \quad (10)$$

Absolute error measures are not scale independent, which makes them unsuitable to compare the prediction accuracy of a forecasting model across different time series. Therefore, they are complemented with the percentage error measures mean absolute percentage error (MAPE) and normalized root mean square error (NRMSE) normalized by the true value:

$$\text{MAPE} = \frac{100}{N} \sum_{t=1}^N \left| \frac{\hat{x}_t - x_t}{x_t} \right|, \quad (11)$$

and

$$\text{NRMSE} = \sqrt{\frac{100}{N} \sum_{t=1}^N \left(\frac{\hat{x}_t - x_t}{x_t} \right)^2}. \quad (12)$$

However, as Hyndman and Koehler [42] pointed out, using x_t as denominator may be problematic as the fraction $\frac{\hat{x}_t - x_t}{x_t}$ is not defined for $x_t = 0$. Therefore, time series containing zero values cannot be assessed with this definition of the MAPE and NRSME.

To overcome the shortage of an undefined fraction in the presence of zero values in the case of MAPE and NRMSE, the mean absolute scaled error (MASE) as proposed by Hyndman and Koehler [42] was used. That is, MAE was normalized with the in-sample mean absolute error of the persistence model forecast:

$$\text{MASE} = \frac{\text{MAE}}{\frac{1}{n-1} \sum_{t=2}^N |x_t - x_{t-1}|}. \quad (13)$$

In summary, in the present research, the forecasting performance of the LSTM RNN and the LASSO were evaluated using MAE, RMSE, MAPE, NRMSE, and MASE.

2.5. Market Simulation

We used a market mechanism with discrete closing times in 15-min intervals. Each consumer and each prosumer submit one order per interval and the asks and bids are matched in a closed double auction that yields a single equilibrium price. The market mechanism was implemented in R. This allows for a flexible and time-efficient analysis of the market outcomes with and without prediction errors.

The simulation of the market mechanism followed five major steps: First, the consumption and production values of each market participant per 15-min interval from 1 October 2017 00:00 to 1 January 2018 00:00 were retrieved. These values are either the true values as yielded by the aggregation of the raw data or the prediction values as estimated by the best performing prediction model. Second, for each market participant, a zero-intelligence limit price was generated by drawing randomly from

the discrete uniform distribution $U\{12.31, 24.69\}$. The lower bound is the German feed-in tariff of $12.31 \frac{\text{EURct}}{\text{kWh}}$ and the upper bound is the average German electricity price in 2016 of $28.69 \frac{\text{EURct}}{\text{kWh}}$ [43]. This agent behavior has been shown to generate efficient market outcomes in double auctions [44] and is rational in so far as electricity sellers would not accept a price below the feed-in tariff and electricity buyers would not pay more than the energy utility's price per kWh. However, this assumes that the agents do not consider any non-price related preferences, such as strongly preferring local renewable energy [6]. Third, for each trading slot (i.e., every 15-min interval), the bids and asks were ordered in price-time precedence. Given the total supply is lower than the total demand, the lowest bid price that can still be served determines the equilibrium price. Given the total supply is higher than the total demand, the overall lowest bid price determines the equilibrium price. In the case of over- or undersupply, the residual amounts are traded at the feed-in ($12.31 \frac{\text{EURct}}{\text{kWh}}$) or the regular household consumer electricity tariff ($28.69 \frac{\text{EURct}}{\text{kWh}}$) with the energy utility. Fourth, the applicable price for each bid and ask was determined and the settlement amounts, resulting from this price and the energy amount ordered, were calculated. In the case of using predicted values for the bids, there was an additional fifth step: After the next trading period, when the actual energy readings were known, any deviations between predictions and true values were settled with the energy utility using the feed-in or household consumer electricity tariff. This led to correction amounts that were deducted or added to the original settlement amounts. For the market simulation, perfect grid efficiency and, hence, no transmission losses were assumed.

3. Data

The raw data used for the present research were provided by Discovery GmbH and are available at [BLEMdata \(github.com/QuantLet/BLEM/tree/master/data\)](https://github.com/QuantLet/BLEM/tree/master/data), hosted at GitHub. Discovery describes itself as a full-range supplier of smart metering solutions offering transparent energy consumption and production data for private and commercial clients [45]. To be able to offer such data-driven services, Discovery smart meters record energy consumption and production near real-time—i.e., in 2-s intervals—and send the readings to Discovery's servers for storage and analysis. Therefore, Discovery has extremely high resolution energy data of their customers at their disposal. This high resolution is in stark contrast to the half-hourly or even hourly recorded data used in previous studies on household energy forecasting (e.g., [21,23,46,47]). To our knowledge, there is no previous research using Discovery smart meter data, apart from Teixeira et al. [48], who used the data as simulation input but not for analysis or prediction.

The data come in 200 individual datasets each containing the meter readings of a single smart meter; 100 datasets belong to pure energy consumers and 100 datasets belong to energy prosumers (households that produce and consumer energy). The meter readings were aggregated to 3-min intervals and range from 1 January 2017 00:00 to 1 January 2018 00:00. This translated into 175,201 observations per dataset. Each observation consists of the total cumulative energy consumption and the total cumulative energy production from the date of installation until time t , current power over all phases installed in the meter at time t and a timestamp in Unix milliseconds.

For further analysis, the power readings were dropped and the first differences of the energy consumption and production readings were calculated. These first differences are equivalent to the energy consumption and production within each 3-min interval between two meter recordings. The result of this computation leaves each dataset with two time series (energy consumption and energy production in kWh) and 175,200 observations.

Figure 3 shows the energy consumption time series of Consumer 082. In the first panel of Figure 3, the consumption per 3-min interval for all of 2017 is shown. Notably, there are two extended periods (in March and June) and three shorter periods (in July, September, and December) with a clearly distinguishable low consumption level and low fluctuation. The most likely explanations for these low, stable energy consumption periods are holidays, in which the household members are on vacation and leave appliances that are on standby or always turned on as the only energy consumers.

The second panel zooms to just one month making daily fluctuation patterns visible. The last panel zooms in to a single day of energy consumption. It exemplifies well a usual pattern of energy consumption: There is low and rather stable energy consumption from midnight until about 07:30, which only fluctuates in a systematic and repeated way due appliances in standby and “always on” appliances, such as a fridge and/or freezer. At around 07:30, the household members probably wake up and the energy consumption spikes for the next 30 min—the lights are turned on, coffee is made, the stove is turned on, and maybe a flow heater is used to shower with hot water. As the household members leave the house (13 May is a Monday), the consumption slowly decreases again. In the evening at about 18:30 the energy consumption spikes again, probably caused by dinner preparations.

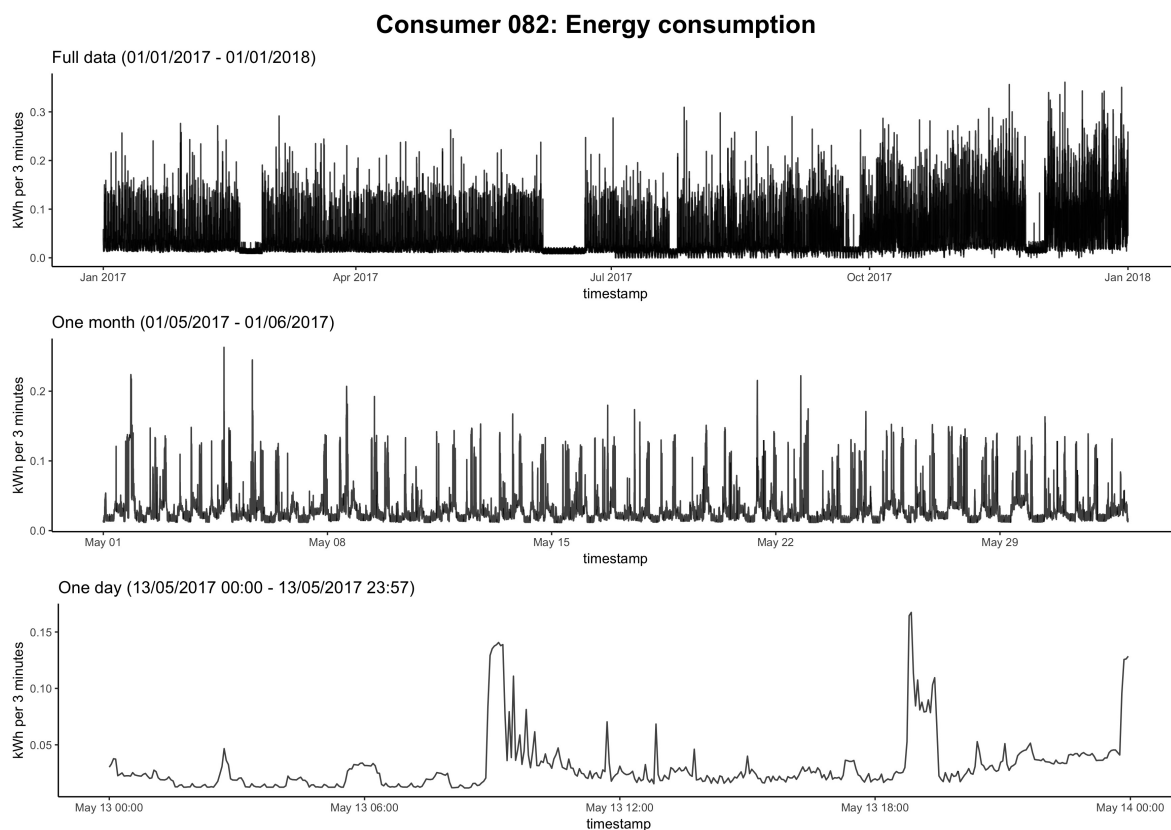


Figure 3. Energy consumption recordings of Consumer 082. The first panel shows the full year 2017, the second panel zooms in to one month (May), and the third panel zooms in to one day (13 May).

[BLEMplotEnergyData \(github.com/QuantLet/BLEM/tree/master/BLEMplotEnergyData\)](https://github.com/QuantLet/BLEM/tree/master/BLEMplotEnergyData)

Out of the 100 consumer datasets, five exhibited non-negligible shares of zero consumption values leading to their exclusion. One consumer dataset was excluded as the consumption time series was flat for the most part of 2017 and one consumer was excluded due to very low and stable consumption values with very rare, extreme spikes. Four more consumers were excluded due to conspicuous regularity in daily or weekly consumption patterns. Lastly, one consumer was excluded not due to peculiarities in the consumption patterns but due to missing data. As the inclusion of this shorter time series would have led to difficulties in the forecasting algorithms, this dataset was excluded as well.

Out of the 100 prosumer datasets, 86 were excluded due to zero total net energy production in 2017. These “prosumers” would not act as prosumers in an LEM as they would never actually supply a production surplus to the market. Of the remaining 14 prosumer datasets, one prosumer dataset was excluded because the total net energy it fed into the grid in 2017 was just 22 kWh. Additionally,

one prosumer dataset was excluded as it only fed energy into the grid in the period from 6 January 2017 to 19 January 2017. For all other measurement points, the net energy production was zero.

Overall, 88 consumer and 12 prosumer datasets remained for the analysis. All datasets include a timestamp and the consumption time series for consumers and the production time series for prosumers with a total of 175,200 data points each.

4. Results

4.1. Evaluation of the Prediction Models

Three prediction methods were used to forecast the energy consumption of 88 consumer households 15 min ahead: a baseline model, a LSTM RNN model, and a LASSO regression. All three prediction models were compared and evaluated using the error measures presented in Section 2.4. The performance of the prediction models was tested on a quarter of the available data. That is, the prediction models were fitted on the consumption values from 1 January 2017 00:00 to 30 September 2017 00:00, which is equivalent to 131,040 data points per dataset. For all 88 consumer datasets, the models were fitted separately resulting in as many distinct LASSO and LSTM prediction models. The fitted models were then used to make energy consumption predictions in 15-min intervals for each household individually on the data from 1 October 2017 00:00 to 1 January 2018 00:00. This equates to 8836 predicted values per dataset per prediction method.

Figure 4 displays the total sum of over- and underestimation errors in kWh of each prediction method per dataset. That is, for each consumer, the total sum of overestimation errors is calculated as summing all differences between true and forecasted value, when the forecasted value is greater than the true value (formally, $\delta_i^o = \sum_{t=1}^N (\hat{x}_{i,t} - x_{i,t}) [(\hat{x}_{i,t} - x_{i,t}) > 0]$; red bars) and the total sum of underestimation errors is calculated as summing all differences between true and forecasted value, when the forecasted value is smaller than the true value (formally, $\delta_i^u = \sum_{t=1}^N (\hat{x}_{i,t} - x_{i,t}) [(\hat{x}_{i,t} - x_{i,t}) < 0]$; blue bars). Thus, the red and blue bars added together depict the total sum of errors in kWh for each prediction method per dataset.

The LASSO technique achieved overall lower total sums of errors than the baseline model. Notably, the sum of underestimation errors is higher across the datasets than the sum of overestimation errors. This points towards a general tendency of underestimating sudden increases in energy consumption by the LASSO technique. The LSTM model on the other hand shows a much higher variability in the sums of over- and underestimation errors. By tendency, the overestimation errors of the LSTM model are smaller than those of the LASSO and baseline model. Nevertheless, the underestimation is much more pronounced in the case of the LSTM model. Especially, some datasets stand out regarding the high sum of underestimation errors. This points towards a much higher heterogeneity in the suitability of the LSTM model to predict consumption values depending on the energy consumption pattern of the specific dataset. The LASSO technique on the other hand seems to be more equally well suited for all datasets and their particular consumption patterns.

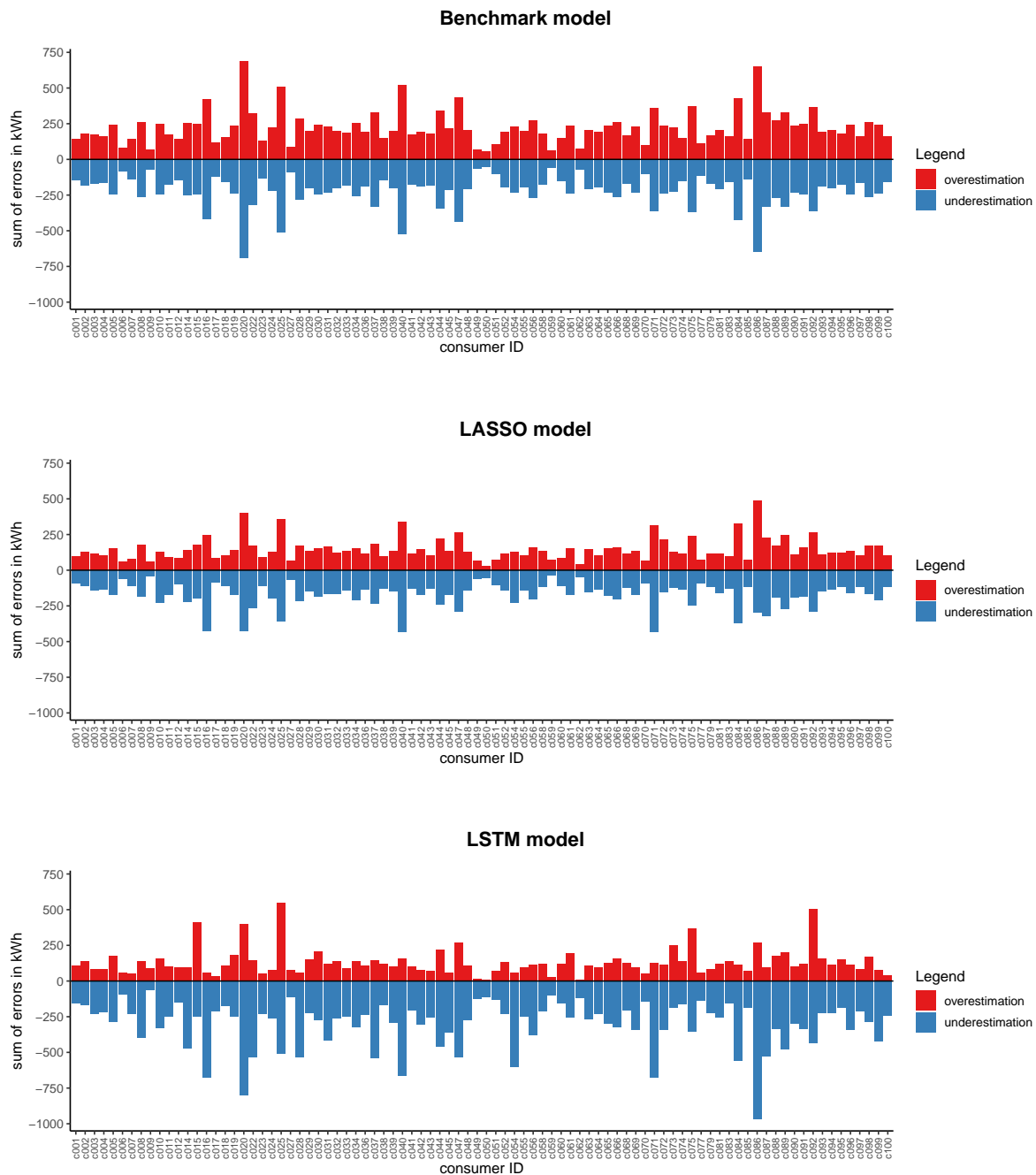


Figure 4. Sum of total over- and underestimation errors of energy consumption per consumer dataset and prediction model. [BLEMplotPredErrors \(github.com/QuantLet/BLEM/tree/master/BLEMplotPredErrors\)](https://github.com/QuantLet/BLEM/tree/master/BLEMplotPredErrors)

The average performance of the three prediction models across all 88 datasets is shown in Table 3. As can be seen, LASSO and LSTM consistently outperformed the baseline model according to MAE, RMSE, MAPE, NRMSE and MASE. The LASSO model performed best overall with the lowest median error measure scores across the 88 consumer datasets.

Table 3. Median of error measures for the prediction of energy consumption across all 88 consumer datasets. [BLEMevaluateEnergyPreds](https://github.com/QuantLet/BLEM/tree/master/BLEMevaluateEnergyPreds) (github.com/QuantLet/BLEM/tree/master/BLEMevaluateEnergyPreds)

Model	MAE	RMSE	MAPE	NRMSE	MASE
LSTM	0.04	0.09	22.22	3.30	0.85
LASSO	0.03	0.05	17.38	2.31	0.57
Benchmark	0.05	0.10	27.98	5.08	1.00
Improvement LSTM (in %)	16.21	12.61	20.57	34.98	14.78
Improvement LASSO (in %)	44.02	48.73	37.88	54.61	43.02

The superior performance of the LASSO model is also clearly visible in Figure 5. This might be surprising, as from a theoretical point of view, a linear model should not outperform a non-linear neural network that fulfills the conditions for a universal approximator for finite input. The most reasonable explanation seems to be that the LSTM RNN model used here missed a good local minimum for a number of datasets and converged to suboptimal parameter combinations. If the main focus of this paper were finding an optimal forecasting algorithm for individual households' short-term energy consumption, this would require further investigation. However, this study focused on the achievable forecasting accuracy with state-of-the-art methods already employed in previous studies. The results imply that it seems unwise to use a general set of hyperparameters on a number of household energy consumption datasets that differ quite substantially in their energy consumption patterns. However, as the LASSO technique employed here achieved an error score that is competitive with comparable research applications, the underperformance of the LSTM RNN compared to the LASSO technique is of no further concern.

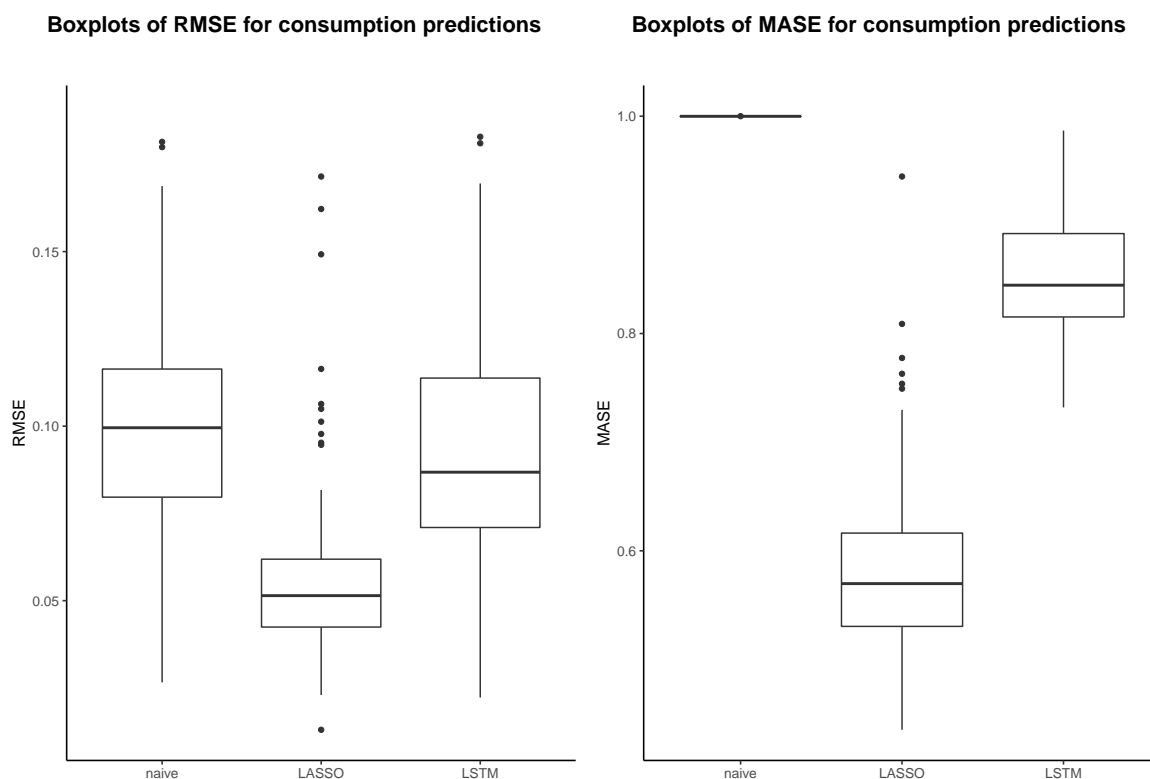


Figure 5. Box plots of RMSE and MASE scores across 88 consumer datasets for the three different prediction models (the upper 3%-quantile of the error measures is cut off for better readability). [BLEMevaluateEnergyPreds](https://github.com/QuantLet/BLEM/tree/master/BLEMevaluateEnergyPreds) (github.com/QuantLet/BLEM/tree/master/BLEMevaluateEnergyPreds)

Interestingly, some consumer datasets exhibit apparently much harder to predict consumption patterns than the other datasets. This is exemplified by the outliers of the box plots, as well as by the heat map displayed in Figure 6. It confirms that there is quite some variation among the same prediction methods across different households. Therefore, one may conclude that there is no “golden industry standard” approach for households’ very short-term energy consumption forecasting. Nevertheless, it is obvious that the LASSO model performed best overall. Hence, the predictions on the last quarter of the data produced by the fitted LASSO model for each consumer dataset were used for the evaluation of the market simulation presented next.

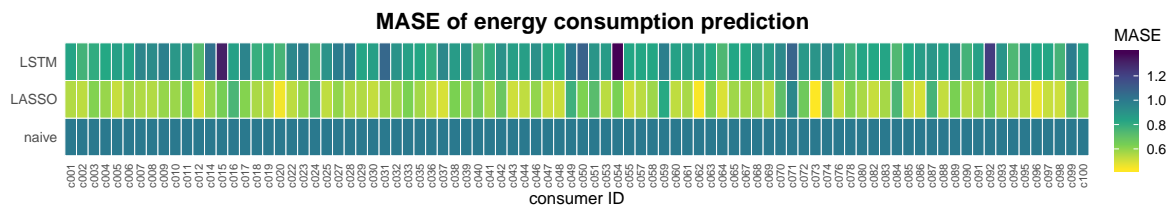


Figure 6. Heat map of MASE scores for the prediction of consumption values per consumer dataset.

 [BLEMevaluateEnergyPreds \(github.com/QuantLet/BLEM/tree/master/BLEMevaluateEnergyPreds\)](https://github.com/QuantLet/BLEM/tree/master/BLEMevaluateEnergyPreds)

4.2. Evaluation of the Market Simulation

The market simulation used the market mechanism of a discrete interval, closed double auction to assess the impact of prediction errors on market outcomes. In total, 88 consumers and 12 prosumer datasets were available. To evaluate different supply scenarios, the market simulation was conducted three times with a varying number of prosumers included. The three scenarios consisted of a market simulation with balanced energy supply and demand, a simulation with severe oversupply and a simulation with severe undersupply. To avoid extreme and unusual market outcomes over the time period of the simulation, two prosumers with high production levels, but long periods of no energy production in the simulation period were not included as energy suppliers in the market. The remaining prosumers were in- or excluded according to the desired supply scenario. That is, the undersupply scenario comprised six prosumers, the balanced supply scenario additionally included one more, and the oversupply scenario included additionally to the balanced supply scenario two more prosumers.

4.2.1. Market Outcomes in Different Supply Scenarios

The difference between supply and demand for each trading period, the equilibrium price of each double auction, and the weighted average price—termed LEM price—is shown in Figure 7. The LEM price is computed in each trading period as the average of the auctions equilibrium price and the energy utilities energy price ($28.69 \frac{\text{EURct}}{\text{kWh}}$) weighted by the amount of kWh traded for the respective price. The three graphs below depicting the market outcomes are results of the market simulation with true consumption values.

As can be seen, the equilibrium price shown in the middle panel of Figure 7 moves roughly synchronous to the over-/undersupply shown in the top panel. As there is by tendency more undersupply in the balanced scenario (the red line in the top panel indicates perfectly balanced supply and demand), the equilibrium price is in most trading periods close to its upper limit and the LEM price is almost always above the equilibrium price. There is by tendency more undersupply due to the fact that four of the relevant prosumer datasets are from producers with large capacities (>10 kWh per 15-min interval) that dominated the remaining prosumers’ production capacity substantially and therefore a more balanced supply scenario could not be created.

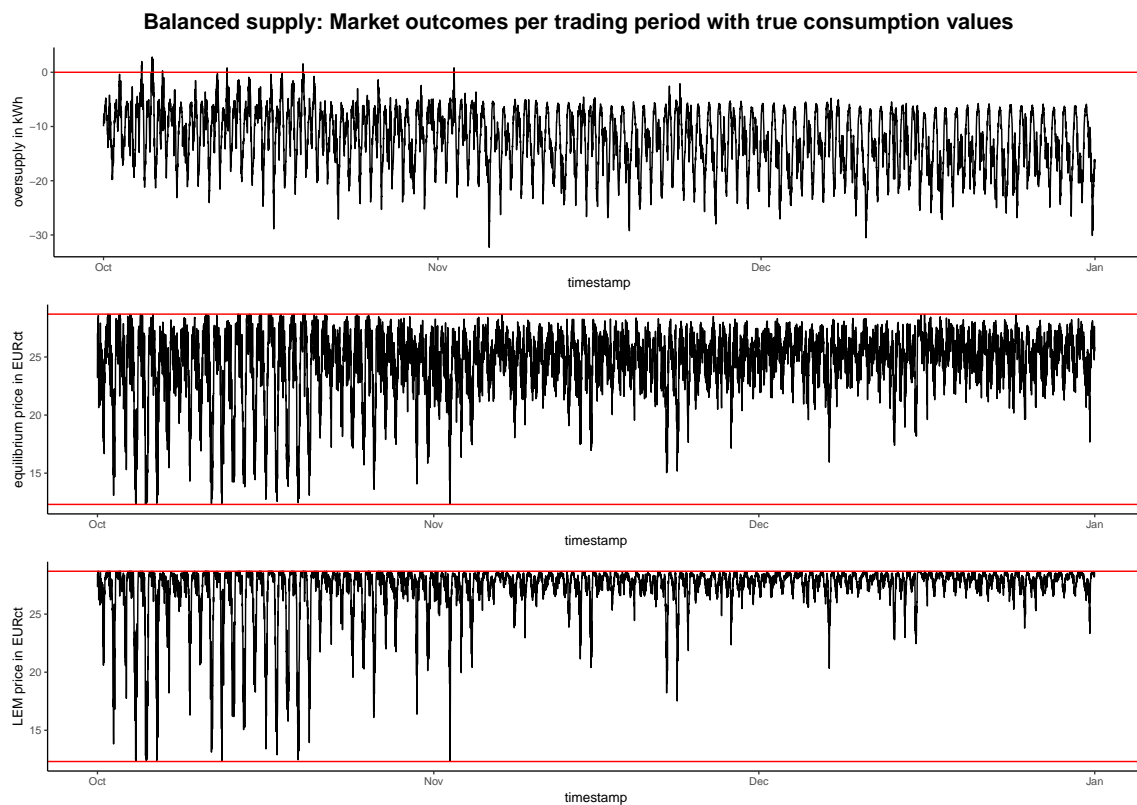


Figure 7. Market outcomes per trading period simulated with true values and a balanced supply scenario. [BLEMmarketSimulation \(github.com/QuantLet/BLEM/tree/master/BLEMmarketSimulation\)](https://github.com/QuantLet/BLEM/tree/master/BLEMmarketSimulation)

This observation is in contrast to the oversupply scenario shown in Figure 8. Here, the prosumers' energy supply surpasses the consumers' energy demand in the majority of trading periods. Accordingly, the equilibrium price in each auction is close to the lower limit of the energy utility's feed-in tariff of $12.31 \frac{\text{EURct}}{\text{kWh}}$. However, trading periods with undersupply lead to visible spikes in the equilibrium price, which are, as expected, even more pronounced in the LEM price. In all other periods, the equilibrium price equals the LEM price as all demand is served by the prosumers and there is no energy purchased from the grid.

Figure 9 shows the market simulation performed in an undersupply scenario. Here, the market outcomes are the opposite to the oversupply scenario: The equilibrium prices move in a band between $20 \frac{\text{EURct}}{\text{kWh}}$ and the upper limit of $28.69 \frac{\text{EURct}}{\text{kWh}}$. The LEM prices are even higher as the deficit in supply has to be compensated by energy purchases from the grid. This means that, the more severe the undersupply is, the more energy has to be purchased from the grid, and the more the LEM price surpasses the equilibrium price.

In summary, one can conclude that the market outcomes are the more favorable to consumers, the more locally produced energy is offered by prosumers. Assuming a closed double auction as market mechanism and zero-intelligence bidding behavior of market participants, oversupply reduces the LEM prices substantially leading to savings on the consumer side. On the other hand, prosumers will favor undersupply in the market as they profit from the high equilibrium prices while still being able to sell their surplus energy generation at the feed-in tariff without a loss compared to no LEM.

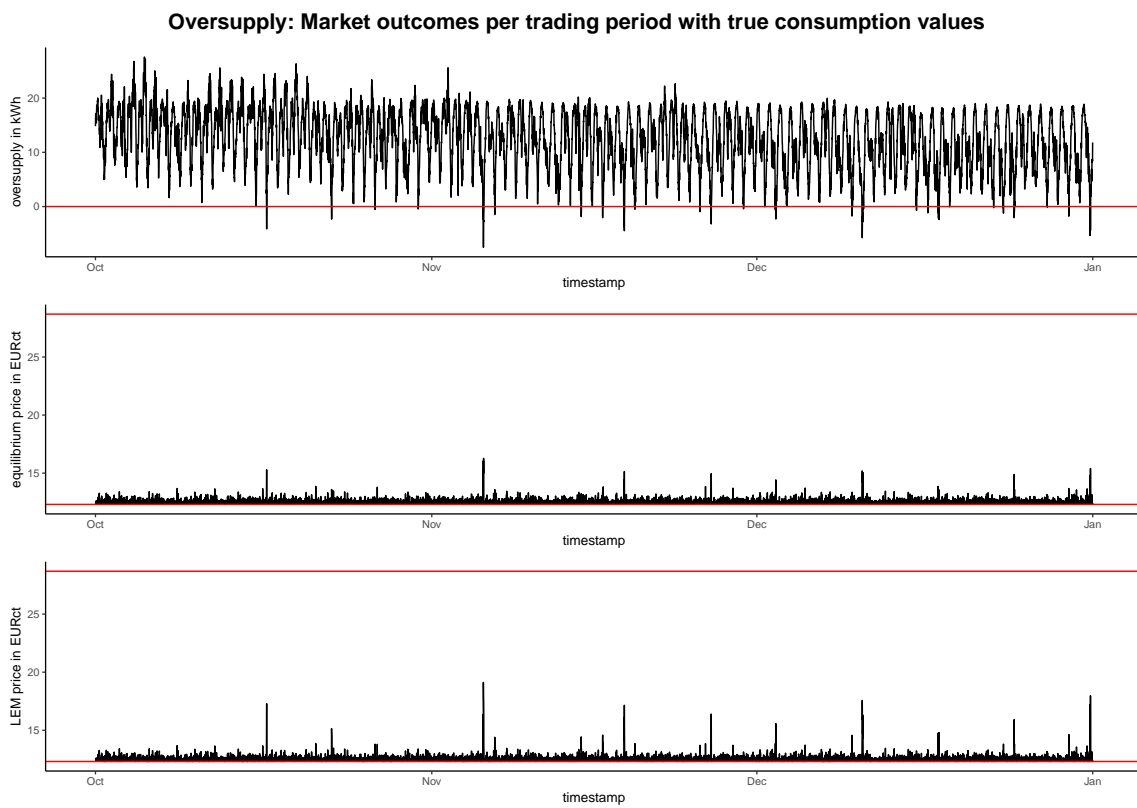


Figure 8. Market outcomes per trading period simulated with true values and an oversupply scenario.
[BLEMmarketSimulation \(github.com/QuantLet/BLEM/tree/master/BLEMmarketSimulation\)](https://github.com/QuantLet/BLEM/tree/master/BLEMmarketSimulation)

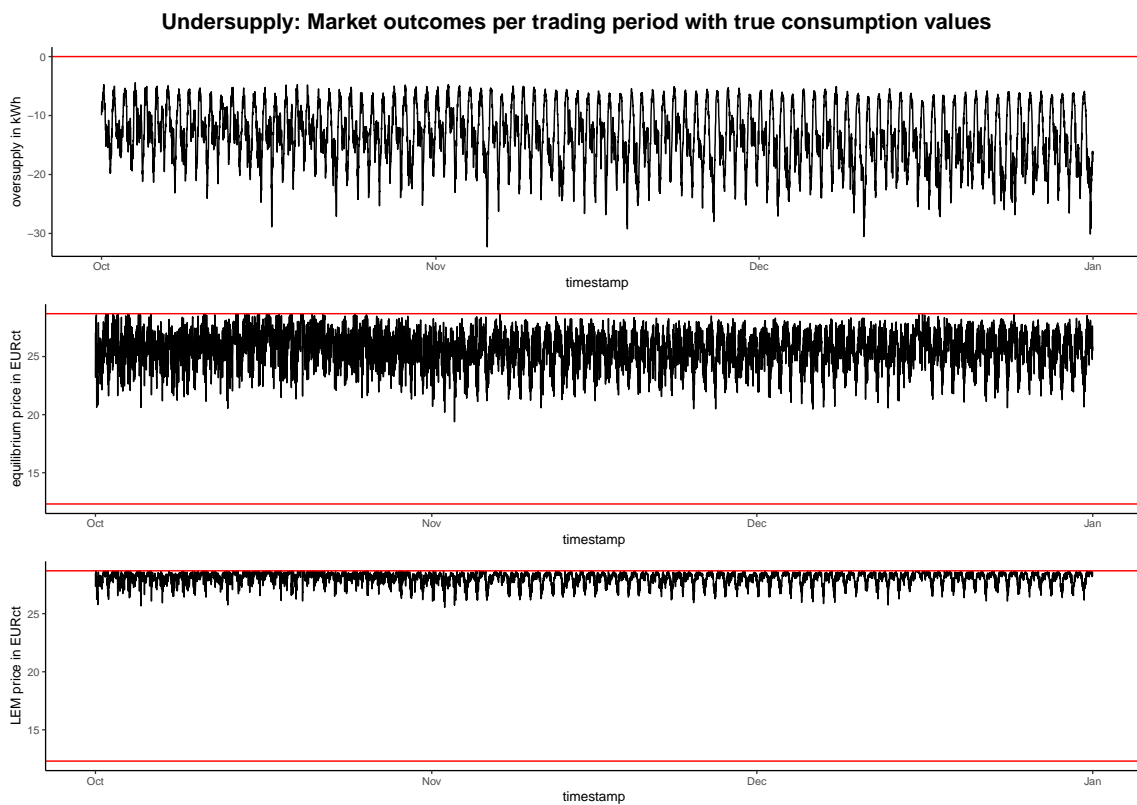


Figure 9. Market outcomes per trading period simulated with true values and an undersupply scenario.
[BLEMmarketSimulation \(github.com/QuantLet/BLEM/tree/master/BLEMmarketSimulation\)](https://github.com/QuantLet/BLEM/tree/master/BLEMmarketSimulation)

4.2.2. Loss to Consumers due to Prediction Errors

To assess the adverse effect of prediction errors on market outcomes, the LASSO-predicted energy consumption values per 15-min interval are used. The predictions of the model served as order amounts in the auction bids. After the true consumption in the respective trading period was observed, payments to settle over- or underestimation errors were made. That is, if a consumer bid with a higher amount than actually consumed, it still bought the full bid amount from the prosumers but had to sell the surplus to the energy utility over the grid at the feed-in tariff. On the other hand, if a consumer bid with a lower amount than actually consumed, it bought the bid amount from the prosumers but had to purchase the surplus energy consumption from the grid at the energy utility's tariff. Thus, prediction errors are costly as the consumer always has to clear the order in less favorable conditions than the equilibrium price provides.


Table 4 contrasts the results of the market simulation with true consumption values with the results of the market simulation with predicted consumption values in three different supply scenarios. The equilibrium and LEM prices almost do not differ within the three scenarios whether the true or predicted consumption values are used. The prices between the scenarios, however, differ substantially. The average total revenue over the three-month simulation period of prosumers is largely unaffected by the use of true or predicted consumption values. This is not surprising as the revenue is a function of the equilibrium price, which is apparently largely unaffected by whether true or predicted consumption values are used, and the electricity produced, which is obviously completely unaffected by whether true or predicted consumption values are used.

Table 4. Average results of the market simulation for three different supply scenarios. Prices are averaged across all trading periods. Revenues and costs for the whole simulation period are averaged across all prosumers and consumers, respectively. [BLEMevaluateMarketSim \(github.com/QuantLe/t/BLEM/tree/master/BLEMevaluateMarketSim\)](https://github.com/QuantLe/t/BLEM/tree/master/BLEMevaluateMarketSim)

Mean	Balanced Supply		Oversupply		Undersupply	
	True	Predicted	True	Predicted	True	Predicted
Equilibrium price (in EURct)	24.64	24.61	12.50	12.49	25.68	25.69
LEM price (in EURct)	27.31	27.28	12.51	12.49	28.08	28.10
Revenue (in EUR)	1113.84	1108.88	3454.62	3451.69	1035.90	1036.12
Cost with LEM (in EUR)	439.26	457.94	200.75	226.61	451.60	470.69
Cost without LEM (in EUR)	459.83	446.93	459.83	446.93	459.83	446.93

What differs according to Table 4, however, is the cost for consumers. The cost without the LEM is on average across all consumers smaller when using predicted consumption values compared to using true consumption values. This can be explained by the LASSO model's tendency to underestimate on the data at hand and because correction payments for the prediction errors are not factored into this number. The average total cost for electricity consumption in the whole simulation period is with an LEM higher when using predicted consumption values compared to using true consumption values. This is due to the above-mentioned need to settle prediction errors at unfavorable terms.

The percentage loss induced by prediction errors is shown in Table 5. Depending on the supply scenario it ranges between about 4.8% and 13.75%. These numbers have to be judged relative to the savings that are brought to consumers by the participation in an LEM. It turns out, that in the balanced supply scenario, the savings due to the LEM are almost completely offset by the loss due to prediction errors. As consumers profit more from an LEM, the lower the equilibrium prices are, this is not the case in the oversupply scenario. Here, the savings are substantial and amount to about 130%, which is almost ten times more than the percentage loss due to the prediction errors. However, the problem of the settlement structure for prediction errors becomes very apparent in the undersupply scenario. Here, the savings due to an LEM are more than offset by the loss due to prediction errors. Consequently, consumers would be better off not participating in an LEM.

Table 5. Average savings for consumers due to the LEM and average loss for consumers due to prediction errors in the LEM.  [BLEMevaluateMarketSim \(github.com/QuantLet/BLEM/tree/master/BLEMevaluateMarketSim\)](https://github.com/QuantLet/BLEM/tree/master/BLEMevaluateMarketSim)

Mean	Balanced Supply	Oversupply	Undersupply
Cost without LEM (in EUR)	459.83	459.83	459.83
Cost predicted values (in EUR)	457.94	226.61	470.69
Cost true values (in EUR)	439.26	200.75	451.60
Savings due to LEM (in %)	4.82	129.08	1.90
Loss due to pred. errors (in %)	−4.80	−13.75	−4.76

This result is visualized in a more differentiated way in Figure 10. The figure shows for each supply scenario, for each consumer, the total energy cost over the whole simulation period in: (1) no LEM; (2) an LEM with the use of predicted consumption values; and (3) an LEM with the use of true consumption values. For each supply scenario, the bottom panel shows the percentage loss due to not participating in the LEM and the loss due to participating and using predicted consumption values compared to participating and using true consumption values. In the balanced scenario, there are some consumers who would make a loss due to the participation in the LEM and relying on predicted values.

For them, the loss due to no LEM (yellow bar) is smaller than the loss due to prediction errors (green bar). However, 56 out of 88 consumer (i.e., 64%) also profit from the participation in the LEM despite the costs induced by prediction errors. Due to the much lower equilibrium prices in the oversupply scenario, the LEM participation here is, despite prediction errors, profitable for all consumers. However, even in this scenario, the savings for the consumers are diminished by more than 10%, which is quite substantial. In contrast, in the undersupply scenario, the loss due to the prediction errors leaves the participation in the LEM for almost all consumers unprofitable. Merely three consumers would profit and have lower costs in an LEM than without an LEM, despite prediction errors.

Overall, it becomes clear that prediction errors significantly lower the economic profitability of an LEM for consumers. This, however, is often argued to be one of the main advantages of LEMs. The result is especially concerning in LEMs where locally produced energy is undersupplied. Here—still assuming the closed double auction market mechanism and zero-intelligence bidding strategies—the savings from the participation in the LEM are marginal. Therefore, the costs induced by prediction errors mostly outweigh the savings from the participation. This results in an overall loss for consumers due to the LEM, which makes the participation economically irrational. Only in cases of substantial oversupply, the much lower equilibrium price, compared to the energy utility's price, compensates for the costs from prediction errors.

In conclusion, this means that LEMs with a discrete interval, closed double auction as market mechanism and a prediction error settlement structure as proposed in [6] combined with the prediction accuracy of state-of-the-art energy forecasting techniques require substantial oversupply in the LEM for it to be beneficial to consumers.



Figure 10. Total energy cost to consumers from 01 October 2018 to 31 December 2017 in the case of no LEM, LEM with true values, and LEM with predicted values in three different supply scenarios.

 [BLEMevaluateMarketSim \(github.com/QuantLet/BLEM/tree/master/BLEMevaluateMarketSim\)](https://github.com/QuantLet/BLEM/tree/master/BLEMevaluateMarketSim)

4.3. Implications for Blockchain-Based Local Energy Markets

In light of these results, it remains open to derive implications and to propose potential adjustments for an LEM market mechanism. After all, there are substantial advantages of LEMs which have been established in various studies and still make LEMs an attractive solution for the challenges brought about by the current energy transition. Adjustments mitigating the negative effect of prediction errors on the profitability of LEMs could address one or more of the following areas: first, the forecasting techniques employed; second, the demand and supply structure of the LEM; and, third, the market mechanism used in the blockchain-based LEM.

The first and most intuitive option is to improve the forecasting accuracy with which the predictions, which serve as the basis of bids and asks, are made. For example, a common approach to reduce the bias of LASSO-based predictions are post-LASSO techniques such as presented by Chernozhukov et al. [49]. Another aspect that seems relevant for the improvement of forecasting models is the evaluation method. Using economic measures for the evaluation of forecasting model performance may address a potential mismatch between statistical measures of forecasting accuracy and the resulting economic profits [50]. However, these approaches most likely result only in small improvements. Thus, the most obvious way to achieve a substantial improvement is the inclusion of more data. More data may hereby refer either to a higher resolution of recorded energy data or to a wider range of data sources such as behavioral data of household members or data from smart appliances. A higher resolution of smart meter readings is already easily achievable. The smart meters installed by Discovery that also supplied the data for the present research are capable of recording energy measurements up to every two seconds. However, data at such a fine granularity requires substantial data storage and processing capacities which are unlikely to be available in an average household. Especially, the training of prediction models with such vast amounts of input data points is computationally very resource intensive. The potential solution of outsourcing this, however, introduces new data privacy concerns that are already a sensible topic in smart meter usage and blockchain-based LEMs (e.g., [8,51]). Increasing the forecasting intervals to 30 or 60 min, as an alternative way to reduce the computational resources needed, would presumably decrease the forecasting accuracy which, in turn, might increase the cost for consumers. However, the effect of this potential solution on the cost for consumers due to forecasting errors seems reasonable to be investigated in future studies. The inclusion of behavioral data into prediction models such as the location of the person within their house and the inclusion of smart appliances' energy consumption (as done by Kong et al. [22]) and running schedules raises important privacy concerns as well. Pooling and using energy consumption data of several households, as done by Shi et al. [23], again introduces privacy concerns as it implies data sharing between households, which in relatively small LEMs cannot be guaranteed to preserve the anonymity of market participants. For all these reasons, it seems unlikely that in the near future qualitative jumps in the prediction accuracy of very short-term household energy consumption or production of individual households will be available.

The second option addresses the demand and supply structure in the blockchain-based LEM. As shown in Section 4.2, the cost induced by prediction errors and their settlement is more than compensated in an oversupply scenario. Hence, employing LEMs only in a neighborhood in which energy production surpasses energy consumption would mitigate the problem of unprofitability due to prediction errors as well. Where this is not possible, participation to the LEM could be restricted, such that oversupply in a majority of trading periods is ensured. However, this might end up in a market manipulation that most likely makes most of LEMs' advantages obsolete. Moreover, it is unclear on what basis the restriction to participate in the market should be grounded.

The third option to mitigate the problem is the market mechanism and the prediction error settlement structure. A simple approach to reduce forecasting errors is to decrease the forecasting horizon. Thus, instead of having 15-min trading periods, which also require 15-min ahead forecast, the trading periods could be shrunk to just 3 min. This would increase the forecasting accuracy, and, thereby, lead to lower costs due to the settlement of prediction errors. However, in a blockchain-based LEM, more frequent market closings come at the cost of more computational resources needed for transaction verification and cryptographic block generation. Depending on the consensus mechanism used for the blockchain, the energy requirements for the computations, which secure transactions and generate new blocks, may be substantial. This, of course, is rather detrimental to the idea of promoting more sustainable energy generation and usage. Nevertheless, using consensus mechanisms based on identity verification of the participating agents may serve as a less computational, and thus energy intensive alternative, which might make shorter trading intervals reasonable. Another, more radical, approach might be to change the market mechanism of closed

double auctions altogether and use an exposed market instead. Hereby, the energy consumption and production is settled in an auction after the true values are known, instead of in advance. This means that market participants submit just limit prices in their bids and asks without related amounts and the offers are matched in an auction in regular time intervals. Then, the electricity actually consumed and produced in the preceding period is settled according to the market clearing price. Related to this approach is a solution, where bidding is based on forecasted energy values, while the settlement is shifted by one period such that the actual amounts can be used for clearing. This approach, however, may introduce the possibility of fraud and market manipulation as agents can try to deliberately bid using false amounts. While in the smart contract developed by Mengelkamp et al. [6] funds needed to back up the bid are held as pledges until the contract is settled (this ensures the availability of the necessary funds to pay the bid), this would be senseless, if settlement is only based on actual consumption without considering the amount specified in the offer. However, the extent of this problem and ways to mitigate it should be assessed from a game theoretical perspective that is out of scope of the present research.

Overall, prediction errors have to be taken into account for future designs of blockchain-based LEMs. Otherwise, they may substantially lower the profitability and diminish the incentive to participate in an LEM for consumers. In addition, the psychological component of having to rely on an unreliable prediction algorithm that may be more or less accurate depending on the household's energy consumption patterns seems unattractive. Even though possible solutions are not trivial and each comes with certain trade-offs, there is room for future improvement of the smart contracts and the market mechanism they reproduce.

5. Conclusions

The present research had the objectives: (1) to evaluate the prediction accuracy achievable for household energy consumption with state-of-the-art forecasting techniques; (2) to assess the effect of prediction errors on an LEM that uses a closed double auction with discrete time intervals as market mechanism; and (3) to infer implications based on the results for the future design of blockchain-based LEMs.

In the performance assessment of currently used forecasting techniques, the LASSO model yielded the best results with an average MAPE across all consumer datasets of 17%. It was subsequently used to make predictions for the market simulation. The evaluation of the market mechanism and prediction error settlement structure revealed that, in a balanced supply and demand scenario, the costs of prediction errors almost completely offset savings brought by the participation in the LEM. In an undersupply scenario, the cost due to prediction errors even surpassed the savings and made market participation uneconomical. The most promising approach to mitigate this problem seemed to be adjustment of the market design, which can be two-fold: either shorter trading periods could be introduced, which would reduce the forecasting horizon, and therefore prediction errors, or the auction mechanism could be altered to not use predicted consumption values to settle transactions.

For the present research, data from a greater number of smart meters and more context information about the data would have been desirable. However, due to data protection legislation, no information regarding locality of the households, household characteristics or the type of power plant prosumer households used could be provided. Thus, unfortunately, no other covariates (e.g., temperature) could be used in the forecasting of energy consumption. In addition, the large-scale differences in the production capacities of the prosumers, contained in the data, complicated the analysis of the market simulation further. Additionally, it is worth mentioning that the market simulation did not account for taxes or fees, especially grid utilization fees, which can be a substantial share of the total electricity cost of households. The simulation also did not take into account compensation costs for blockchain miners that reimburses them for the computational cost they bear.

Evidently, future research concerned with blockchain-based LEMs should take into account the potential cost of prediction errors. Furthermore, to our knowledge, there has been no simulation

of a blockchain-based LEM with actual consumption and production data conducted. Doing so on a private blockchain with the market mechanism coded in a smart contract should be the next step for the assessment of potential technological and conceptual weaknesses.

In conclusion, previous research has shown that blockchain technology and smart contracts combined with renewable energy production can play an important role in tackling the challenges of climate change. The present research, however, emphasizes that advancement on this front cannot be made without a holistic approach that takes all components of blockchain-based LEMs into account. Simply assuming that reasonably accurate energy forecasts for individual households will be available once the technical challenges of implementing an LEM on a blockchain are solved, may steer research into a wrong direction and bears the risk of missing the opportunity to quickly move into the direction of a more sustainable and less carbon-intensive future.

Author Contributions: Conceptualization, M.K. and W.K.H.; Data curation, M.K.; Formal analysis, W.K.H.; Methodology, M.K.; Software, M.K.; Supervision, W.K.H.; Validation, M.K. and W.K.H.; Visualization, M.K.; Writing—original draft, M.K.; and Writing—review and editing, M.K. and W.K.H.

Funding: This research received no external funding.

Acknowledgments: We would like to thank Discovery GmbH for the kind provision of their smart meter data, the Humboldt Lab for Empirical and Quantitative Research (LEQR) at the School of Business and Economics, Humboldt-University Berlin for the kind provision of computing resources, and the IRTG 1792 at the School of Business and Economics, Humboldt University of Berlin for valuable support.

Conflicts of Interest: The authors declare no conflict of interest.

Data Availability: All data and algorithms are freely available through www.quantlet.de with the keyword BLEM and at GitHub: github.com/QuantLet/BLEM.

Abbreviations

The following abbreviations are used in this manuscript:

LEM	Local energy market
LASSO	Least absolute shrinkage and selection operator
RNN	Recurrent neural network
LSTM	Long short-term memory
ML	Machine learning
GPU	Graphical processing unit
CPU	Central processing unit
CV	Cross-validation
SD	Standard deviation
MAE	Mean absolute error
RMSE	Root mean square error
MAPE	Mean absolute percentage error
NRMSE	Normalized root mean square error
MASE	Mean absolute scaled error

References

1. Sinn, H.W. Buffering volatility: A study on the limits of Germany's energy revolution. *Eur. Econ. Rev.* **2017**, *99*, 130–150. [[CrossRef](#)]
2. Bayer, B.; Matschoss, P.; Thomas, H.; Marian, A. The German experience with integrating photovoltaic systems into the low-voltage grids. *Renew. Energy* **2018**, *119*, 129–141. [[CrossRef](#)]
3. BSW-Solar. *Statistische Zahlen der deutschen Solarstrombranche (Photovoltaik)*; Bundesverband Solarwirtschaft e.V.: Berlin, Germany, 2018.
4. Weron, R. *Modeling and forecasting electricity loads and prices: A statistical approach*; John Wiley & Sons: Chichester, UK, 2006.
5. Rutkin, A. Blockchain-Based Microgrid Gives Power to Consumers in New York. Available online: news scientist.com/article/2079334-blockchain-based-microgrid-gives-power-to-consumers-in-new-york/ (accessed on 13 July 2019).

6. Mengelkamp, E.; Gärttner, J.; Rock, K.; Kessler, S.; Orsini, L.; Weinhardt, C. Designing microgrid energy markets—A case study: The Brooklyn Microgrid. *Appl. Energy* **2018**, *210*, 870–880. [[CrossRef](#)]
7. Lamparter, S.; Becher, S.; Fischer, J.G. An Agent-based Market Platform for Smart Grids. In Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS): Industry Track, Toronto, ON, Canada, 10–14 May 2010; pp. 1689–1696.
8. Buchmann, E.; Kessler, S.; Jochem, P.; Böhm, K. The Costs of Privacy in Local Energy Markets. In Proceedings of the 2013 IEEE 15th Conference on Business Informatics, Vienna, Austria, 15–18 July 2013; pp. 198–207.
9. Block, C.; Neumann, D.; Weinhardt, C. A Market Mechanism for Energy Allocation in Micro-CHP Grids. In Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008), Waikoloa, HI, USA, 7–10 January 2008; pp. 172–183.
10. Mengelkamp, E.; Notheisen, B.; Beer, C.; Dauer, D.; Weinhardt, C. A blockchain-based smart grid: Towards sustainable local energy markets. *Comput. Sci. Res. Dev.* **2018**, *33*, 207–214. [[CrossRef](#)]
11. Stadler, M.; Cardoso, G.; Mashayekh, S.; Forget, T.; DeForest, N.; Agarwal, A.; Schönbein, A. Value streams in microgrids: A literature review. *Appl. Energy* **2016**, *162*, 980–989. [[CrossRef](#)]
12. Mengelkamp, E.; Gärttner, J.; Weinhardt, C. Intelligent Agent Strategies for Residential Customers in Local Electricity Markets. In Proceedings of the Ninth International Conference on Future Energy Systems (e-Energy '18), Karlsruhe, Germany, 12–15 June 2018; pp. 97–107.
13. Koirala, B.P.; Koliou, E.; Friege, J.; Hakvoort, R.A.; Herder, P.M. Energetic communities for community energy: A review of key issues and trends shaping integrated community energy systems. *Renew. Sustain. Energy Rev.* **2016**, *56*, 722–744. [[CrossRef](#)]
14. Hvelplund, F. Renewable energy and the need for local energy markets. *Energy* **2006**, *31*, 2293–2302. [[CrossRef](#)]
15. Ilic, D.; Silva, P.G.D.; Karnouskos, S.; Griesemer, M. An energy market for trading electricity in smart grid neighbourhoods. In Proceedings of the 2012 6th IEEE International Conference on Digital Ecosystems and Technologies (DEST), Campione d'Italia, Italy, 18–20 June 2012; pp. 1–6.
16. Rosen, C.; Madlener, R. An auction design for local reserve energy markets. *Decis. Support System.* **2013**, *56*, 168–179. [[CrossRef](#)]
17. Mengelkamp, E.; Gärttner, J.; Weinhardt, C. Decentralizing Energy Systems Through Local Energy Markets: The LAMP-Project. In Proceedings of the Multikonferenz Wirtschaftsinformatik (MKWI), Lüneburg, Germany, 6–9 March 2018; pp. 924–930.
18. Wang, Y.; Chen, Q.; Hong, T.; Kang, C. Review of smart meter data analytics: Applications, methodologies, and challenges. *IEEE Trans. Smart Grid* **2018**, *10*, 1–24. [[CrossRef](#)]
19. Burger, C.; Kuhlmann, A.; Richard, P.; Weinmann, J. *Blockchain in the Energy Transition. A Survey among Decision-Makers in the German Energy Industry*; Report; ESMT Berlin, Berlin, Germany, 2016.
20. Münsing, E.; Mather, J.; Moura, S. Blockchains for decentralized optimization of energy resources in microgrid networks. In Proceedings of the 2017 IEEE Conference on Control Technology and Applications (CCTA), Mauna Lani, HI, USA, 27–30 August 2017; pp. 2164–2171.
21. Arora, S.; Taylor, J.W. Forecasting electricity smart meter data using conditional kernel density estimation. *Omega* **2016**, *59*, 47–59. [[CrossRef](#)]
22. Kong, W.; Dong, Z.Y.; Jia, Y.; Hill, D.J.; Xu, Y.; Zhang, Y. Short-Term Residential Load Forecasting Based on Resident Behaviour Learning. *IEEE Trans. Power Syst.* **2018**, *33*, 1087–1088. [[CrossRef](#)]
23. Shi, H.; Xu, M.; Li, R. Deep learning for household load forecasting—A novel pooling deep RNN. *IEEE Trans. Smart Grid* **2018**, *9*, 5271–5280. [[CrossRef](#)]
24. Li, P.; Zhang, B.; Weng, Y.; Rajagopal, R. A sparse linear model and significance test for individual consumption prediction. *IEEE Trans. Power Syst.* **2017**, *32*, 4489–4500. [[CrossRef](#)]
25. Bansal, A.; Rompikuntla, S.K.; Gopinadhan, J.; Kaur, A.; Kazi, Z.A. Energy Consumption Forecasting for Smart Meters. In Proceedings of the third International Conference on Business Analytics and Intelligence (BAI) 2015, Bangalore, India, 17–19 December 2015; pp. 1–20.
26. Diagne, M.; David, M.; Lauret, P.; Boland, J.; Schmutz, N. Review of solar irradiance forecasting methods and a proposition for small-scale insular grids. *Renew. Sustain. Energy Rev.* **2013**, *27*, 65–76. [[CrossRef](#)]
27. Gan, D.; Wang, Y.; Zhang, N.; Zhu, W. Enhancing short-term probabilistic residential load forecasting with quantile long-short-term memory. *J. Eng.* **2017**, *2017*, 2622–2627. [[CrossRef](#)]
28. Chollet, F.; Allaire, J. *Deep Learning with R*; Manning Publications Co.: Shelter Island, NY, USA, 2018.

29. Van der Meer, D.W.; Widén, J.; Munkhammar, J. Review on probabilistic forecasting of photovoltaic power production and electricity consumption. *Renew. Sustain. Energy Rev.* **2018**, *81*, 1484–1512. [[CrossRef](#)]
30. Chen, K.; Chen, K.; Wang, Q.; He, Z.; Hu, J.; He, J. Short-term Load Forecasting with Deep Residual Networks. *IEEE Trans. Smart Grid* **2018**, 1–10. [[CrossRef](#)]
31. Hornik, K.; Stinchcombe, M.; White, H. Multilayer feedforward networks are universal approximators. *Neural Netw.* **1989**, *2*, 359–366. [[CrossRef](#)]
32. Bengio, Y.; Simard, P.; Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* **1994**, *5*, 157–166. [[CrossRef](#)]
33. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)]
34. Gers, F.A.; Schmidhuber, J.; Cummins, F. Learning to Forget: Continual Prediction with LSTM. *Neural Comput.* **2000**, *12*, 2451–2471. [[CrossRef](#)]
35. Lipton, Z.C.; Berkowitz, J.; Elkan, C. A Critical Review of Recurrent Neural Networks for Sequence Learning. *arXiv* **2015**, arXiv:1506.00019v4.
36. Graves, A., Supervised Sequence Labelling. In *Supervised Sequence Labelling with Recurrent Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 5–13.
37. Chollet, F.; Allaire, J. R Interface to Keras. 2017. Available online: <https://github.com/rstudio/keras> (accessed on 30 September 2018).
38. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Statist. Soc. Ser. B (Methodol.)* **1996**, *58*, 267–288. [[CrossRef](#)]
39. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **2010**, *33*, 1–22. [[CrossRef](#)]
40. Hoff, T.; Perez, R.; Kleissl, J.; Renne, D.; Stein, J. Reporting of irradiance modeling relative prediction errors. *Prog. Photovolt. Res. Appl.* **2013**, *21*, 1514–1519. [[CrossRef](#)]
41. Zhang, J.; Florita, A.; Hodge, B.M.; Lu, S.; Hamann, H.F.; Banunarayanan, V.; Brockway, A.M. A suite of metrics for assessing the performance of solar power forecasting. *Sol. Energy* **2015**, *111*, 157–175. [[CrossRef](#)]
42. Hyndman, R.J.; Koehler, A.B. Another look at measures of forecast accuracy. *Int. J. Forecast.* **2006**, *22*, 679–688. [[CrossRef](#)]
43. Heidjann, J. *Strompreise in Deutschland - Vergleichende Analyse der Strompreise für 1437 Städte in Deutschland*; StromAuskunft - Alles über Strom, Münster, Germany, 2017.
44. Gode, D.K.; Sunder, S. Allocative Efficiency of Markets with Zero-Intelligence Traders: Market as a Partial Substitute for Individual Rationality. *J. Political Econ.* **1993**, *101*, 119–137. [[CrossRef](#)]
45. Discovery GmbH. *Intelligente Stromzähler und Messsysteme*; Discovery GmbH: Heidelberg, Germany, 2018.
46. Auder, B.; Cugliari, J.; Goude, Y.; Poggi, J.M. Scalable Clustering of Individual Electrical Curves for Profiling and Bottom-Up Forecasting. *Energies* **2018**, *11*, 1893. [[CrossRef](#)]
47. Gerossier, A.; Girard, R.; Kariniotakis, G.; Michiorri, A. Probabilistic day-ahead forecasting of household electricity demand. *CIREN - Open Access Proc. J.* **2017**, *2017*, 2500–2504. [[CrossRef](#)]
48. Teixeira, B.; Silva, F.; Pinto, T.; Santos, G.; Praça, I.; Vale, Z. TOOCC: Enabling heterogeneous systems interoperability in the study of energy systems. In Proceedings of the 2017 IEEE Power Energy Society General Meeting, Chicago, IL, USA, 16–20 July 2017; pp. 1–5.
49. Chernozhukov, V.; Härdle, W.K.; Huang, C.; Wang, W. LASSO-Driven Inference in Time and Space. *arXiv* **2018**, arXiv:1806.05081v3.
50. Maciejowska, K.; Nitka, W.; Weron, T. Day-Ahead vs. Intraday—Forecasting the Price Spread to Maximize Economic Benefits. *Energies* **2019**, *12*, 631. [[CrossRef](#)]
51. Greveler, U.; Justus, B.; Loehr, D. Forensic content detection through power consumption. In Proceedings of the 2012 IEEE International Conference on Communications (ICC), Ottawa, ON, Canada, 10–15 June 2012; pp. 6759–6763.





ELSEVIER

Contents lists available at ScienceDirect

Energy Economics

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

Regularization approach for network modeling of German power derivative market[☆]



Shi Chen^{a,*}, Wolfgang Karl Härdle^{b,c,d}, Brenda López Cabrera^b

^aChair of Statistics and Econometrics, Karlsruher Institut für Technologie, Blücherstr.17, Karlsruhe 76185, Germany

^bSchool of Business and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, Berlin 10099, Germany

^cDepartment of Mathematics and Physics, Charles University Prague, Ke Karlovu 2027, 12116 Praha 2, Czech

^dWISE, Xiamen University, 422 Siming Road, Xiamen 361005, China

ARTICLE INFO

Article history:

Received 9 August 2018

Received in revised form 15 May 2019

Accepted 28 June 2019

Available online 4 July 2019

JEL classification:

C1

Q41

Q47

Keywords:

Regularization

Energy risk transmission

Connectedness

Network

German power derivative market

ABSTRACT

In this paper we propose a regularization approach for network modeling of German power derivative market. To deal with the large portfolio, we combine high-dimensional variable selection techniques with dynamic network analysis. The estimated sparse interconnectedness of the full German power derivative market, clearly identify the significant channels of relevant potential risk spillovers. Our empirical findings show the importance of interdependence between different contract types, and identify the main risk contributors. We further observe strong pairwise interconnections between the neighboring contracts especially for the spot contracts trading in the peak hours, its implications for regulators and investors are also discussed. The network analysis of the full German power derivative market helps us to complement a full picture of system risk, and have a better understanding of the German power market functioning and environment.

© 2019 Elsevier B.V. All rights reserved.

1. Introduction

Affordable and reliable energy supply is essential for industrial growth. Achieving these in times of growing demand, raw materials shortage and climate change pose challenges. Germany's power system for the industry and the consumers is undergoing radical change; this transformation is being driven by the restructuring of electricity supply and by intense competition between suppliers (see BMWi (2016), Spiecker et al. (2014), Seifert et al. (2016), Grossi et al. (2017), Sinn (2017) among others). However, the ongoing expansion of renewable energy and the phase-out of nuclear energy for power generation will change the composition of the electricity mix, which in return, will generate pricing signals affecting the electricity trading (e.g. Benhmad and Percebois (2018), Ketterer (2014), Ballester and Furió (2015), Paraschiv et al. (2014)). As we know, electricity is the commodity that should be supplied immediately. Unlike coal,

oil, gas or other typical commodities, electricity cannot be stored. This results in the price of electricity being volatile and very dependent on a secure supply. To hedge against the uncertainty that has arisen in the market, we study the system-wide market risk of the whole German power derivative market. Therefore energy companies may invest in both electricity spot and derivative markets to diversify their existing portfolios. As electricity grids worldwide also begin relying more heavily on renewable energy sources, analysis of German power market thus provides useful insights for power generation companies and transmission organizations across the globe.

However, the number of variables and relevant factors is typically huge. A properly designed subset selection has to be employed to identify the most informative power contracts representing energy market risk. The German power market is highly interconnected with a dense and wide range of electricity contracts, this motivates us to build up an ultra high-dimensional network and investigate its sparse property. To better understand the interaction between power contracts, the iterated sure independence screening (iterated-SIS, Fan et al., 2009) method combined with regularization estimators are applied to estimate the sparse web of connections. Our network of interest is constructed in the context of time series based

[☆] The financial support from the Deutsche Forschungsgemeinschaft via IRTG 1792 "High Dimensional Non Stationary Time Series" and SCHI - 1127/2 is gratefully acknowledged.

* Corresponding author.

E-mail address: shi.chen@kit.edu (S. Chen).

on vector autoregression (VAR) models, the iterated-SIS method is important when building VAR models since the number of parameters to estimate increases quadratically in the number of variables included. To quantify the associations between individual power contract and energy exchange market, the network we constructed is obtained from the forecast error variance decomposition (FEVD) based on VAR estimates in the framework of Koop et al. (1996) and Pesaran and Shin (1998). This kind of connectedness measure is also used by Diebold and Yilmaz (2009), Diebold and Yilmaz (2014) for conceptualizing and empirically measuring weighted, directed networks at a variety of levels. They proposed this variance decomposition networks as tools for quantifying and ranking the systemic risk of individual component in a portfolio.

In this paper we investigate the concept of connectedness in a realistic high-dimensional framework, which is important for system-wide risk measurement and management. We aim to obtain a sparse network in which nodes represent power contracts and links represent the magnitude of connectedness, local shocks and events can therefore be easily amplified and turned into global events. While estimates of the network yield the qualitative links between power contracts, individual impact from specific contract can be estimated and speculated accordingly. Hence the risk contribution from the market component can be identified, this will help us to learn more about the German power market functioning and environment. Following Diebold and Yilmaz (2014), Demirel et al. (2018), Hautsch et al. (2014) and similar connectedness literature, the risk refers to the uncertainty that has arisen in the system, and it measures the amount of future uncertainty caused by other component in the system. For example, the market uncertainty may be caused by economic and financial uncertainties, or weather conditions, energy prices and regulation. While the systemic risk network yields qualitative information on risk channels and roles of assets within the constructed portfolio, the risk is quantified by its contribution to the forecast error variance of other components in the whole system. Therefore we are able to capture the systemic risk of the component by summing up its total contribution. Understanding the risk transmission channels for investors is of great importance, for example, our results show that day-ahead spot power contracts bidding between 9 am and 13 am are in the core of the German energy power market, the key derivatives in connecting markets can be identified.

The rest of the paper proceeds as follows. Section 2 reviews the relevant literature and introduces the German energy market. In Section 3, we describe in details how the regularization approach is applied to estimate the large portfolio and how the network is constructed. Section 4 reports the data and discusses the model selection result. Section 5 presents a static analysis of full-sample connectedness. Section 6 provides the empirical results of the dynamic network. Finally Section 7 concludes.

2. Background

2.1. Related literature

As a tradable commodity electricity is relatively new, its dynamic properties have been analyzed with many different approaches, some recent contributions are Weron (2007), Geman and Roncoroni (2006), Bierbrauer et al. (2007) and Knittel and Roberts (2005). There is also a strand of literature that analyzes the multivariate behavior of electricity prices. For examples, Higgs (2009) examines the inter-relationships of wholesale spot electricity prices across four Australasian markets by a multivariate GARCH model. Henriques and Sadosky (2008) develop a four variable VAR model to explain the dependence structure of a variety of energy equities, where they find that shocks to technology actually have a larger impact on the stock prices of alternative energy companies than do oil prices.

Castagneto-Gissey et al. (2014) study the interactions of a representative sample of 13 European (EU) electricity spot prices with dynamic Granger-causal networks.

We take our starting point in the energy literature based on a vector regression framework, where the network of interest is constructed based on Diebold and Yilmaz (2009), Diebold and Yilmaz (2014). Most relevant studies explore the relationship between oil and energy equity prices in terms of volatility spillovers, they estimate the implied volatility linkages across markets as a source of future uncertainty. For example Du et al. (2011) conduct a Bayesian analysis to explain volatility spillovers among crude oil and various economic factors. Arouri et al. (2012) investigate the volatility spillovers between oil and stock markets in Europe using VARGARCH approach. Sadosky (2012) applies a multivariate GARCH model to estimate the volatility spillovers between oil prices and the stock prices of clean energy/technology companies. Joo and Park (2017) examine the time-varying causal relationship between the stock and crude oil price uncertainties using a DCC GARCH-in-Mean specification. More empirical work are Diebold and Yilmaz (2012), Reboredo (2014), Maghyereh et al. (2016), Awartani et al. (2016), Zhang (2017), Apergis et al. (2017), among others. There is also a fairly sizable literature exploring the electricity market integration using vector regression framework, for example Worthington et al. (2005), Zachmann (2008), Bunn and Gianfreda (2010), Balaguer (2011), Böckers and Heimeshoff (2014), Castagneto-Gissey et al. (2014).

However, relatively little research has focused on the systemic directional interaction between energy equities. A recent study by Lundgren et al. (2018) is the first to analyze the connectedness network among different energy asset classes, their analysis examines the connectedness network among renewable energy stock, four investments, and uncertainties. Demirel et al. (2018) use Lasso method to select, shrink and estimate a high-dimensional network. Other empirical work are with more focus on financial banking contexts, like Yi et al. (2018) use the VARX-L framework developed by Nicholson et al. (2017) to conduct static and dynamic volatility spillovers among cryptocurrencies. More relevant work are Wang et al. (2018), Acharya et al. (2012), Hautsch et al. (2014), Giglio et al. (2016), Babus (2016), Brownlees and Engle (2016), Acharya et al. (2017) among others.

However, almost all existing energy literature are based only on moderate dimensions. This motivates us to examine the systemic risk transmission channels in a realistic high-dimensional framework. The main argument is that the contracts trading in both spot and derivative markets share the same underlying contracts, and it is therefore natural to consider a high-dimensional portfolio. High-dimensional statistical problems arise from diverse fields of scientific research and technological development, including energy markets. The traditional idea of best subset selection methods is computationally too expensive for many modern statistical applications. Variable selection techniques have therefore been successfully developed in recent years and they indeed play a pivotal role in contemporary statistical learning and techniques. Researchers have proposed various regularized estimators with different penalty terms, a preminent example being the least absolute selection and shrinkage operator (Lasso) of Tibshirani (1996). In recent years, Lasso has been extended to high-dimensional case, see Bickel et al. (2009). Other popular methods contribute to the literature, such as smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), adaptive Lasso of Zou (2006), elastic net estimator of Zou and Hastie (2005), Dantzig selector of Candes and Tao (2007). In an ultra high-dimensional case where the dimensionality of the model is allowed to grow exponentially in the sample size, it is helpful to begin with screening to delete some significantly irrelevant variables from the model. Fan and Lv (2008) introduce a method called sure independence screening for this goal. Even when the regularity conditions may not hold, Fan et al. (2009) extend the iterated-SIS method to work by

iteratively performing feature selection to recruit a small number of features. Furthermore, the asymptotic properties of Lasso for high-dimensional time series have been considered by Loh and Wainwright (2011) and Wu et al. (2016). Kock and Callot (2015) establish the high-dimensional VAR estimation with focus on Lasso and adaptive Lasso. Basu et al. (2015) investigate the theoretical properties of regularized estimates in sparse high-dimensional time series models when the data are generated from a multivariate stationary Gaussian process.

We address the above high-dimensional problem by combining a regularization approach with classic VAR model. In doing so, we contribute to the energy literature in several ways: First, we extend the current literature to investigate the sparse linkage between energy equities in a very large portfolio. Second, we identify the main risk contributor to help investors diversify their existing portfolio rather than having large holdings of individual electricity contract. Third, our study may further be extended by including renewable energy assets and oil price across European energy market; this may provide a better understanding of the overall energy market. In addition, our study can also be applied to investigate electricity market integration by estimating the total connectedness for a wide range of components.

2.2. Overview of German power market

2.2.1. German electricity market

The German electricity market is Europe's largest, with annual power consumption of around 530 TWh and a generation capacity of 184 GW. As a net energy exporter, the export capacity of Germany is expected to continue to grow as planned interconnections expand cross-border transmission capacity with several neighboring countries. Germany has significant interconnection capacity with neighboring EU member states as well. It is interconnected with Austria, Switzerland, the Czech Republic, Denmark, France, Luxembourg, the Netherlands, Poland, and Sweden. To maintain stable and reliable supply of electricity, the so-called Transmission system operators (TSOs) keep control over available power. Primary control, secondary control, and tertiary control reserve are procured by the respective TSOs within a non-discriminatory control power market in accordance with the requirements of the Federal Cartel Office. Demand for control energy is created when the sum of power generated varies from the actual load caused by unforeseeable weather fluctuations in the case of renewable energies.

Electricity is traded on the exchange and over the counter. Standardized products are bought and sold in a transparent process on the exchange, which, for Germany, is the European Energy Exchange EEX in Leipzig, the European Energy Exchange EPEX SPOT in Paris and the Energy Exchange Austria (EXAA) in Vienna. The European Energy Exchange (EEX) is the leading energy exchange in Europe. It develops, operates and connects secure, liquid and transparent markets for energy and commodity products. Contracts on power, coal and CO₂ emission allowances as well as freight and agricultural products are traded or registered for clearing on EEX. EPEX SPOT, Powernext, Cleartrade Exchange (CLTX) and Gaspoint Nordic are also members of EEX Group. The German wholesale electricity market is broadly made up of three elements, a forward market, a day-ahead market and an intra-day market. These submarkets generate the pricing signal in which electricity production and consumption align to. The objective of this paper is to analyze the interaction of different future contracts traded in the forward market, whether forward market is influenced by market power of spot prices traded in EPEX market.

Electricity providers and electricity purchasers submit their bids in their national day-ahead market zones. The exchange price on the day-ahead market is determined jointly for coupled markets. Electricity providers and electricity purchasers submit their bids in

their national day-ahead market zones. In an iterative process, the demand for electricity in the market zone is served by the lowest price offers of electricity from all the market areas until the capacity of the connections between the market zones (cross-border inter-connectors) is fully utilized. As long as the cross-border inter-connectors have sufficient capacity, this process aligns the prices in the coupled market areas. On account of market coupling, the national power demand is covered by the international offers with lowest prices. The upshot is that on the whole less capacity is required to meet the demand. As shown in Fig. 2.1, Phelix Future, as the product traded in Germany, is a financial derivatives contract settling against the average power spot market prices of future delivery periods for the German/Austrian market area.

2.2.2. Phelix futures

Electricity supply deliveries in the forward market can be negotiated up to seven years in advance, but for liquidity reasons typically only look out three years, and in fact one year ahead futures are traded at most. The Phelix Future is a financial derivatives contract referring to the average power spot market prices of future delivery periods of the German/Austrian market area.

As the most liquid contract and benchmark for European power trading, the underlying of these future contracts is the Physical Electricity Index determined daily by EPEX Spot Exchange for base and peak load profiles. To be more specific, the Phelix Base contract is the average price of the hours 1 to 24 for electricity traded on spot market, while the Phelix Peak is the average price of the hours 9 to 20 for electricity traded on spot market. EEX offers continuous trading and trade registration of financially fulfilled Phelix Futures, with Day/Weekend Futures, Week Futures, Month Futures, Quarter Futures and Year Futures available.

The time series of Phelix day base and Phelix day peak prices are displayed in Fig. 2.2. Phelix day peak exhibits a larger volatility and more pronounced spikes than the Phelix day base. This is not surprising, since the Phelix day peak corresponds to hours with high and variable demand. Both price series exhibit positive skewness and an excess kurtosis of about 1, implying a heavy-tailed unconditional distribution that is skewed to the right.

In addition, the Phelix market is also successfully connected to other European power markets. The products of Location Spread enable members to trade price differences between markets, thus enabling participants to benefit from improved liquidity and tighter spreads, for instance, Phelix/French Power, Italian/Phelix Power, Phelix/Nordic Power and Phelix/Swiss Power. For the empirical work of this paper, we use the Phelix Future data to find price drivers and important variables in the big system we construct. The decision-making mechanism of energy companies will also be explored.

3. Econometric model

3.1. Basic model

When there is a high-dimensional portfolio consisting of various power derivative contracts, standard methods are intractable and therefore require novel statistical methods. Here we are interested in addressing the following research questions, how all these contracts interact with each other? Which variables are crucial for the whole system? However, due to the large number of variables in the system, some sparsity assumption must be imposed for the sake of an accurate estimate. The large dimensionality in our model comes from not only the varieties of power derivative products, but also the large lag order in VAR model to avoid the correlation of error terms.

The standard VAR(p) model with lag order p is constructed according to Lütkepohl (2005),

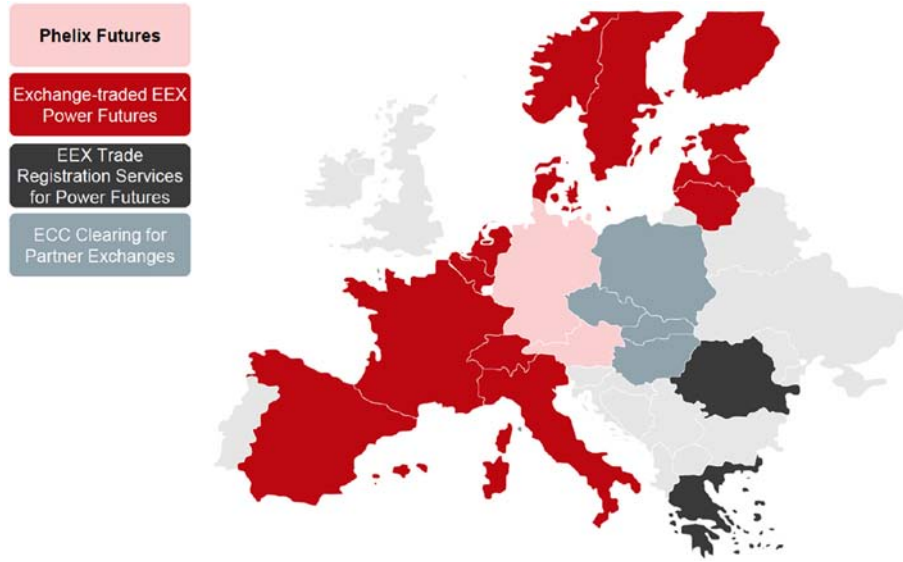


Fig. 2.1. The distribution of European power derivatives in EEX market. Source: EEX website.

$$y_t = \nu + A_1 y_{t-1} + A_2 y_{t-2} + \dots + A_p y_{t-p} + u_t$$

$$= \nu + (A_1, A_2, \dots, A_p) (y_{t-1}^T, y_{t-2}^T, \dots, y_{t-p}^T)^T + u_t \quad (3.1)$$

where $y_t = (y_{1t}, y_{2t}, \dots, y_{kt})^T$ is a $(K \times 1)$ random vector consisting of K variables at time t , $t = 1, \dots, T$. A_i are unknown $(K \times K)$ coefficient matrices. ν is a $(K \times 1)$ vector of intercept terms, $u_t = (u_{1t}, u_{2t}, \dots, u_{kt})^T$ is a K -dimensional innovation process.

Define

$$Y = (y_1, y_2, \dots, y_T)$$

$$B = (\nu, A_1, A_2, \dots, A_p)$$

$$Z_t = (1, y_t, y_{t-p+1})^T$$

$$Z = (Z_0, Z_1, \dots, Z_{T-1}) \quad (3.2)$$

For multivariate case, rewrite Eq. (3.1) as

$$Y = BZ + U \quad (3.3)$$

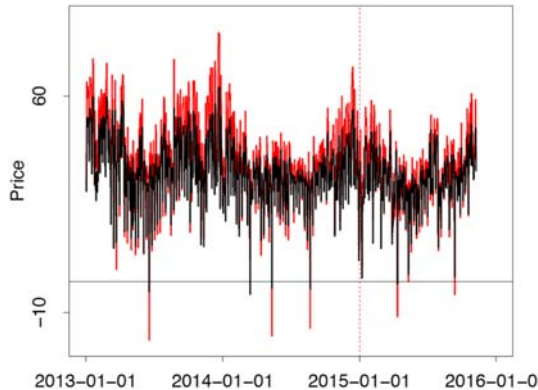


Fig. 2.2. Phelix day base (black) and Phelix day peak (red) index from 2013-01-01 to 2015-10-31. The red dotted line marks the end of the in-sample period.

where $U = (u_1, u_2, \dots, u_T)$. The compact form of Eq. (3.3) is

$$vec(Y) = (Z^T \otimes I_K) vec(B) + vec(U)$$

$$y = (Z^T \otimes I_K) \beta + u = X\beta + u \quad (3.4)$$

The dimension of model (3.4) to be estimated is pK^2 and the number of observations is KT . The ratio $\frac{Kp}{T}$ could be large due to the reasons mentioned earlier, which deteriorates the accuracy of final estimate. Worse still, if $Kp > T$, the model becomes high-dimensional with more unknown parameters than observations. Therefore we require techniques other than traditional OLS method. That's why most existing energy literature are based only on moderate dimensions.

Here we use variable selection technique, for example Lasso, to estimate the model. Besides, under normal assumption of error term, the upper bound of error in estimation is positively correlated in $\frac{\log(K^2p)}{T}$, part of oracle inequality. The estimation results can be further developed by adding one more step of sure independent screening (SIS) before variable selection step. Another advantage of SIS is that it could mitigate the problem caused by multicollinearity, which is common in time series setting. The techniques introduced in the proceeding paragraph are of great importance in the sense that the true underlying model has a sparse representation.

3.2. Regularization estimator and iterated-SIS algorithm

3.2.1. Regularization approach

Variable selection is an important tool for the linear regression analysis. A popular method is the Lasso estimator of Tibshirani (1996), which can be viewed to simultaneously perform model selection and parameter estimation. Related literature includes bridge regression studied by Frank and Friedman (1993) and Fu (1998), the least angle regression of Efron et al. (2004) and adaptive Lasso proposed by Zou (2006). Another remarkable example is a smoothly clipped absolute deviation (SCAD) penalty for variable selection proposed by Fan and Li (2001), they proved its oracle properties.

Let us start with considering model estimation and variable selection for Eq. (3.4),

$$y = X\beta + u \quad (3.5)$$

The least square estimate is obtained via minimizing $\|y - X\beta\|^2$, where the ordinary least squares (OLS) gives nonzero estimates $\omega = X^T y$ to all coefficients. Normally the best subset selection is implemented to select significant variables, but the traditional idea of best subset selection methods is computationally too expensive for many statistical applications. Therefore the penalized least square with a penalty term that is separable with respect to the estimated parameter $\hat{\beta}$ is considered here. In this paper we consider two popular estimators, Lasso and SCAD.

The Lasso is a regularization technique for simultaneous estimation and variable selection, with the estimator given by,

$$\begin{aligned} \hat{\beta}_{LASSO} &= \arg \min_{\beta} \|y - X\beta\|^2 + \lambda \sum_{j=1}^p |\beta_j| \\ &= \arg \min_{\beta} \|y - \sum_{j=1}^p x_j \beta_j\|^2 + \lambda \sum_{j=1}^p |\beta_j| \end{aligned} \tag{3.6}$$

where λ is the tuning parameter. The second term in Eq. (3.6) is known as the ℓ_1 -penalty. The idea behind Lasso is the coefficients shrink toward 0 as λ increases. When λ is sufficiently large, some of the estimated coefficients are exactly zero. The estimation accuracy comes from the trade-off between estimation variance and the bias. Lasso is the penalized least square estimates with the ℓ_1 penalty in the general least squares and likelihood settings. Furthermore, the ℓ_2 penalty results in a ridge regression and ℓ_p penalty will lead to a bridge regression.

We proceed to the SCAD method. In the present context, the SCAD estimator is given by,

$$\hat{\beta}_{SCAD} = \begin{cases} \text{sgn}(\omega)(|\omega| - \lambda)_+ & \text{when } |\omega| \leq 2\lambda \\ \frac{((a-1)\omega - \text{sgn}(\omega)a\lambda)}{a-2} & \text{when } 2\lambda < |\omega| \leq a\lambda \\ \omega & \text{when } |\omega| > a\lambda \end{cases} \tag{3.7}$$

where $a > 2$ is an additional tuning parameter. The continuous differentiable penalty function for SCAD estimator is defined by,

$$p'_\lambda(\beta) = \lambda \left\{ I(\beta \leq \lambda) + \frac{(a\lambda - \beta)_+}{(a-1)\lambda} I(\beta > \lambda) \right\} \quad \text{for } a > 2 \text{ and } \beta > 0 \tag{3.8}$$

To sum up, both estimators are members of this penalized likelihood family. Lasso has better performance when the noise to signal ratio is large, but this approach creates bias. SCAD can generate variable selection results without generating excess biases.

3.2.2. Iterated-SIS algorithm for estimation

SIS method is proposed by Fan and Lv (2008) to select important variables in ultra high-dimensional linear models. The proposed two-stage procedure can perform better than other methods in the sense of statistical learning problems. The SIS method based on the concept of sure screening, is defined as the correlation learning which filters out the features that have weak correlation with the

response. By sure screening, all the important variables survive after variable screening with probability tending to 1. Fan et al. (2009) improve iterated-SIS to a general pseudo-likelihood framework by allowing feature deletion in the iterative process. Fan et al. (2010) further extend the SIS model and consider an independent learning by ranking the maximum marginal likelihood estimator or maximum marginal likelihood itself for generalized linear models. Here we combine the VAR(p) model and SIS algorithm to find out the key elements in a big system. The basic idea of SIS is introduced in the following.

Let $\omega = (\omega_1, \omega_2, \dots, \omega_p)^T$ be a p -vector that is obtained by component-wise regression, i.e.,

$$\omega = X^T y \tag{3.9}$$

where y is n vector of response and X is a $n \times p$ data matrix. ω is a vector of marginal correlations of predictors with the response of predictors with the response variable, rescaled by the standard deviation of the response.

When there are more predictors than observation, LS (least square) estimator is noisy, that's why ridge regression is considered. Let $\omega^\lambda = (\omega_1^\lambda, \dots, \omega_p^\lambda)^T$ be a p -vector obtained by ridge regression, i.e.,

$$\omega^\lambda = (X^T X + \lambda I_p)^{-1} X^T y \tag{3.10}$$

where $\lambda > 0$ is a regularization parameter. Obviously, when $\lambda \rightarrow 0$, $\omega^\lambda \rightarrow \hat{\beta}_{LS}$ and $\lambda \rightarrow \infty$, $\lambda \omega^\lambda \rightarrow \omega$. The component-wise regression is a specific case of ridge regression with $\lambda = \infty$.

The iterated-SIS algorithm applied for estimating the VAR(p) model is as follows:

1. Apply SIS for initial screening, reduce the dimensionality to a relative large scale d ;
2. Apply a lower dimensional model selection method (such as Lasso, SCAD) to the sets of variables selected by SIS;
3. Apply SIS to the variables selected in the previous step;
4. Repeat steps 2 and 3 until the set of selected variables do not decrease.

3.3. Connectedness measure

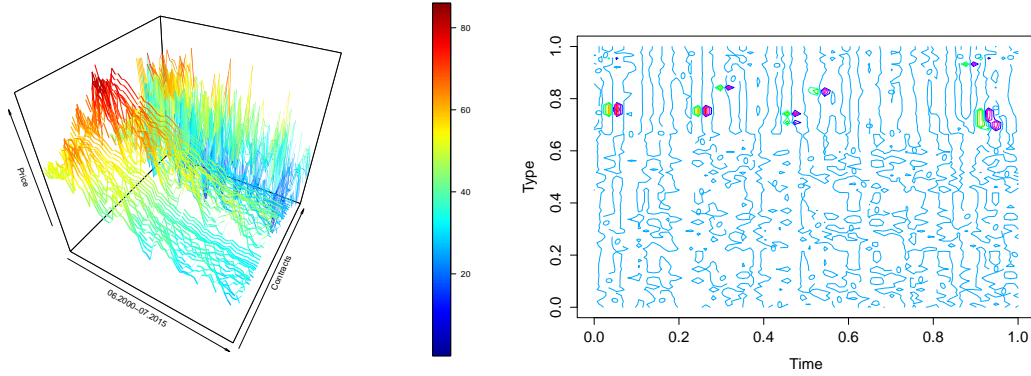
We construct our network using the fashionable directional connectedness measure proposed by Diebold and Yilmaz (2014). The connectedness is measured by cross-sectional variance decomposition, where the forecast error variance of variable is decomposed into parts attributed to the various variables in the system.

The interactions between the variables, i.e., the directional connectedness measure $\theta_{ij}(q)$ is given by,

$$\theta_{ij}(q) = \frac{\sigma_{jj}^{-1} \sum_{q=0}^{Q-1} (e_i^T \hat{B}_q \Sigma e_j)^2}{\sum_{q=0}^{Q-1} (e_i^T \hat{B}_q \Sigma \hat{B}_q^T e_i)} \tag{3.11}$$

Table 3.1
Connectedness table of interest.

	x_1	x_2	...	x_n	From others
x_1	$\theta_{11}(q)$	$\theta_{12}(q)$...	$\theta_{1n}(q)$	$\sum_{j=1}^n \theta_{1j}(q), j \neq 1$
x_2	$\theta_{21}(q)$	$\theta_{22}(q)$...	$\theta_{2n}(q)$	$\sum_{j=1}^n \theta_{2j}(q), j \neq 2$
\vdots	\vdots	\vdots		\vdots	\vdots
x_n	$\theta_{n1}(q)$	$\theta_{n2}(q)$...	$\theta_{nn}(q)$	$\sum_{j=1}^n \theta_{nj}(q), j \neq n$
To others	$\sum_{i=1}^n \theta_{i1}(q), i \neq 1$	$\sum_{i=1}^n \theta_{i2}(q), i \neq 2$...	$\sum_{i=1}^n \theta_{in}(q), i \neq n$	$\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \theta_{ij}(q), i \neq j$



(a) Ribbon plot of prices over 90 contracts

(b) Contour plot of log return

Fig. 4.1. Overview of dataset.

where q is the lag order, e_t is a $pK^2 \times 1$ selection vector with unity as its i -th element and zeros elsewhere. $\Sigma = E(u_t u_t^T)$, is the covariance matrix of the non-orthogonalized VAR(p) in Eq. (3.1), with σ_{ij} the corresponding j -th diagonal element of Σ . \hat{B}_i are the coefficient matrices of Eq. (3.12).

With iterated-SIS algorithm to estimate the sparse VAR structure, we can acquire its moving average (MA) transformation,

$$y_t = \sum_{i=0}^{\infty} B_i u_{t-i} \tag{3.12}$$

where the coefficient matrices B_i obey $B_i = \sum_{j=1}^{iy} B_{i-j} A_j$, with $B_0 = I_K$ and $A_j = 0$ for $j > p$. $A_j, j = 1, 2, \dots, p$ is the coefficient matrix of VAR(p) model.

To measure the persistent effect of a shock on the behavior of a series, we aim to acquire the population connectedness Table 3.1, according to Diebold and Yilmaz (2014).

The rightmost column gives the “from” effect of total connectedness, and the bottom row gives the “to” effect. In particular, the directional connectedness “from” and “to” associated with the forecast error variation θ_{ij} for specific power contract when the arising shocks transmit from one asset to the other. These two connectedness

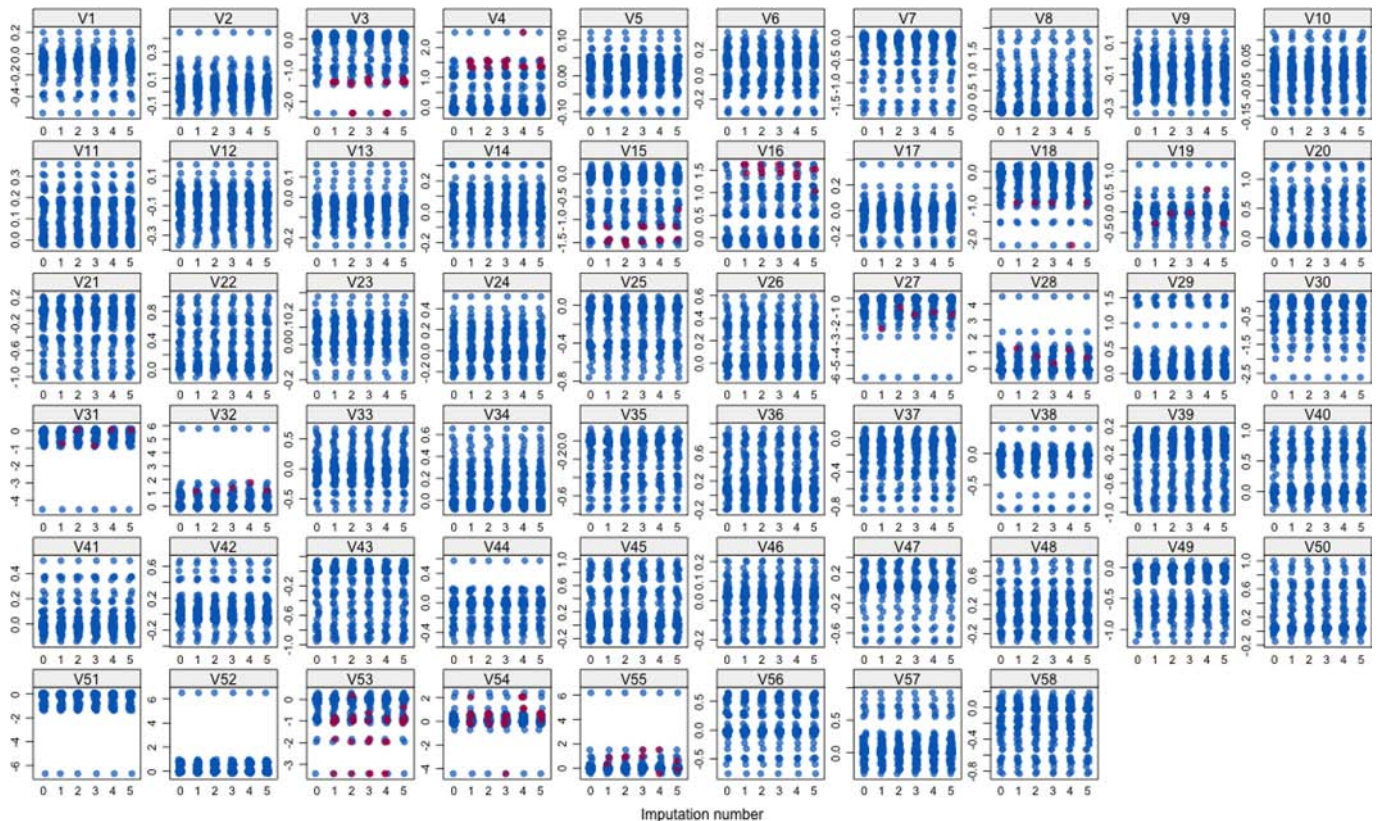


Fig. 4.2. Pattern of imputed data.

Table 4.1
Model selection results according to the AIC, BIC and HQ criteria.

Model	AIC	HQ	BIC
Iterated-SIS-Lasso, $p = 1$	4.5686	4.7249	5.7864
Iterated-SIS-Lasso, $p = 2$	4.5006	4.6426	5.6076
Iterated-SIS-Lasso, $p = 3$	6.9854	7.2315	10.0345
Iterated-SIS-SCAD, $p = 1$	4.5714	4.7277	5.7892
Iterated-SIS-SCAD, $p = 2$	6.1043	6.1043	9.5782
Iterated-SIS-SCAD, $p = 3$	7.2559	7.6820	10.5770

estimators can be obtained by adding up the row or column elements, the pairwise directional connectedness from j to i is given by,

$$C_{i \leftarrow j} = \theta_{ij}(q) \tag{3.13}$$

The total directional connectedness “from” $C_{i \leftarrow \cdot}$ (others to i), “to” $C_{\cdot \leftarrow j}$ (j to others) and the corresponding net connectedness are defined as

$$\begin{aligned}
 C_{i \leftarrow \cdot} &= \sum_{j=1}^n \theta_{ij}(q), i \neq j \\
 C_{\cdot \leftarrow j} &= \sum_{i=1}^n \theta_{ij}(q), i \neq j \\
 C_i &= C_{to} - C_{from} = C_{\cdot \leftarrow i} - C_{i \leftarrow \cdot}
 \end{aligned}
 \tag{3.14}$$

The “to” connectedness measures its total contribution to the forecast error variance of other components in the system, and therefore captures the systemic risk of the component.

4. Data and model selection

4.1. Data sets

As introduced in Section 2.2, EEX offers continuous trading data of Phelix Futures. The available load profiles are base, peak

and off-peak. The available products with different maturities have five kinds: Day/Weekend Futures, Week Futures, Month Futures, Quarter Futures and Year Futures. Nevertheless the products of Day/Weekend Futures and Week Futures only have the off-peak load data, for all other contracts base and peak only. Here we recall the underlying of the Phelix Futures data, the Phelix Base contract is average price of the hours 1 to 24 for electricity traded on spot market, while the Phelix Peak is the average price of the hours 9 to 20 for electricity traded on spot market. Therefore we involve the products of spot prices as well. The contracts of spot prices are diversified in Hours from 00–01 h up to 23–24 h, and in Blocks of Base Monthly, off-peak 01–08, off-peak 21–24, Peak Monthly. The dataset we constructed is provided by Bloomberg, we have 90 kinds of contracts in total. The time span is from 30.09.2010 to 31.07.2015. All the contracts are listed in Table 8.1 with detailed information in Table 8.2.

To remove the redundant variable, we apply screening technique to select variables using the Phelix Futures consisting of different contracts and over different maturities. To implement the VAR model, first order difference of the data in Fig. 4.1a is required to transform non-stationary data to stationary time series. The contour plot of the constructed dataset is depicted in Fig. 4.1b.

In the market of Phelix Futures, final settlement at negative price is also possible. There are some missing values after transforming the original data to stationary time series by first order difference. To deal with the missing data, some quick fixes such as mean-substitution may be fine in some cases. While such simple approaches usually introduce bias into the data, for instance, applying mean substitution leaves the mean unchanged (which is desirable) but decreases variance, which may be undesirable. In our paper, we impute missing values with plausible values drawn from a distribution using an approach proposed by Van Buuren and Oudshoorn (2000). The patterns of missing data for the original dataset and imputation dataset are compared in Fig. 4.2. The distributions of the variables are shown as individual points, the imputed data for each imputed dataset is shown in magenta while the density of the observed data is shown in blue. The distributions are expected to be

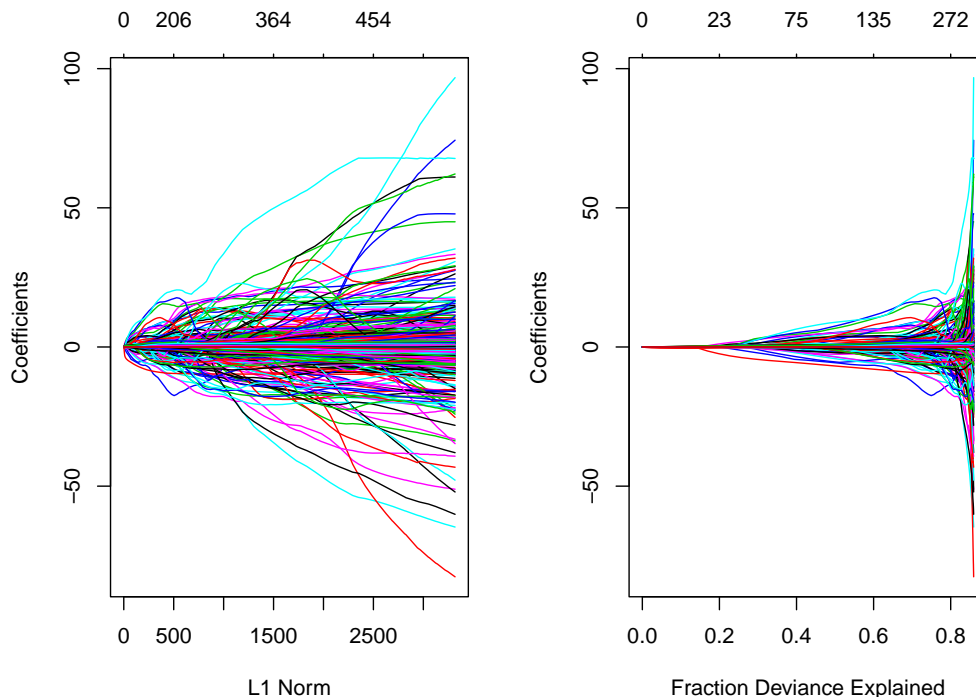


Fig. 4.3. Iterated-SIS-Lasso VAR(2) estimation results.

Table 5.1
The connectedness table for the aggregated network, including 11 types of contracts.

	GI	GT	HP	GJ	HI	NE	POA	PDA	PBA	LPXBHR	LPXBHxx	From
GI	1.75	0.46	0.86	1.46	0.44	0.71	1.48	0.48	0.80	0.15	0.15	8.74
GT	0.46	2.33	0.98	0.58	2.41	0.93	0.42	2.03	0.84	0.30	0.41	11.70
HP	0.73	0.84	5.27	0.63	0.51	4.78	0.70	1.21	4.19	0.06	0.03	18.95
GJ	1.46	0.58	0.73	1.80	0.63	0.64	1.09	0.51	0.65	0.13	0.13	8.36
HI	0.44	2.41	0.60	0.63	2.65	0.63	0.37	1.92	0.49	0.34	0.48	10.96
NE	0.61	0.79	4.78	0.55	0.54	5.09	0.55	1.05	3.34	0.08	0.12	17.52
POA	1.48	0.42	0.81	1.09	0.37	0.64	1.60	0.48	0.79	0.19	0.18	8.04
PDA	0.48	2.03	1.42	0.51	1.92	1.23	0.48	1.99	1.27	0.26	0.32	11.91
PBA	0.68	0.72	4.19	0.56	0.42	3.34	0.67	1.09	3.88	0.20	0.12	15.88
LPXBHR	0.80	1.24	0.50	0.67	1.35	0.50	0.90	1.12	1.01	7.86	9.81	25.79
LPXBHxx	0.13	0.35	0.03	0.11	0.41	0.12	0.16	0.28	0.13	2.70	3.86	8.28
To	9.03	12.18	20.17	8.59	11.65	18.62	8.41	12.16	17.38	12.28	15.63	146.12
Net	0.29	0.48	1.22	0.23	0.70	1.10	0.37	0.25	1.50	-13.50	7.35	

outperforms iterated-SIS-SCAD algorithm. Therefore the iterated-SIS-Lasso algorithm is selected for constructing the corresponding connectedness measure.

5. Static analysis of power market connectedness

5.1. The static network across contracts

The graph of our full-sample energy market network is depicted in Fig. 5.1. We observe the cluster phenomena in this graph, which motivates us to study the connectedness between contracts within and across 11 different types of energy contracts. In general, the contracts that belong to the same type tend to appear inside the same cluster. We find out several pairs of strong connections between different types of contracts, for example, the upper-left area reveals that the LPXBHR-type and LPXBHxx-type are massively connected. In addition, a cluster consisting of HP-type (Phelix Base Year Option), NE-type (Phelix Peak Year Future) and PBA-type (Phelix Off-Peak Year Future) indicates the closer relationship among these contracts; this implies that the Year derivative contracts are closer to each other while the month future and quarter future remain distinct.

Given the cluster phenomena observed in the full sample connectedness graph, we aggregate the pairwise connectedness of the contracts that belong to the same type. The aggregated network is then reported in Table 5.1, within a total of 11 types of contracts (more details can be found in Table 8.2). The off-diagonal elements represent the cross contractType connectedness, while within contractType connectedness are on the diagonal.

Recall that for each component in the system, the “to” connect-edness measures its total contribution to the forecast error variance

of other components in the system, making possible to capture the systemic risk of the component. In this table, it is obvious that the HP-type, NE-type and PBA-type are main contributors to systemic-wide risk. HP-type contracts have stronger links both from and to the other contract types. This is potentially interesting because, although HP-type, NE-type are important for the whole market as shown in Fig. 5.2, their net connectedness are negligible, with 4.52% and 4.08% of the total market power contracts. The reason is that these two types of contracts (HP, NE) are mutually closely interconnected, they also have “from” impacts to the system and thus offsetting their risk contributions. Moreover, the strongest pairwise connectedness is the impact of LPXBHxx-type on the LPXBHR-type, however the inverse impact is not significant. We also observe very strong net impacts from LPXBHxx-type on the LPXBHR-type contracts.

5.2. Determining important market component

In terms of magnitude for different contract types reported in Table 5.1, the “net” directional connectedness is distributed rather tightly, in total 77.21% of LPXBHR-type and LPXBHxx-type. We first investigate pairwise connectedness across 24 LPXBHR-type contracts (full connectedness table is available in Table 8.3).

Fig. 5.3 visualizes the connectedness network according to different trading hours, the numbered nodes correspond to the trading hours from 00–01 to 23–24 h. Some blocks of high connectedness are successfully detected and represented by darker purple color, i.e. the peak spot contracts with trading hours ranging from 09–13 h to 16–20 h. We summarize the “from”, “to” and “net” effects for the 24

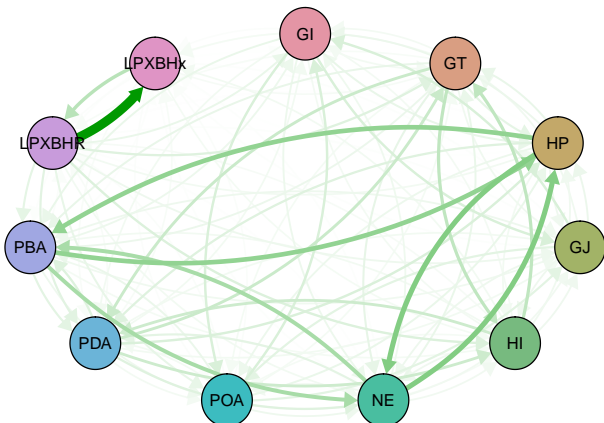


Fig. 5.2. The graph for network across 11 different contract types.

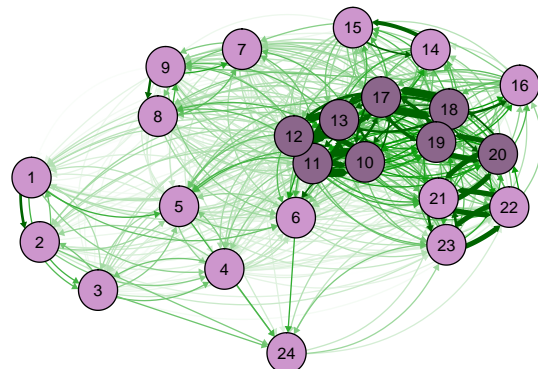


Fig. 5.3. The connectedness network graph based on different trading hours for LPXBHR-type spot contracts. The number in each node corresponds to the trading hours ranging from 00–01 to 23–24 h, e.g. 12 refers to the LPXBHR12 contract with trading hours 11–12 h. The LPXBHR-type power contracts that bidding from 09–13 h to 1620 h are represented with darker purple color because of their higher “to”-connectedness.

Table 5.2

Summary of “from”, “to” and “net” connectedness for all LPXBHR-type contracts. The contracts are ranked by their system risk measure “to”-connectedness.

Power contract	“from”-connectedness	“to”-connectedness	“net”-connectedness
LPXBHR11.Index	12.48	11.08	-1.40
LPXBHR10.Index	12.27	10.83	-1.45
LPXBHR12.Index	11.83	10.55	-1.27
LPXBHR13.Index	11.55	10.28	-1.27
LPXBHR17.Index	11.46	10.05	-1.41
LPXBHR19.Index	11.27	10.05	-1.23
LPXBHR18.Index	11.06	9.79	-1.27
LPXBHR20.Index	10.61	9.73	-0.88
LPXBHR21.Index	9.51	9.31	-0.19
LPXBHR22.Index	8.77	8.51	-0.26
LPXBHR23.Index	7.80	8.48	0.68
LPXBHR14.Index	7.48	8.16	0.68
LPXBHR16.Index	7.02	7.16	0.14
LPXBHR15.Index	6.51	7.15	0.64
LPXBHR01.Index	5.99	6.99	0.99
LPXBHR09.Index	5.64	6.54	0.90
LPXBHR06.Index	5.60	6.42	0.83
LPXBHR05.Index	5.35	6.30	0.94
LPXBHR04.Index	5.28	6.19	0.91
LPXBHR08.Index	5.13	5.65	0.52
LPXBHR07.Index	5.04	4.98	-0.06
LPXBHR02.Index	4.66	4.87	0.21
LPXBHR03.Index	4.36	4.85	0.49
LPXBHR24.Index	2.07	4.84	2.77

LPXBHR-type contracts in descending order of importance, and rank the contacts by their system risk in Table 5.2. Our finding clearly shows that, the impacts from day-ahead spot power contracts bidding between 09:00 and 13:00 have the highest “to”-connectedness values, and therefore have the strongest impacts on the other day-ahead spot contracts. The peak spot contracts (trading hours 09–13 h, 16–20 h) have very large negative “net”-connectedness, revealing that they are main risk source in determining the electricity price.

The pairwise directional impacts between LPXBHR-type and LPXBHxx-type are plotted in Fig. 5.4, the colors of the nodes are the same as Fig. 5.1 (the connectedness table of the impacts from the six LPXBHxx-type contracts to the 24 LPXBHR-type spot contracts is available in the appendix, see Table 8.4). We find a cluster of LPXBHR10, LPXBHR11, LPXBHR12, LPXBHR13, LPXBHRP, LPXBHRB

and LPXBHB. The LPXBHB (Base hours 00: 00–14:00) contract has significant impacts on the spot contracts from hours 09 to 13, while the impacts from the LPXBHP (Peak Hours 08: 00–20:00) contract is negligible. In addition, both LPXBHRB (Baseload) and LPXBHRP (Peakload) contracts exhibit strong interconnectedness with the spot contracts from hours 09 to 13, however only the LPXBHRP (Peakload) contract affects the spot prices trading from hours 16 to 18. Moreover, the graph exhibits strong mutual links between some of the spot contracts, for example, the pairs of LPXBHR10 and LPXBHR11, LPXBHR11 and LPXBHR12, LPXBHR12 and LPXBHR13, LPXBHR17 and LPXBHR18 among others.

The network graph in Fig. 5.5 illustrates the pairwise directional connectedness between HP-type and NE-type contracts. Compared with Fig. 5.1, we can see that, the links between these two types are

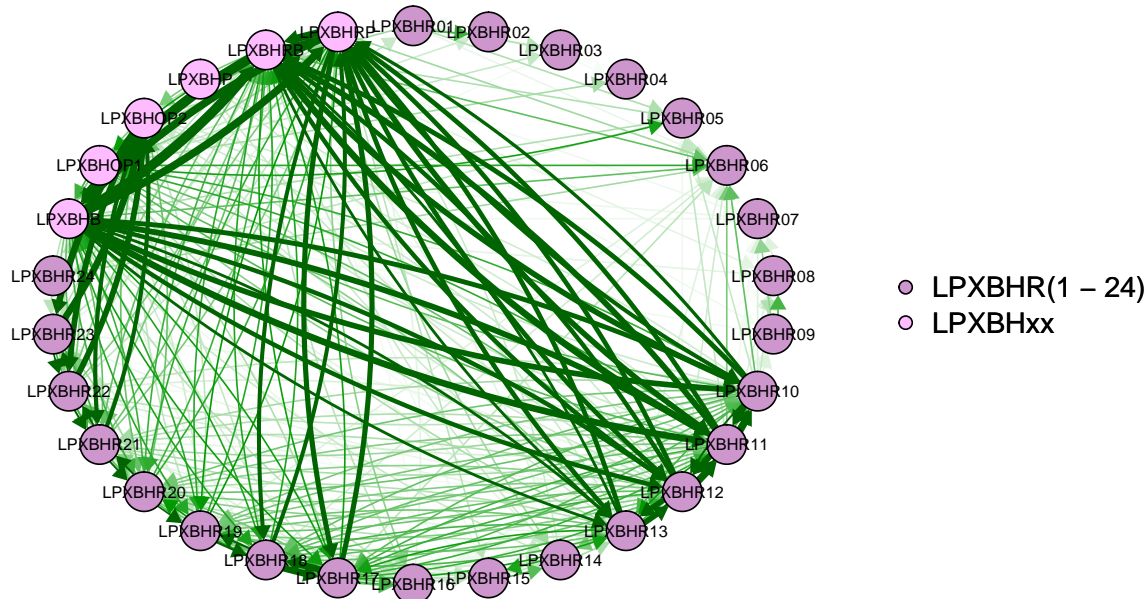


Fig. 5.4. The network graph for LPXBHR-type and LPXBHxx-type power contracts.

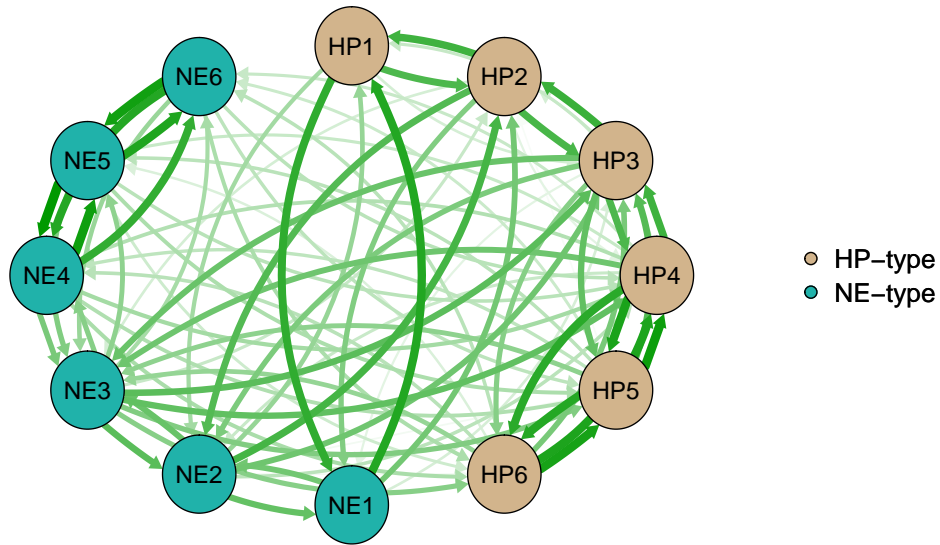


Fig. 5.5. The network graph for HP and NE-type power contracts.

very strong. NE-type contracts are the Year Futures with maturities up to six years, the underlying of these contracts is the average price of hours 9 to 20 for electricity traded on the spot market. HP-type contracts are the European style options on the Phelix Base Future provided by EEX, the underlying of Phelix Base is the average price of the hours 1 to 24 for electricity traded on the spot market. It is calculated for all calendar days of the year as the simple average of the Auction prices for the hours 1 to 24 in the market area Germany/Austria. This figure shows the similar connectedness pattern between HP1 and NE1 contracts.

6. Dynamic analysis of power market connectedness

We now study the dynamic network using rolling estimation. The number of observations used in the rolling sample to compute prediction is 36 or correspondingly three years, and we examine

dynamic evolution of the network for the following one year. In each window, we repeat model selection and conduct iterated-SIS algorithm to obtain the sparse estimates.

6.1. System-wide connectedness over time

We first calculate full sample system-wide connectedness for each window by summing up the total directional connectedness whether “from” or “to”. In general, the full sample system-wide connectedness reflects the overall uncertainty that has arisen in the system. The dynamic evolution path is plotted in the left panel of Fig. 6.1, with the peak of the system-wide connectedness is December. In most regions of Germany, the coldest days of the year are from around mid December to late January, which results in more work for the heating systems. In particular, the amount of electricity used for the decorations increases dramatically over the holiday season in

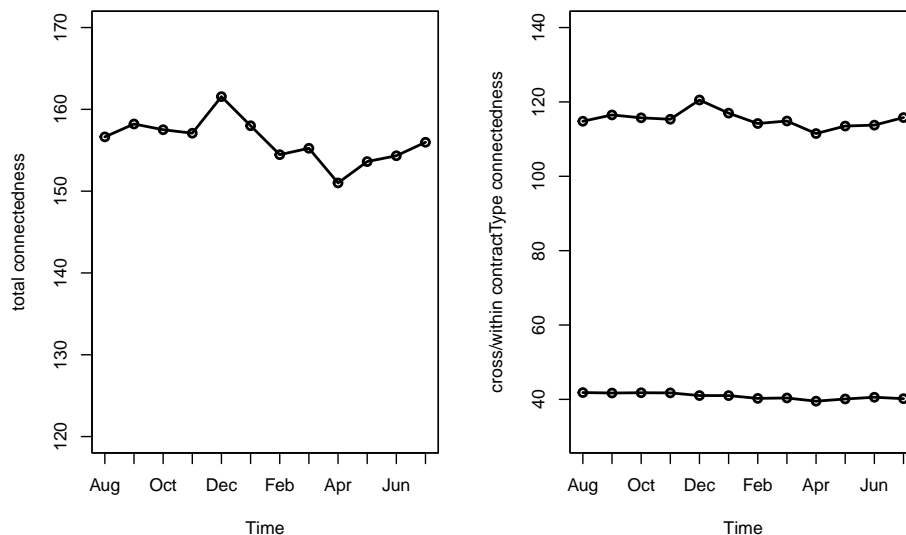


Fig. 6.1. The time-varying network for the system-wide connectedness from August 2014 to July 2015. The left panel is the time-varying full sample connectedness, it can be decomposed into two parts: the cross contractType connectedness (upper curve in the right panel) and the within contractType connectedness (lower curve in the right panel).

Table 6.1

The summary statistics of the connectedness measures for the aggregated network over one year, including 11 types of contracts. For each connectedness measure, we report the corresponding mean(15% quantile–85% quantile) and the standard deviation.

	"to"- connectedness		"from"- connectedness		"net"- connectedness	
	Mean	S.d.	Mean	S.d.	Mean	S.d.
GI	9.29 (9.02–9.68)	0.37	9.08 (8.85–9.33)	0.27	0.22 (0.06–0.36)	0.16
GT	12.40 (12.09–12.57)	0.26	11.92 (11.80–12.05)	0.15	0.49 (0.24–0.69)	0.19
HP	21.02 (20.44–21.52)	0.46	19.93 (19.43–20.41)	0.43	1.09 (1.03–1.15)	0.06
GJ	8.49 (8.16–8.90)	0.32	8.31 (7.99–8.69)	0.32	0.18 (0.12–0.22)	0.06
HI	11.92 (11.53–12.20)	0.34	11.16 (10.98–11.27)	0.18	0.75(0.52–0.92)	0.18
NE	19.59 (19.24–19.92)	0.34	18.49 (18.07–18.89)	0.41	1.11 (1.02–1.20)	0.09
POA	8.73 (8.28–9.29)	0.49	8.37 (8.15–8.70)	0.29	0.36 (0.13–0.61)	0.24
PDA	12.35 (12.12–12.49)	0.20	12.15 (11.95–12.29)	0.19	0.20 (-0.03–0.40)	0.20
PBA	18.87 (17.61–20.39)	1.35	17.60 (16.28–19.18)	1.39	1.27 (1.19–1.33)	0.10
LPXBHR	15.16 (13.62–16.55)	1.46	29.92 (26.25–33.36)	3.33	-14.76 (-16.59–12.63)	1.89
LPXBHxx	18.31 (16.74–19.86)	1.51	9.21 (8.43–9.85)	0.71	9.10 (8.30–9.88)	0.82

December. After December, the value of the system-wide connectedness decreased to the lowest value of April, and it then increased gradually with increasing temperature until September.

To measure the system-wide interaction, we further decompose the full sample system-wide connectedness into two parts: cross contractType connectedness and within contractType connectedness as shown in the right panel of Fig. 6.1. The cross contractType connectedness sums up the directional connectedness between the contracts coming from different contract types, while within contractType connectedness is the sum of directional connectedness of all the contracts within the same contract type. The higher values of cross contractType connectedness (upper curve) indicate that it is the main driving force for system-wide connectedness. However the within-contractType connectedness becomes less important, its values remain around 40 for the whole period.

We further conduct a robustness risk by computing full sample system-wide connectedness $C_{h,t}$ at different forecast horizons $h_1 = 8, h_2 = 9$. Then the dynamic networks of the resulting time-varying $C_{h_1,t}, C_{h_2,t}$ are compared with the above full sample system-wide connectedness $C_{h_0,t}, h_0 = 10$. To achieve this, we

apply a Welch two sample t-test on $H_0 : C_{h_0,t} = C_{h_1,t}$ against $H_a : C_{h_0,t} \neq C_{h_1,t}$. The p-value of the Welch t-test is 0.8994. We therefore cannot reject the null hypothesis, indicating that there is not enough evidence of a difference between the (true) averages of the two groups at the usual significance level. Moreover, the p-value of the Welch two sample t-test on $H_0 : C_{h_0,t} = C_{h_2,t}$ against $H_a : C_{h_0,t} \neq C_{h_2,t}$ is 0.9898. Both large p-values suggest the robustness of dynamic network estimation. The time-varying evolution of full sample system-wide connectedness estimated at different forecast horizons can be found in Fig. 8.1.

6.2. System-wide cross-contractType connectedness

To investigate the cross contractType connectedness, we aggregate the pairwise connectedness of the contracts that belong to the same type. The resulting time-varying networks across different contract types are summarized in Table 6.1. For each contract type, we report the mean and standard deviation (s.d.), the numbers inside the parentheses represent the 15% and 85% quantiles respectively. In general, the results are consistent with the

Table 6.2

The connectedness table reflects how the LPXBHxx-type contracts may affect spot contract according to different trading hours. In this table we report the mean of the time-varying networks, together with the s.d. inside the parentheses. Each element measures the directional pairwise connectedness from the jth contract of LPXBHxx-type to the ith LPXBHR-type, i.e. $C_{LPXBHR_i \leftarrow LPXBHxx_j}$. The numbers larger than 0.8 are marked with bold font.

	LPXBHB	LPXBHOP1	LPXBHOP2	LPXBHP	LPXBHRB	LPXBHRP
LPXBHR01.Index	0.28 (0.01)	0.45 (0.01)	0.11 (0.01)	0.15 (0.01)	0.28 (0.01)	0.19 (0.01)
LPXBHR02.Index	0.30 (0.01)	0.42 (0.02)	0.16 (0.01)	0.17 (0.01)	0.30 (0.01)	0.22 (0.01)
LPXBHR03.Index	0.34 (0.02)	0.54 (0.02)	0.14 (0.01)	0.16 (0.01)	0.34 (0.02)	0.25 (0.02)
LPXBHR04.Index	0.40 (0.02)	0.56 (0.02)	0.17 (0.01)	0.21 (0.01)	0.40 (0.02)	0.32 (0.01)
LPXBHR05.Index	0.43 (0.02)	0.62 (0.02)	0.24 (0.01)	0.26 (0.02)	0.43 (0.02)	0.34 (0.01)
LPXBHR06.Index	0.51 (0.02)	0.60 (0.03)	0.33 (0.01)	0.21 (0.01)	0.51 (0.02)	0.43 (0.02)
LPXBHR07.Index	0.42 (0.01)	0.38 (0.01)	0.29 (0.01)	0.11 (0.01)	0.42 (0.01)	0.38 (0.01)
LPXBHR08.Index	0.43 (0.01)	0.37 (0.01)	0.34 (0.01)	0.10 (0.01)	0.43 (0.01)	0.39 (0.01)
LPXBHR09.Index	0.61 (0.02)	0.49 (0.02)	0.34 (0.01)	0.14 (0.01)	0.61 (0.02)	0.58 (0.02)
LPXBHR10.Index	0.86 (0.02)	0.68 (0.02)	0.47 (0.01)	0.38 (0.01)	0.86 (0.02)	0.82 (0.02)
LPXBHR11.Index	0.85 (0.02)	0.66 (0.02)	0.44 (0.01)	0.45 (0.01)	0.85 (0.02)	0.83 (0.02)
LPXBHR12.Index	0.82 (0.02)	0.60 (0.02)	0.40 (0.01)	0.50 (0.01)	0.82 (0.02)	0.82 (0.02)
LPXBHR13.Index	0.78 (0.02)	0.53 (0.02)	0.34 (0.01)	0.51 (0.01)	0.78 (0.02)	0.82 (0.02)
LPXBHR14.Index	0.51 (0.01)	0.40 (0.01)	0.17 (0.01)	0.37 (0.01)	0.51 (0.01)	0.56 (0.02)
LPXBHR15.Index	0.46 (0.01)	0.38 (0.01)	0.16 (0.01)	0.30 (0.01)	0.46 (0.01)	0.50 (0.02)
LPXBHR16.Index	0.56 (0.02)	0.34 (0.01)	0.29 (0.01)	0.22 (0.01)	0.56 (0.02)	0.61 (0.02)
LPXBHR17.Index	0.71 (0.02)	0.44 (0.01)	0.28 (0.01)	0.29 (0.01)	0.71 (0.02)	0.81 (0.02)
LPXBHR18.Index	0.71 (0.02)	0.44 (0.01)	0.32 (0.01)	0.28 (0.01)	0.71 (0.02)	0.80 (0.02)
LPXBHR19.Index	0.70 (0.02)	0.43 (0.01)	0.45 (0.01)	0.34 (0.01)	0.70 (0.02)	0.73 (0.02)
LPXBHR20.Index	0.67 (0.02)	0.40 (0.01)	0.60 (0.02)	0.33 (0.01)	0.67 (0.02)	0.66 (0.02)
LPXBHR21.Index	0.58 (0.02)	0.33 (0.01)	0.78 (0.02)	0.29 (0.01)	0.58 (0.02)	0.53 (0.01)
LPXBHR22.Index	0.49 (0.01)	0.31 (0.01)	0.84 (0.02)	0.18 (0.01)	0.49 (0.01)	0.39 (0.01)
LPXBHR23.Index	0.43 (0.01)	0.22 (0.01)	0.84 (0.02)	0.33 (0.01)	0.43 (0.01)	0.36 (0.01)
LPXBHR24.Index	0.22 (0.01)	0.12 (0.01)	0.57 (0.02)	0.16 (0.01)	0.22 (0.01)	0.18 (0.01)

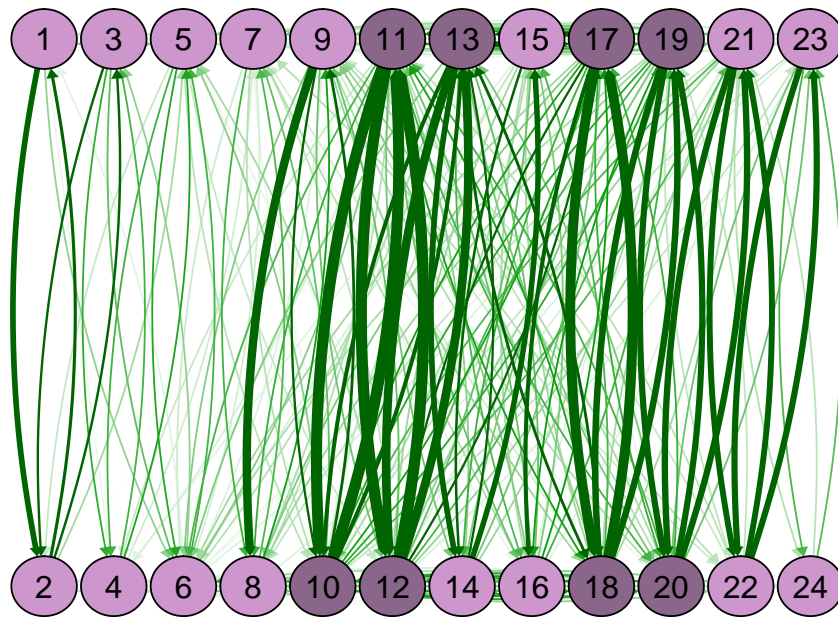


Fig. 6.2. The averaged pairwise connectedness among the LPXBHR-type contracts over the rolling period. The numbered nodes correspond to the underlying trading hours for each contract. The colors are the same as Fig. 5.3, the contracts with higher connectedness values are represented with darker purple color.

findings in Section 5.1. The HP-type, NE-type and PBA-type contracts have relatively higher values of both “to”-connectedness and “from”-connectedness, but their net impacts in the system are very low. Moreover, the LPXBHxx-type contracts have significant net impacts on the LPXBHR-type contracts, similar results are discussed in Section 5.2.

According to different trading hours, we further discuss how LPXBHxx-type contracts may affect the LPXBHR-type contracts traded in the power derivative market. The mean and s.d. for the time-varying connectedness $C_{LPXBHR_i \leftarrow LPXBHxx_j}$ are reported in Table 6.2. The spot contracts based on bid hours from 09 to 13 are closely related to the contracts of LPXBHB, LPXBHRB and LPXBHRP. We also find that the LPXBHRP affects the spot contracts from hours 16 to 19 most. To control for a relatively stable Germany spot electricity price, the risk transmission channels among the contracts are not negligible. Investments in LPXBHB, LPXBHRB and LPXBHRP contracts help to limit the potential risk of loss when there are adverse movements of spot prices. This result may provide a guide for policy maker, energy companies and investors.

We also compute the averaged pairwise connectedness among the spot contracts based on different trading hours, the interaction between the LPXBHR-type contracts is depicted in Fig. 6.2. We observe strong pairwise interconnection between the neighboring contracts. For example, the LPXBHR12 has stronger connection with its nearest two neighbors LPXBHR11 and LPXBHR13, but contract LPXBHR13 does not have strong linkage with LPXBHR14. The most influential contracts are identified as the contracts based on trading hours from 09–13 to 16–20. This is potentially interesting as it provides pricing signals affecting the electricity trading.

7. Conclusion

This paper combines regularization approach with dynamic network analysis in an ultra high-dimensional setting. We empirically analyze the sparse interconnectedness of the full German power derivative market, clearly identify the significant channels of relevant potential risk spillovers and thus complement the full picture of system risk. As we know, electricity is not storable and may be affected by various factors on the supply and demand side of

the market, such as policy changes, weather conditions and external economic uncertainties. Nowadays Germany is transforming its power system towards renewable energy, analysis of German power derivative market thus provides useful insights for power generation companies and transmission organizations across the globe.

Our empirical findings suggest that the Phelix base year options and peak year futures are the main contributors to the system risk. However these two types of contracts are mutually closely interconnected, they also have high “from” impacts received from the system and thus offsetting their risk contributions. In addition, the connectedness between different contract types are more significant compared with the connectedness among contracts within the same type. Therefore it is important for policy makers and investors to take the interdependence between different contract-types into account. Other interesting conclusions arise from the connectedness of spot and future contracts, when we examine each contract according to different trading hours. For spot contracts, we observe strong pairwise interconnections between the neighboring contracts especially for the contracts trading in the peak hours, this provides investors pricing signals affecting the electricity trading. In addition, the monthly base future (BHB), monthly peak-load(BHRP)/baseload(BHRB) are identified as main driving force for the peak-hour spot contracts. This evidence has implications for regulators to control for a relatively stable Germany spot electricity price. For energy companies and investors, it is important to diversify their existing portfolio rather than having large holdings of individual electricity contract, for example, investments in LPXBHB, LPXBHRB and LPXBHRP contracts help to limit the potential risk of loss when there are adverse movements of spot prices. Another important characteristic of electricity is seasonality, this characteristic is reflected in our dynamic network analysis. We observe the dynamic evolution of full-sample system-wide connectedness in case of weather condition and external uncertainty, for example the full-sample system-wide connectedness increased gradually with increasing temperature from April until September. In general, with the wide range of power derivative contracts trading in the German electricity market, we are able to identify, estimate the risk contribution of individual power contract; this helps us to have a better understanding of the German power market functioning and environment.

Appendix A

Table 8.1
Phelix Futures data traded at EEX.

Symbol	Types			
GI	GI1.Comdty	GI2.Comdty	GI3.Comdty	GI4.Comdty
	GI5.Comdty	GI6.Comdty	GI7.Comdty	
GT	GT1.Comdty	GT2.Comdty	GT3.Comdty	GT4.Comdty
	GT5.Comdty	GT6.Comdty	GT7.Comdty	
HP	HP1.Comdty	HP2.Comdty	HP3.Comdty	HP4.Comdty
	HP5.Comdty	HP6.Comdty		
GJ	GJ1.Comdty	GJ2.Comdty	GJ3.Comdty	GJ4.Comdty
	GJ5.Comdty	GJ6.Comdty	GJ7.Comdty	
HI	HI1.Comdty	HI2.Comdty	HI3.Comdty	HI4.Comdty
	HI5.Comdty	HI6.Comdty	HI7.Comdty	
NE	NE1.Comdty	NE2.Comdty	NE3.Comdty	NE4.Comdty
	NE5.Comdty	NE6.Comdty		
POA	POA1.Comdty	POA2.Comdty	POA3.Comdty	POA4.Comdty
	POA5.Comdty	POA6.Comdty	POA7.Comdty	
PDA	PDA1.Comdty	PDA2.Comdty	PDA3.Comdty	PDA4.Comdty
	PDA5.Comdty	PDA6.Comdty	PDA7.Comdty	
PBA	PBA1.Comdty	PBA2.Comdty	PBA3.Comdty	PBA4.Comdty
	PBA5.Comdty	PBA6.Comdty		
LPXBHR	LPXBHR01.Index	LPXBHR02.Index	LPXBHR03.Index	LPXBHR04.Index
	LPXBHR05.Index	LPXBHR06.Index	LPXBHR07.Index	LPXBHR08.Index
	LPXBHR09.Index	LPXBHR10.Index	LPXBHR11.Index	LPXBHR12.Index
	LPXBHR13.Index	LPXBHR14.Index	LPXBHR15.Index	LPXBHR16.Index
	LPXBHR17.Index	LPXBHR18.Index	LPXBHR19.Index	LPXBHR20.Index
	LPXBHR21.Index	LPXBHR22.Index	LPXBHR23.Index	LPXBHR24.Index
LPXBHxx	LPXBHBM1.Index	LPXBHOP1.Index	LPXBHOP2.Index	LPXBHPMI.Index
	LPXBHRBS.Index	LPXBHRPK.Index		

Table 8.2
Selected contracts from the file "Products 2016" provided by [EuropeanEnergyExchangeEEXAG](#).

No.	Symbol	Description
1	GI1.Comdty - GI7.Comdty	Phelix Base Month Option, and the respective next six delivery months
2	GT1.Comdty - GT7.Comdty	Phelix Base Quarter Option, and the respective next six delivery quarters
3	HP1.Comdty - HP6.Comdty	Phelix Base Year Option, and the respective next five delivery years
4	GJ1.Comdty - GJ7.Comdty	Phelix Peak Month Future, and the respective next six delivery months
5	HI1.Comdty - HI7.Comdty	Phelix Peak Quarter Future, and the respective next six delivery quarters
6	NE1.Comdty - NE6.Comdty	Phelix Peak Year Future, and the respective next five delivery years
7	POA1.Comdty - POA7.Comdty	Phelix Off-Peak Month Future, and the respective next six delivery months
8	PDA1.Comdty - PDA7.Comdty	Phelix Off-Peak Quarter Future, and the respective next six delivery quarters
9	PBA1.Comdty - PBA6.Comdty	Phelix Off-Peak Year Future, and the respective next five delivery years
10	LPXBHR01.Index - LPXBHR24.Index	EEX Day-ahead Spot Market with Bid Type from 00-01 to 23-24h, e.g. LPXBHR14.Index is EEX Day-ahead Spot price based on bid hours from 13 -14.
11	LPXBHRxx.Index	EEX Day-ahead Spot Market with different Bid Types: LPXBHB.Index is Base Monthly 00-14h; LPXBHOP1.Index is Off Peak1 01-08h; LPXBHOP2.Index is Off Peak2 21-24h; LPXBHP.Index is Peak Monthly 08 - 20h; LPXBHRB.Index is Baseload; LPXBHRP.Index is Peakload.

Table 8.3
Population connectedness table for LPXBHR contracts.

	LPXBHR01	LPXBHR02	LPXBHR03	LPXBHR04	LPXBHR05	LPXBHR06	LPXBHR07	LPXBHR08
LPXBHR01	1.00	0.71	0.52	0.35	0.53	0.43	0.13	0.14
LPXBHR02	0.42	0.59	0.47	0.36	0.31	0.18	0.06	0.07
LPXBHR03	0.27	0.41	0.54	0.42	0.36	0.23	0.12	0.12
LPXBHR04	0.19	0.32	0.41	0.53	0.38	0.25	0.21	0.15
LPXBHR05	0.25	0.24	0.30	0.33	0.47	0.37	0.19	0.15
LPXBHR06	0.18	0.12	0.18	0.18	0.34	0.41	0.22	0.23
LPXBHR07	0.10	0.08	0.13	0.20	0.22	0.29	0.45	0.40
LPXBHR08	0.10	0.08	0.13	0.14	0.18	0.30	0.45	0.54
LPXBHR09	0.11	0.08	0.14	0.15	0.18	0.35	0.58	0.70
LPXBHR10	0.25	0.23	0.23	0.33	0.49	0.66	0.46	0.49
LPXBHR11	0.27	0.28	0.26	0.35	0.50	0.61	0.42	0.46
LPXBHR12	0.26	0.29	0.23	0.30	0.42	0.48	0.28	0.31
LPXBHR13	0.23	0.30	0.24	0.30	0.39	0.45	0.21	0.27
LPXBHR14	0.16	0.15	0.17	0.23	0.26	0.30	0.13	0.13
LPXBHR15	0.22	0.18	0.17	0.19	0.22	0.27	0.14	0.13
LPXBHR16	0.07	0.06	0.10	0.13	0.19	0.34	0.36	0.41

Table 8.3 (continued)

	LPXBHR01	LPXBHR02	LPXBHR03	LPXBHR04	LPXBHR05	LPXBHR06	LPXBHR07	LPXBHR08
LPXBHR17	0.13	0.13	0.18	0.22	0.35	0.51	0.29	0.36
LPXBHR18	0.13	0.11	0.15	0.20	0.32	0.47	0.24	0.27
LPXBHR19	0.15	0.11	0.12	0.24	0.32	0.45	0.33	0.30
LPXBHR20	0.10	0.07	0.08	0.17	0.24	0.38	0.32	0.28
LPXBHR21	0.06	0.07	0.05	0.12	0.17	0.28	0.27	0.25
LPXBHR22	0.11	0.13	0.11	0.12	0.19	0.31	0.23	0.26
LPXBHR23	0.10	0.12	0.06	0.07	0.11	0.18	0.12	0.14
LPXBHR24	0.00	0.00	0.01	0.01	0.01	0.00	0.00	0.00
	LPXBHR09	LPXBHR10	LPXBHR11	LPXBHR12	LPXBHR13	LPXBHR14	LPXBHR15	LPXBHR16
LPXBHR01	0.16	0.25	0.27	0.26	0.23	0.13	0.08	0.03
LPXBHR02	0.07	0.15	0.19	0.20	0.21	0.07	0.04	0.03
LPXBHR03	0.10	0.13	0.14	0.13	0.14	0.07	0.05	0.05
LPXBHR04	0.12	0.21	0.23	0.20	0.20	0.11	0.10	0.15
LPXBHR05	0.12	0.24	0.25	0.22	0.21	0.11	0.08	0.11
LPXBHR06	0.20	0.29	0.27	0.23	0.22	0.13	0.12	0.15
LPXBHR07	0.38	0.27	0.26	0.20	0.16	0.09	0.09	0.16
LPXBHR08	0.53	0.28	0.27	0.19	0.17	0.08	0.08	0.09
LPXBHR09	0.73	0.33	0.31	0.22	0.19	0.10	0.11	0.09
LPXBHR10	0.45	0.98	0.94	0.82	0.70	0.43	0.37	0.43
LPXBHR11	0.43	0.94	0.98	0.92	0.81	0.44	0.37	0.43
LPXBHR12	0.31	0.84	0.94	1.00	0.90	0.46	0.36	0.44
LPXBHR13	0.27	0.71	0.83	0.90	1.00	0.62	0.51	0.43
LPXBHR14	0.13	0.39	0.41	0.41	0.54	0.83	0.77	0.27
LPXBHR15	0.14	0.31	0.32	0.30	0.40	0.67	0.71	0.27
LPXBHR16	0.41	0.39	0.38	0.34	0.32	0.22	0.26	0.59
LPXBHR17	0.33	0.68	0.72	0.72	0.78	0.50	0.51	0.65
LPXBHR18	0.24	0.66	0.66	0.65	0.70	0.48	0.50	0.71
LPXBHR19	0.27	0.63	0.63	0.60	0.63	0.48	0.47	0.62
LPXBHR20	0.26	0.64	0.61	0.58	0.52	0.34	0.32	0.49
LPXBHR21	0.24	0.56	0.54	0.53	0.44	0.25	0.22	0.35
LPXBHR22	0.24	0.51	0.47	0.42	0.35	0.17	0.14	0.32
LPXBHR23	0.14	0.40	0.41	0.44	0.39	0.22	0.17	0.27
LPXBHR24	0.00	0.03	0.04	0.07	0.07	0.02	0.01	0.03
	LPXBHR17	LPXBHR18	LPXBHR19	LPXBHR20	LPXBHR21	LPXBHR22	LPXBHR23	LPXBHR24
LPXBHR01	0.13	0.13	0.15	0.10	0.06	0.11	0.10	0.00
LPXBHR02	0.08	0.07	0.08	0.07	0.14	0.17	0.22	0.38
LPXBHR03	0.09	0.08	0.07	0.06	0.11	0.12	0.15	0.42
LPXBHR04	0.16	0.16	0.18	0.15	0.17	0.16	0.18	0.36
LPXBHR05	0.17	0.15	0.16	0.14	0.19	0.18	0.22	0.50
LPXBHR06	0.22	0.20	0.21	0.19	0.25	0.24	0.27	0.55
LPXBHR07	0.20	0.19	0.23	0.22	0.20	0.18	0.16	0.17
LPXBHR08	0.20	0.16	0.18	0.18	0.21	0.20	0.17	0.20
LPXBHR09	0.24	0.17	0.20	0.19	0.17	0.18	0.10	0.00
LPXBHR10	0.67	0.65	0.61	0.63	0.55	0.50	0.39	0.04
LPXBHR11	0.71	0.65	0.62	0.60	0.53	0.46	0.40	0.05
LPXBHR12	0.72	0.65	0.61	0.58	0.53	0.42	0.43	0.08
LPXBHR13	0.78	0.70	0.63	0.52	0.44	0.35	0.39	0.08
LPXBHR14	0.42	0.41	0.41	0.30	0.22	0.16	0.21	0.07
LPXBHR15	0.38	0.37	0.36	0.25	0.18	0.13	0.15	0.05
LPXBHR16	0.46	0.46	0.42	0.36	0.28	0.26	0.20	0.02
LPXBHR17	1.00	0.93	0.74	0.62	0.46	0.36	0.27	0.02
LPXBHR18	0.93	1.00	0.83	0.68	0.46	0.38	0.30	0.02
LPXBHR19	0.74	0.83	1.00	0.81	0.59	0.45	0.46	0.05
LPXBHR20	0.62	0.68	0.81	1.00	0.82	0.64	0.56	0.08
LPXBHR21	0.46	0.46	0.59	0.82	1.00	0.82	0.71	0.27
LPXBHR22	0.36	0.38	0.45	0.64	0.82	1.00	0.81	0.23
LPXBHR23	0.27	0.30	0.46	0.56	0.71	0.81	1.00	0.38
LPXBHR24	0.03	0.03	0.06	0.08	0.22	0.19	0.31	0.82

Table 8.4

The "To" impacts from the six LPXBHxx-type power contracts to the 24 LPXBHR-type contracts. The numbers larger than 0.8 are marked in bold font.

	LPXBHB	LPXBHOP1	LPXBHOP2	LPXBHP	LPXBHRB	LPXBHRP
LPXBHR01	0.29	0.49	0.10	0.10	0.29	0.18
LPXBHR02	0.30	0.48	0.12	0.14	0.30	0.18
LPXBHR03	0.28	0.50	0.07	0.04	0.28	0.18
LPXBHR04	0.38	0.54	0.10	0.12	0.38	0.27
LPXBHR05	0.53	0.78	0.15	0.14	0.53	0.40
LPXBHR06	0.66	0.75	0.24	0.08	0.66	0.55
LPXBHR07	0.43	0.43	0.20	0.01	0.43	0.36
LPXBHR08	0.44	0.40	0.20	0.01	0.44	0.39
LPXBHR09	0.42	0.36	0.20	0.01	0.42	0.37

(continued on next page)

Table 8.4 (continued)

	LPXBHB	LPXBHOP1	LPXBHOP2	LPXBHP	LPXBHRB	LPXBHRP
LPXBHR10	0.91	0.67	0.49	0.29	0.91	0.87
LPXBHR11	0.93	0.68	0.48	0.37	0.93	0.90
LPXBHR12	0.88	0.59	0.48	0.49	0.88	0.88
LPXBHR13	0.84	0.54	0.42	0.48	0.84	0.88
LPXBHR14	0.49	0.29	0.22	0.33	0.49	0.56
LPXBHR15	0.43	0.23	0.17	0.21	0.43	0.51
LPXBHR16	0.52	0.29	0.31	0.14	0.52	0.59
LPXBHR17	0.79	0.48	0.35	0.24	0.79	0.89
LPXBHR18	0.77	0.44	0.37	0.24	0.77	0.87
LPXBHR19	0.75	0.43	0.50	0.30	0.75	0.81
LPXBHR20	0.72	0.37	0.65	0.31	0.72	0.73
LPXBHR21	0.65	0.33	0.87	0.28	0.65	0.61
LPXBHR22	0.61	0.36	0.91	0.16	0.61	0.50
LPXBHR23	0.51	0.26	0.93	0.34	0.51	0.44
LPXBHR24	0.07	0.11	0.41	0.10	0.07	0.06

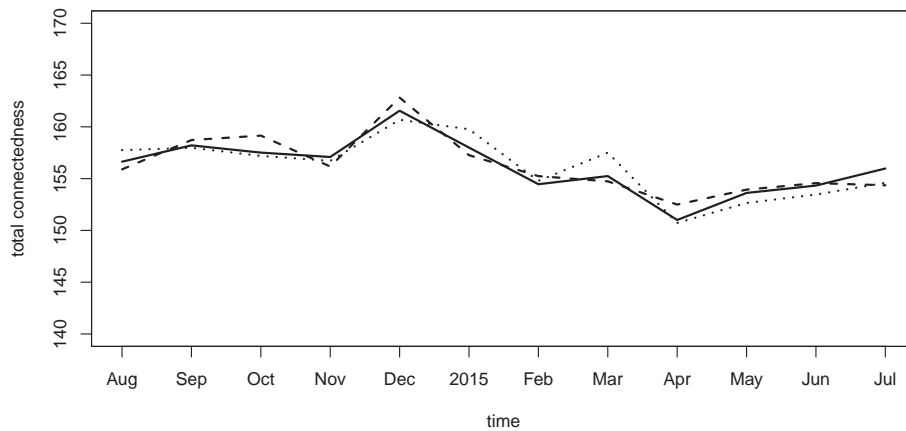


Fig. 8.1. The time-varying full sample system-wide connectedness estimated at different forecast horizons, from August 2014 to July 2015. The solid line is $C_{h_0,t}$, the dashed line is $C_{h_1,t}$, and the dotted line is $C_{h_2,t}$.

References

- Acharya, V., Engle, R., Richardson, M., 2012. Capital shortfall: a new approach to ranking and regulating systemic risks. *Am. Econ. Rev.* 102 (3), 59–64.
- Acharya, V.V., Pedersen, L.H., Philippon, T., Richardson, M., 2017. Measuring systemic risk. *Rev. Financ. Stud.* 30 (1), 2–47.
- Apergis, N., Baruník, J., Lau, M.C.K., 2017. Good volatility, bad volatility: what drives the asymmetric connectedness of Australian electricity markets? *Energy Econ.* 66, 108–115.
- Arouri, M.E.H., Jouini, J., Nguyen, D.K., 2012. On the impacts of oil price fluctuations on European equity markets: volatility spillover and hedging effectiveness. *Energy Econ.* 34 (2), 611–617.
- Awartani, B., Aktham, M., Cherif, G., 2016. The connectedness between crude oil and financial markets: evidence from implied volatility indices. *J. Commod. Mark.* 4 (1), 56–69.
- Babus, A., 2016. The formation of financial networks. *RAND J. Econ.* 47 (2), 239–272.
- Balaguer, J., 2011. Cross-border integration in the European electricity market. Evidence from the pricing behavior of Norwegian and Swiss exporters. *Energy Policy* 39 (9), 4703–4712.
- Ballester, C., Furió, D., 2015. Effects of renewables on the stylized facts of electricity prices. *Renew. Sust. Energ. Rev.* 52, 1596–1609.
- Basu, S., Michailidis, G., et al. 2015. Regularized estimation in sparse high-dimensional time series models. *Ann. Stat.* 43 (4), 1535–1567.
- Benhmad, F., Percebois, J., 2018. Photovoltaic and wind power feed-in impact on electricity prices: the case of Germany. *Energy Policy* 119, 317–326.
- Bickel, P.J., Ritov, Y., Tsybakov, A.B., 2009, 08. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* 37 (4), 1705–1732. <https://doi.org/10.1214/08-AOS620>.
- Bierbrauer, M., Menn, C., Rachev, S.T., Trück, S., 2007. Spot and derivative pricing in the EEX power market. *J. Bank. Financ.* 31 (11), 3462–3485.
- BMW, 2016. Zahlen und fakten. URL <http://www.bmwi.de/DE/Themen/Energie/Strommarkt-der-Zukunft/zahlen-fakten.html>.
- Böckers, V., Heimeshoff, U., 2014. The extent of European power markets. *Energy Econ.* 46, 102–111.
- Brownlees, C., Engle, R.F., 2016. Srisk: a conditional capital shortfall measure of systemic risk. *Rev. Financ. Stud.* 30 (1), 48–79.
- Bunn, D.W., Gianfreda, A., 2010. Integration and shock transmissions across European electricity forward markets. *Energy Econ.* 32 (2), 278–291.
- Candes, E., Tao, T., 2007. The Dantzig selector: statistical estimation when p is much larger than n . *Ann. Stat.* 2313–2351.
- Castagneto-Gissey, G., Chavez, M., Fallani, F.D.V., 2014. Dynamic Granger-causal networks of electricity spot prices: a novel approach to market integration. *Energy Econ.* 44, 422–432.
- Demirer, M., Diebold, F.X., Liu, L., Yilmaz, K., 2018. Estimating global bank network connectedness. *J. Appl. Econ.* 33 (1), 1–15.
- Diebold, F.X., Yilmaz, K., 2009. Measuring financial asset return and volatility spillovers, with application to global equity markets. *Econ. J.* 119 (534), 158–171.
- Diebold, F.X., Yilmaz, K., 2012. Better to give than to receive: predictive directional measurement of volatility spillovers. *Int. J. Forecast.* 28 (1), 57–66.
- Diebold, F.X., Yilmaz, K., 2014. On the network topology of variance decompositions: measuring the connectedness of financial firms. *J. Econ.* 182 (1), 119–134.
- Du, X., Cindy, L.Y., Hayes, D.J., 2011. Speculation and volatility spillover in the crude oil and agricultural commodity markets: a Bayesian analysis. *Energy Econ.* 33 (3), 497–503.
- Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. 2004. Least angle regression. *Ann. Stat.* 32 (2), 407–499.
- Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* 96 (456), 1348–1360.
- Fan, J., Lv, J., 2008. Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* 70 (5), 849–911.
- Fan, J., Samworth, R., Wu, Y., 2009. Ultrahigh dimensional feature selection: beyond the linear model. *J. Mach. Learn. Res.* 10 (Sep), 2013–2038.
- Fan, J., Song, R., et al. 2010. Sure independence screening in generalized linear models with np-dimensionality. *Ann. Stat.* 38 (6), 3567–3604.
- Frank, L.E., Friedman, J.H., 1993. A statistical view of some chemometrics regression tools. *Technometrics* 35 (2), 109–135.
- Fu, W.J., 1998. Penalized regressions: the bridge versus the Lasso. *J. Comput. Graph. Stat.* 7 (3), 397–416.
- Geman, H., Roncoroni, A., 2006. Understanding the fine structure of electricity prices. *J. Bus.* 79 (3), 1225–1261.
- Giglio, S., Kelly, B., Pruitt, S., 2016. Systemic risk and the macroeconomy: an empirical evaluation. *J. Financ. Econ.* 119 (3), 457–471.

- Grossi, L., Heim, S., Waterson, M., 2017. The impact of the German response to the Fukushima earthquake. *Energy Econ.* 66, 450–465.
- Hautsch, N., Schaumburg, J., Schienle, M., 2014. Financial network systemic risk contributions. *Review of Finance* 19 (2), 685–738.
- Henriques, I., Sadorsky, P., 2008. Oil prices and the stock prices of alternative energy companies. *Energy Econ.* 30 (3), 998–1010.
- Higgs, H., 2009. Modelling price and volatility inter-relationships in the Australian wholesale spot electricity markets. *Energy Econ.* 31 (5), 748–756.
- Joo, Y.C., Park, S.Y., 2017. Oil prices and stock markets: does the effect of uncertainty change over time? *Energy Econ.* 61, 42–51.
- Ketterer, J.C., 2014. The impact of wind power generation on the electricity price in Germany. *Energy Econ.* 44, 270–280.
- Knittel, C.R., Roberts, M.R., 2005. An empirical examination of restructured electricity prices. *Energy Econ.* 27 (5), 791–817.
- Kock, A.B., Callot, L., 2015. Oracle inequalities for high dimensional vector autoregressions. *J. Econ.* 186 (2), 325–344.
- Koop, G., Pesaran, M.H., Potter, S.M., 1996. Impulse response analysis in nonlinear multivariate models. *J. Econ.* 74 (1), 119–147.
- Loh, P.-L., Wainwright, M.J., 2011. High-dimensional regression with noisy and missing data: provable guarantees with non-convexity. *Advances in Neural Information Processing Systems*. pp. 2726–2734.
- Lundgren, A.I., Milicevic, A., Uddin, G.S., Kang, S.H., 2018. Connectedness network and dependence structure mechanism in green investments. *Energy Econ.* 72, 145–153.
- Lütkepohl, H., 2005. *New introduction to multiple time series analysis*. Springer Science & Business Media.
- Maghyereh, A.I., Awartani, B., Bouri, E., 2016. The directional volatility connectedness between crude oil and equity markets: new evidence from implied volatility indexes. *Energy Econ.* 57, 78–93.
- Nicholson, W.B., Matteson, D.S., Bien, J., 2017. Varx-l: structured regularization for large vector autoregressions with exogenous variables. *Int. J. Forecast.* 33 (3), 627–651.
- Paraschiv, F., Erni, D., Pietsch, R., 2014. The impact of renewable energies on EEX day-ahead electricity prices. *Energy Policy* 73, 196–210.
- Pesaran, H.H., Shin, Y., 1998. Generalized impulse response analysis in linear multivariate models. *Econ. Lett.* 58 (1), 17–29.
- Reboredo, J.C., 2014. Volatility spillovers between the oil market and the European union carbon emission market. *Econ. Model.* 36, 229–234.
- Sadorsky, P., 2012. Correlations and volatility spillovers between oil prices and the stock prices of clean energy and technology companies. *Energy Econ.* 34 (1), 248–255.
- Seifert, S., Cullmann, A., von Hirschhausen, C., 2016. Technical efficiency and CO2 reduction potentials: an analysis of the German electricity and heat generating sector. *Energy Econ.* 56, 9–19.
- Sinn, H.-W., 2017. Buffering volatility: a study on the limits of Germany's energy revolution. *Eur. Econ. Rev.* 99, 130–150.
- Spiecker, S., Eickholt, V., Weber, C., 2014. The impact of carbon capture and storage on a decarbonized German power market. *Energy Econ.* 43, 166–177.
- Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* 58 (1), 267–288. ISSN 00359246.
- Van Buuren, S., Oudshoorn, C., 2000. *Multivariate Imputation by Chained Equations. MICE V1.0 user's manual*. Leiden: TNO Preventie en Gezondheid.
- Wang, G.-J., Xie, C., Zhao, L., Jiang, Z.-Q., 2018. Volatility connectedness in the Chinese banking system: do state-owned commercial banks contribute more? *J. Int. Financ. Mark. Inst. Money* 57, 205–230.
- Weron, R., 2007. *Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach*. 403. John Wiley & Sons.
- Worthington, A., Kay-Spratley, A., Higgs, H., 2005. Transmission of prices and price volatility in Australian electricity spot markets: a multivariate garch analysis. *Energy Econ.* 27 (2), 337–350.
- Wu, W.-B., Wu, Y.N., et al. 2016. Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electron. J. Stat.* 10 (1), 352–379.
- Yi, S., Xu, Z., Wang, G.-J., 2018. Volatility connectedness in the cryptocurrency market: is bitcoin a dominant cryptocurrency? *Int. Rev. Financ. Anal.* 60, 98–114.
- Zachmann, G., 2008. Electricity wholesale market prices in Europe: convergence? *Energy Econ.* 30 (4), 1659–1671.
- Zhang, D., 2017. Oil shocks and stock markets revisited: measuring connectedness from a global perspective. *Energy Econ.* 62, 323–333.
- Zou, H., 2006. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* 101 (476), 1418–1429.
- Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat Methodol.)* 67 (2), 301–320.

Industry Interdependency Dynamics in a Network

Context ^{*}

Ya Qian,[†] Wolfgang Karl Härdle,[‡] Cathy Yi-Hsuan Chen [§]

April, 2017

Abstract

This paper contributes to model the industry interconnecting structure in a network context. General predictive model (Rapach et al. 2016) is extended to quantile LASSO regression so as to incorporate tail risks in the construction of industry interdependency networks. Empirical results show a denser network with heterogeneous central industries in tail cases. Network dynamics demonstrate the variety of interdependency across time. Lower tail interdependency structure gives the most accurate out-of-sample forecast of portfolio returns and network centrality-based trading strategies seem to outperform market portfolios, leading to the possible 'too central to fail' argument.

Keywords: dynamic network, interdependency, general predictive model, quantile LASSO, connectedness, centrality, prediction accuracy, network-based trading strategy

JEL classification: C32, C55, C58, G11, G17

^{*}The financial support from the Deutsche Forschungsgemeinschaft, Humboldt-Universität zu Berlin and IRTG 1972 'High Dimensional Non Stationary Time Series' is gratefully acknowledged.

[†]Corresponding author. Humboldt-Universität zu Berlin, C.A.S.E.-Center of Applied Statistics and Economics, Unter den Linden 6, 10099 Berlin, Germany. Email: qianya@hu-berlin.de

[‡]Humboldt-Universität zu Berlin, C.A.S.E.-Center of Applied Statistics and Economics, Unter den Linden 6, 10099 Berlin, Germany. Research fellow in Sim Kee Boon Institute for Financial Economics, Singapore Management University, 90 Stamford Road, 6th Level, School of Economics, Singapore 178903. Research fellow in Wang Yanan Institute for Studies in Economics(WISE), A307,Economics Building, Xiamen University, Xiamen, 361005 China. Research fellow in National Chiao Tung University, No. 1001, Daxue Rd, East District, Hsinchu City, Taiwan 300.

[§]Humboldt-Universität zu Berlin, C.A.S.E.-Center of Applied Statistics and Economics, Unter den Linden 6, 10099 Berlin, Germany.

1 Introduction

Interdependency among different assets is always the key topic of portfolio management. From the very beginning of portfolio theory (Markowitz 1952), correlation between every two assets is considered to be one of the most important factors in portfolio construction. Ever since the recent financial crisis, studies in interdependencies in the context of risk management have increased rapidly with most of them showing a great interest in the dependency structure within financial sector, i.e. financial contagion (Rodriguez 2007, May & Arinaminpathy 2010, Hasman 2013, Georg 2013, Acemoglu et al. 2015). However, as broad asset allocation including industry assets becomes more and more popular, interdependency among industries started to attract more attention as well. Some research takes the perspective of the interdependency among financial sector and other real economy sectors (Baur 2012, Chiu et al. 2015, Claessens et al. 2012). With no exception, work in this direction concentrates on the effect of financial sector on other real economy sectors, not the way around. Whilst if we consider an easy example containing three corporations, being them oil (X), car manufactory (Y) and autos dealer (Z) from the perspective of supply chain, we have to say that interconnectivity across different industries (not limited to financial to others) is pretty common, as the work done by Rapach et al. (2016). They use the one-period predictive model to establish return predictability among different industries as a depiction of industry interdependency among various sectors and claim the interdependency is pretty widespread among each other. Nevertheless, the industry interdependency in an extreme or stress situation hasn't been addressed intensively. One may imagine that the interdependency may not necessarily show a monotonic linearity w.r.t the quantile level being considered. We therefore contribute to the extant literature by aiming this extreme interdependency which can be referable in the industry portfolio in a market downturn. Tail event based quantile regression with LASSO regularization is implemented here, which is cast into a dynamic network context.

We study the industry interdependency from the network point of view for mainly three

important reasons: First, it has been proved in literature as an excellent tool to depict interconnectivities. Real network analysis includes the work of Schweitzer et al. (2009), by taking a socioeconomic perspective, they argue a network architecture built upon trade, *R&D* alliances, ownership or credit-debt relationships can vividly study the strategic behavior of the interacting agents. Gençay et al. (2015) use North American supplier-customer network data of public companies to assess counterparty risk and detect counterparty network effects as significant determinants of credit spreads. In the empirical part of Zhu, Pan, Li, Liu & Wang (2016), they test the 'Chinese Twitter' - Xinlang Weibo social network and observe a significant network effect in Chinese social activities. Zhu, Wang, Wang & Härdle (2016) extend Zhu, Pan, Li, Liu & Wang (2016) into the quantile regression framework to consider tail risks. They then exert the quantile network autoregressive model to describe Chinese stocks' interconnecting behaviors on the basis of common shared ownership information. For artificial networks, statistical methods need to be used to construct linkages. Based on vector autoregressive (VAR) model, Diebold & Yilmaz (2014) propose a generalized variance decomposition to define a weighted directed network. They apply their method to US financial institutions and it turns out to be coordinated pretty well with the 2008 financial crisis. Similarly, Billio et al. (2011) use linear as well as nonlinear Granger-Causality tests to construct pairwise connections in the network and apply it to monthly returns of different sectors of finance department stocks. Their empirical results show the advantage of network models in measuring the systematic risk levels. Chan-Lau et al. (2016) adopt a default correlation model to construct the forward-looking partial default correlations, which turn out to be the network element. Depending on their network construction, they study the systemic risk of over 1000 exchange-traded banks in the global network framework and argued that connectivity hasn't been paid enough emphasis in Financial Stability Board. Härdle et al. (2016) propose a nonlinear semiparametric quantile regression method on CoVaR to construct a tail-event driven network in order to study the systemic risk among different financial sectors and conclude that the interconnectedness is growing during financial crisis period with largest systemic risk receivers and emitters being the most systemically

important.

The second reason for adopting network methodology due to industry characteristics themselves. As it is not just limited to the dyadic relationship, industry interconnectivity has more complicated dependency structure. It focuses on concentration. In a bunch of various companies from different industries, some companies are more important in the sense that they are connected to more of the others with others just being less important locating in the periphery. An obvious example concerning concentration is the banking industry acts actively in the group of companies which need funding. None of these funding-needed companies can thrive without the financing from the banking sector. In addition, Baxamusa et al. (2015) provided the empirical evidence that in the customer-supplier network, the more central the firm, the lower its returns from the acquisition activities. And this concentration can be easily established in the network framework.

Thirdly, it is not hard to properly specify the node set and edge set in our research question, which are the basic elements of a network structure. Fama-French industry portfolios are the nodes, while for edges establishment, we have to use appropriate statistics methods. Former popular calibration methods on interdependency analysis include correlation analysis (see Chiang et al. (2016), for instance), vector autoregressive models (e.g. Diebold & Yilmaz (2014)) and copula based methodology (Poshakwale & Mandal (2016)). However, to include the information of return predictability and further use the network to do forecasts, we construct the edges as the one-month ahead return predictive model parameters. As Rapach et al. (2016) argue, there is a significant relationship between their general predictive model and the US production network. We therefore claim the setup of using predictive model parameters as edges is reasonable.

The motivation to incorporate tail risks are the consideration of the parallelity between industry portfolios and financial stocks. Lots of research has investigated the importance of downside risk in financial stocks and comparatively, beyond just financial sector, we

conjecture that industry interdependencies are also affected by extreme situations. Therefore, we refer to Tibshirani (1996) and Li & Zhu (2008) to introduce quantile LASSO techniques upon the industry return general predictive model of Rapach et al. (2016). We have the penalizing techniques come into play here to solve high dimensionality problem. For different quantiles, we can model median level as well as tail level interconnectivities to fulfill the purpose of comparison. In a nutshell: Tail event based quantile regression with LASSO regularization is cast into a dynamic network context. Our main contribution is to extend the general predictive regression framework into the tail case using quantile LASSO regularization to construct networks under different quantile levels and compare median and tail-centered return data of industry portfolios in order to show the increase of the connection during extreme periods. Based on the differences between normal and extreme markets, we would also like to compare the prediction accuracy of the one-month forward return. For utilizing the network information into financial markets, we will construct network-based trading strategies of industry portfolios as well in order to see whether the markets can be beaten. Lastly, by studying the dynamic structure of industry portfolio network across time, we would like to discern some evolution pattern of this industry network.

The remaining of our work is organized as follows: Section 2 describes the econometric model that we are using to construct the industry network; Some basic concepts of network structure and its key parameters are given in Section 3; Section 4 shows the empirical network analysis of 49 industry portfolios obtained from Kenneth French's data library. Analysis in this part includes the construction of whole network as well as dynamic networks, prediction accuracy computation and performance of network-based trading strategy; Section 5 concludes and summarizes. Tables and figures are organized in Appendices at the end.

2 Econometric Modelling of Industry Interdependency

Rapach et al. (2016) propose a general predictive model of industry returns to study the interdependency among 30 industries. They compare their model to American production network, concluding that their model could represent a good construction of industry network. However, only the sector return is considered in their work, while we all agree that in stress situations, tail events carry information on the network infrastructure. This motivates to extend the general predictive model to different tails, e.g. quantile levels, to investigate the role tail risks playing in industry portfolios.

The general predictive model proposed by Rapach et al. (2016) is given as follows:

$$r_{i,t+1} = \beta_{0,i} + \sum_{j=1}^N \beta_{i,j} r_{j,t} + \varepsilon_{i,t+1}, t = 1, \dots, T - 1 \quad (1)$$

where $r_{i,t}$ is the monthly return of industry portfolio i at time t ; N is the total number of industries and $\varepsilon_{i,t}$ is the white noise error term.

The generalized quantile regression is described as: $\{X, Y\} = \{x_i, y_i\}_{i=1}^n$, $x_i = (x_{i1}, \dots, x_{ip})^\top \in \mathbb{R}^p$, $\tau \in (0, 1)$.

$$Y = X\beta + \varepsilon \quad (2)$$

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n \rho_\tau(y_i - x_i^\top \beta) \quad (3)$$

where $\rho_\tau(\cdot)$ is an asymmetric loss function:

$$\rho_\tau(u) = |u|^\alpha |\mathbf{I}(u \leq 0) - \tau|, \alpha \geq 1 \quad (4)$$

with $\alpha = 1$ and $\alpha = 2$ corresponding to a quantile and expectile regression respectively, see Breckling & Chambers (1988). The aforementioned general predictive model (1) is a special case of the generalized quantile regression (2), if we set $y_i = r_{i,t+1}$, $x_i = (1, r_t)$, $\alpha = 1$, $\tau = 0.5$, where $r_t = \{r_{j,t}\}_{j=1}^N$.

For large dimension p one runs into singularity problems and a plethora of too many small

coefficients. Prediction accuracy and model interpretability become so big problems with large p that standard ordinary least squares (OLS) turns out to be valid no longer. Standard techniques for improving the OLS estimates contain subset selection and ridge regression while neither of which solves the two problem simultaneously. The way to go here is the implementation of Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani 1996):

$$\hat{\beta}(LASSO) = \arg \min_{\beta} \sum_{i=1}^n (y_i - x_i^{\top} \beta)^2 + \lambda \sum_{i=1}^p |\beta_i| \quad (5)$$

where λ is a nonnegative regularization parameter, p is the total number of possible covariates that explains Y and the second term $\sum_{i=1}^p |\beta_i| = \|\beta\|_1$ is the l_1 norm.

Combining the idea of tail event QR with LASSO leads us to

$$\hat{\beta}(qLASSO) = \arg \min_{\beta_0, \beta} \sum_{t=1}^{T-1} \rho_{\tau}(r_{i,t+1} - \beta_0 - r_t^{\top} \beta) + \lambda \|\beta\|_1 \quad (6)$$

where r_t denotes the return vector of all industries at time t ; β the vector of coefficients of the regression and β_0 the intercept. The l_1 -norm quantile LASSO model can be referred to Li & Zhu (2008).

As is known from Härdle & Simar (2015), the solution to (6) yields a finite subset of nonzero elements of the $\hat{\beta}(qLASSO)$ vector. The coefficients in this 'active set' may be called 'prominent' since all other coefficients are actually zero.

Later in Section 4 we will use quantile LASSO regression method to build the network across different industry portfolios. Before going to that, we would like to give a brief introduction of network structure, the main graphic tool in our analysis.

3 Network Structure

A binary set $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ represents the network structure of a system where \mathcal{V} denotes the collection of vertices (also called nodes) in the system and \mathcal{E} stands for the collection

of edges (or linkages, etc) between each pair of the vertices. In our application, vertices \mathcal{V} are the industries and the edges \mathcal{E} are constructed as the prominent quantile LASSO coefficients. A network is always corresponding to an adjacency matrix which specifies the edges between each pair of the nodes. At the very beginning, adjacency matrix is simply a symmetric binary matrix but it is later extended to weighted and asymmetric ones, which is exactly the case in our later empirical analysis.

Given a network \mathcal{G} , two important and interesting questions are always asked: first, how to measure the graph level connectivity of the network? Second, which parameter gives us insight into the relative importance of each vertice?

3.1 Revisit *Fagiolo (2007)*

The answer to the first question is the concept of 'connectedness'. Connectedness is a measure specified in network analysis depicting the degree of interdependency among all nodes and the whole network connectedness is achieved by averaging all the node-specific connectedness. In the context of graph theory, connectedness is usually referred to as clustering coefficient, which measures the inherent tendency of nodes clustering together. The global version of clustering coefficient gives the measure of the connectedness of the whole network. The most common definition is designed for undirected and binary adjacency matrices, following Fagiolo (2007), we use four patterns (cycle, middleman, in and out) to depict directed networks.

- *cycle*: there is a cyclical relation among i and any two of its neighbors ($i \rightarrow j \rightarrow h \rightarrow i$ or viceversa);
- *middleman*: when one of i 's neighbors reach a third neighbor directly with an outward edge or indirectly with i as a medium;
- *in*: i has two inward edges;
- *out*: i has two outward edges.

For an asymmetrix binary adjacency matrix, *in-degree*, *out-degree* *total-degree* and *bilateral-degree* of node i are defined as:

$$\begin{aligned} d_i^{in} &= \sum_{j \neq i} a_{ji} = (A^\top)_i \mathbf{1} \\ d_i^{out} &= \sum_{j \neq i} a_{ij} = A_i \mathbf{1} \\ d_i^{tot} &= d_i^{in} + d_i^{out} = (A^\top + A)_i \mathbf{1} \\ d_i^{\leftrightarrow} &= \sum_{j \neq i} a_{ij} a_{ji} = A_{ii}^2 \end{aligned}$$

where A^\top is the transpose of A , A_i the i_{th} row of A , A_{ii} the i_{th} diagonal element of A , $\mathbf{1}$ is the N -dimensional column vector $(1, 1, \dots, 1)^\top$.

Based on above notations, the number of all possible triangles that node i could form (T_i^D):

$$T_i^D = d_i^{tot}(d_i^{tot} - 1) - 2d_i^{\leftrightarrow}$$

For weighted adjacency matrix W , the four patterns of clustering coefficient are defined as:

$$\begin{aligned} C_i^{cyc} &= \frac{(W^{[1/3]^3})_{ii}}{d_i^{in} d_i^{out} - d_i^{\leftrightarrow}} \\ C_i^{mid} &= \frac{(W^{[1/3]} W^{[1/3]^\top} W^{[1/3]})_{ii}}{d_i^{in} d_i^{out} - d_i^{\leftrightarrow}} \\ C_i^{in} &= \frac{(W^{[1/3]^\top} W^{[1/3]^2})_{ii}}{d_i^{in} (d_i^{in} - 1)} \\ C_i^{out} &= \frac{(W^{[1/3]^2} W^{[1/3]^\top})_{ii}}{d_i^{out} (d_i^{out} - 1)} \end{aligned}$$

where $W^{[1/3]}$ denotes the matrix with each element generated as the cubic roots of W 's

elements. To get overall connectedness, we just average these C_i^* via:

$$C^* = N^{-1} \sum_{i=1}^N C_i^*$$

where * stands for elements in *cyc*, *mid*, *in*, *out*.

3.2 Centrality Measures

For the second, the answer is centrality. Centrality basically answers the question 'what characterizes the important vertices?' There are various kinds of centrality definitions. In the simplest cases, *Degree centrality* measures how many ties each node has and assigns the biggest value of importance to the node which has the largest number of ties. A more complex extension of degree centrality is to consider the directions of linkages in directed networks. Therefore we have 'in' as well as 'out' degree centralities. Though simple and easily to exert, degree centrality assigns equal values to all the edges. Baveias (1950) defined the *closeness centrality* of a node as the average length of the shortest path between the node and all other nodes. Freeman (1977) introduced the *betweenness centrality* that measures the number of times that the node plays as a bridge along the shortest path between any other two nodes. A more appropriate version of these two centrality measures is to incorporate the concept of 'cost' in which case we define the shortest path in the sense of actual length instead of number of nodes. Since the version with 'cost' pays attention to the actual distance between each pair of node, it is more applicable to real world networks, for instance, transportation networks. As for differentiating the relative importance of different nodes, these two measures contributes little. A good measure to incorporate relative importance of different nodes is *Eigenvector centrality*. For a network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, *eigenvector centrality* of node v - $C_E(v)$ equals

$$C_E(v) = \frac{1}{\lambda} \sum_{t \in M(v)} C_E(t) = \frac{1}{\lambda} \sum_{t \in \mathcal{G}} a_{v,t} C_E(t) \quad (7)$$

where λ is the maximum eigenvalue of the adjacency matrix A ; $M(v)$ the set of neighbors of v and $a_{v,t}$ the element of A in row v and column t . According to this measure, a node in a network is important if it is linked to other important nodes. Hence one does not consider the edges between every pair of the nodes as equally important, one assigns different importance value through the first eigenvector. Similar to degree centrality, when in directed graphs, we can have 'in' and 'out' eigenvector centrality measures to discriminate the 'receiving' and 'emitting' effects respectively. Since industries need to be treated differently and receiving and emitting effects have to be set apart, we are going to adopt eigenvector centrality for the weighted, directed graphs in the empirical part.

4 Empirical Results

4.1 Data

Monthly return of the 49 industry portfolios constructed by Kenneth R. French is used as our data sample. The data is available from Kenneth French's webpage¹. As monthly data is mostly often used in industry portfolio analysis, we select it from January 1970 to January 2017 with 565 observations in total.

As quantile-quantile (QQ) plots (Figure 1 to Figure 5) of the 49 industry portfolios show, compared to normal distributions, tail behaviors exist in most industries, which justifies our analysis of focusing on industry network structure at different tail levels.

4.2 Whole sample network

We now come to the network construction based on Equation (6). The edge between node i and node j exists if and only if the lagged return of industry j (i) is selected by LASSO as the significant predictor of the return of industry i (j). The edges are constructed as directional: if i (j) helps predicting j (i), then the edge goes from i to j (j to i); if i

¹http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

helps predicting j and j helps predicting i as well, then the edge between i and j has both arrows. Furthermore, the edges in our industry network are also weighted. And the weight assigned to each edge depends on the absolute value of the beta coefficients. To sum up, the adjacency matrix corresponding to our industry interdependency network is the absolute value of beta coefficients.

To discern the probable varieties between median network and tail-event driven network, we set $\tau = (0.05, 0.5, 0.95)$, denoting the crisis, stable and boom situations respectively. Figure 6 depicts the whole sample (1970.01-2017.01) industry networks under these three situations. In every subfigure, we locate the 49 industries in a circle and fix their positions in favor of convenient comparison among different τ levels. The nodes in the network have different sizes according to their 'in' and 'out' eigenvector centrality scores respectively, as introduced aforementioned. We arrange the network plots with eigenvector centrality from the 'in' direction in the left panel and 'out' the right panel of each subfigure in Figure 6. Specifically, the leading industries possess larger sizes in our industry network. The grey arrows with directions within the network circle suggest the intensity of the interconnectivity of the industry network. Comparing the density of edges in these three figures, we can clearly reach the conclusion that in extreme cases, the industry network connection increases a lot when comparing to stable situation, which means the whole economy becomes more connected in extreme cases, as is listed in Table 2. Meanwhile, the leading industries change as well. To see this perspective more precisely, we list the top leading industries which has eigenvector centrality score larger than 0.200 (both 'in' and 'out') under each situation in Table 3. Comparing the leading industries in different cases, we have several interesting findings:

- a. More central industries are identified under extreme cases with more even centrality score distribution. It is a signal of intercorrelation rise among various industries in extremes.
- b. Financial-related industries (banking, insurance and trading) play important roles as risk emitters under whatever market situations (banking has rank 6 as risk emit-

ters when $\tau = 0.95$ with centrality score 0.186). It is determined by the nature of financial-related industries.

- (a) Banking evolves as a leading risk emitter when extreme events happen. It complies with our common knowledge that when markets go to extremes, banking affects other industries more.
 - (b) Insurance arises as the NO.1 risk emitter when market goes down. It is a reflection of the market sentiment.
 - (c) However, when market goes down, financial trading also becomes a crucial risk receiver which to some degree, signifies the hard time that traders have to endure during financial crisis.
- c. Coal remains leading risk emitter in various markets. It may be accounted to the relative position of coal industry in the supply chain. As an upstream industry, coal price can actually affect the return of many others which need it as raw material. Furthermore, no matter what the market looks like, basic production still has to be done.
- d. Gold always stays as top risk receiver for which we may argue from the role gold plays as a financial hedging instrument.
- e. The leading industries detected has little to do with industry size according to our empirical analysis.

4.3 Dynamic networks

In the last section we see the differences in whole sample networks under different τ_s . However the information is limited since it is a static picture. One step further, we can gain more insight by investigating the dynamics of networks. Through this, we expect to discover some potential patterns in industry networks. As introduced in Sec 3.1, we plot the 4 clustering coefficients under three τ_s in a moving window framework. We compute the network structures using quantile LASSO for samples of every three years'

data. The data sample starts from 1970.01, then we move the window forward every three months to compute the next one, that is, the first sample is 1970.01-1972.12, the second one 1970.04-1973.03, so on so forth. Finally, we get 177 data points of each clustering coefficients and their plots are shown in Figure 7 (arranged by different τ_s) and Figure 8 (arranged by different clustering patterns in a directed network).

The dynamic networks vary across time are shown in Fig 7 and Fig 8. Under each τ levels, the 'cycle' and 'out' clustering are lower than 'middleman' and 'in' clustering in our industry network. Besides, 'cycle' and 'out' also possess smaller volatilities. However, the four have similar pattern under each τ across time. Around 1997-2002, 2007-2010 and 2012-2013 there are sharp ups and downs in all connectedness measures no matter whether we take tail risks into account. These time slots also approximately match the economic crisis in history. When comparing each of the four connectedness measures under various τ_s , we once again validate that median level network is less connected than those in tail cases, from the perspective of network dynamics. When zooming in those special time periods, we have an interesting finding that lowertail connectedness behave oppositely with uppertail and median ones in periods 2000-2002, 2011-2012, 2016 thereafter but move more simultaneously during the recent financial crisis during 2007-2008. This, probably, can be explained by the relative important positions financial-related industries possess in the industry network. When the financial-related industries triggered the crisis in 2007 and 2008, lots of industries are affected which leads to the comovement of the connectedness under different τ_s . This argument also complies with centrality analysis in section 4.2.

4.4 Three additions based on specific network construction

4.4.1 Prediction Directions

As far as we discussed, we use the absolute value of the beta coefficients as the adjacency matrix inasmuch as to comply with the canonical definition of adjacency matrix. However, this setting has the drawback of ignoring the signs of coefficients, which is very useful

in return prediction. As is contained in the quantile LASSO regression (6), not only do we know the magnitude of industry return prediction network (the absolute values of the betas), but also the information on directions of industry return predictability (the signs of the beta coefficients). Therefore we add the information of whole sample interdependency in the image plots in Fig 9 and in Table 4 and of rolling window predictability in Table 5, which, in our opinion, are valuable additions to the canonical network analysis. In order for later return prediction, the rolling window is constructed as follows: we first use the data from 1970.01-1995.12 to construct the first 'network' composed of beta coefficients (here with signs). After the first step, we enlarge our estimation window with one more month every time, i.e. in the second step we compute the beta coefficients network using data from 1970.01-1996.01, and so on so forth. Finally we get 253 beta networks of the industry interdependency.

In all three image plots (Figure 9), the horizontal represents the predictive power one industry getting from all others (receiving) and vertical stands for the predictive power that industry to others (emitting). Comparing these three image plots in Table 4, we can see that in common, the LASSO method selects quite a small set of significant predictors out of the 2401 cells in total. Besides, the differences are quite obvious. The difference between normal market and extreme markets is that when market changes from stable to extremes, a larger quantity of significant connections are detected (either in crisis or in booms). In network language, we say the entire industries become more interconnected in stress situations. Furthermore, there are still some differences between these two extremes. First of all, more negative connections are detected in boom than that in crisis while more positive ones in crisis than in boom, indicating a higher and non-diversified tail risk and a difficulty w.r.t industry diversification in the market crisis. The benefit of industry diversification is diminished in this situation, implying an inevitable tail risk. Second of all, the average connectedness is positive when $\tau = 0.05$ and negative when $\tau = 0.95$. It sends out the signal that bad market conditions tend to affect most industries in the same direction while good ones is more favorable for portfolio management since the average connectedness is negative. The averaged dynamic directions of the beta

coefficients under rolling window framework depict a similar pattern in Table 5 as that in Table 4. Combining them together, we claim it's hard to diversify industry portfolios in crisis, indicating a higher and non-diversified tail risk, which, when translates into risk-return relationship, tells us that we can expect higher returns with the tolerance of higher risks, if we construct appropriate trading strategies. This point is going to be shown in 4.4.3.

4.4.2 Prediction Performance

To make use of the industry network structure that we constructed before and the prediction directions information, a dynamic network structure within a rolling window framework is used to predict the one-month ahead industry returns. We here compare the performance of the interdependencies under different τ levels in the use of predicting future returns. Based on the predicted and actual monthly industry returns, we calculate the root mean squared error (RMSE) of the three models, i.e. with different τ_s , as a measure of the prediction accuracy. Specifically, we first use the data from 1970.01-1995.12 to construct the first 'network' composed of beta coefficients(here with signs). We multiply this beta matrix with the industry monthly returns in 1995.12 as the prediction of the industry monthly returns in 1996.01. After the first step, we enlarge our estimation window with one more month every time, i.e. in the second step we compute the beta coefficients network using data from 1970.01-1996.01 and multiply the new beta matrix with the returns in 1996.01 as a prediction of those in 1996.02, and so on so forth. Finally we get 253 predictions of the monthly returns for each industry. Lastly, we compute square of the differences between the predicted returns and the real ones and average them over the length of time series, i.e., 253. Figure 10 shows the average RMSEs of the 49 industries in our data sample for the three τ_s .

As expected, beta 'network' dynamics under extreme cases have better prediction performance (smaller out-of-sample forecast error) than that in normality. In general, for these three cases, we claim that lower tail prediction achieves the highest accuracy (the smallest out-of-sample forecast error). Therefore, incorporating tail risks contributes to

industry return prediction which justifies the necessity of extending extant studies which focus only on median level to different quantiles.

4.4.3 Network centrality-based trading strategies

As we can see from the aforementioned discussion, some vital characteristics about the industry interdependency network have been investigated. However, we still lack the tactic to make use of the network information to profit from financial markets. We are going into this direction here. Network centrality-based trading strategies are considered due to the importance of the concept of centrality as well as our specific construction of industry network. As discussed earlier, centrality measures the relative importance of different industries in the network structure, which, in our framework, is based on one-month-ahead return predictability. Therefore the most central nodes demonstrate the relatively important roles of influencing or being influenced by others more than the rest in the sense of return predictability. That is, industries with higher 'out' ('in') eigenvector centrality scores can also be called 'prediction emitters' ('prediction receivers') which affect more of others (are affected more by others) in the sense of return prediction.

We conjecture that highly centralized industries are more likely to outperform market portfolios due to the similar effect as 'too connected to fail', we call it 'too central to fail'. Centralized industries are more connected to other central industries and hence possess more complicated risk structures, which in turn, lead to higher excess returns. However, we would also like to conjecture that the centrality-based trading portfolios have no abnormal return, i.e., no mispricing about the industry portfolios. To verify our assumption, we do the same rolling window computation as described in Subsection 4.4.2 to generate a time series of 253 networks under each τ level and find the first and least leading industries of each network. Then we construct the 4 centrality-based trading strategies and balance them every month with the updated beta network. Finally, we calculate the average of annualized cumulative log-returns of these strategies and of the MKT and report their *t-stat* to decide whether the excess returns are significant. Also, we regress the monthly excess returns of these strategies to Fama-French three risk factors

so as to check the existence of abnormal returns. Details are reported in Table 6 and Table 7. The 4 trading strategies we constructed are as follows:

- long the top 10 leading 'in' centralized industries (HI strategy)
- long the top 10 leading 'out' centralized industries (HO strategy)
- long the bottom 10 leading 'in' centralized industries (TI strategy)
- long the bottom 10 leading 'out' centralized industries (TO strategy)

The empirical results mostly authenticate the assumption under various market situations: except for HI in median case, all other more centralized industry portfolios outperforms less centralized ones and outperform market portfolios as well. Furthermore, they have no significant abnormal return when regressing on Fama-French three risk factors. Even in market with huge downside risk, the more centralized portfolios gain a sizeable excess return. Therefore, our empirical analysis stands up for the assumption of 'too central to fail' of the industry networks, which rests on complicated risk structures of more centralized industry portfolios.

5 Conclusion

This study extends the general predictive model of industry portfolios to different quantile levels so as to incorporate tail risks in interdependency measurement and construct network analysis under different market situations. By comparing median level ($\tau = 0.50$) with upper ($\tau = 0.95$) and lower ($\tau = 0.05$) tail networks, we find out that interdependency across USA industries increases a lot during extreme market situations in the whole period from 1970.01 to 2017.01. Similar results are achieved under dynamic network framework - connectedness in stress situations is always higher than that in normality. The time series of four connectedness measures corresponding to weighted directed networks show significant varieties of the interdependency structure dynamics. In addition, leading industries vary as well when market switches from stable to highly volatile. An

obvious finding is that financial-related industries evolve as leading ones under stress situations which highlights their role in bad times. At last, three more additions to network analysis are summarized in the last subsection. First, prediction directions differ under different τ_s : lower-tail case involves more positive coefficients while upper-tail circumstance has more negative ones, which reflects the movement in the same direction of different assets in crisis, i.e. asset returns are highly affected in the same direction when market goes down. While in promising situations, they tend to be negatively related therefore it is more effective to do risk management during these periods since we can easily find negative-related assets. Second, when quantifying the one-month-ahead industry return prediction accuracy using dynamic coefficients networks under different quantile levels, as expected, the lower tail case gives the best prediction performance in the sense of RMSE. Third, four trading strategies based on network centrality dynamics are constructed and compared with market portfolio. Empirical results report significant sizeable excess returns for more centralized industry portfolios, which outperform less centralized ones and market portfolio, even in bad market situations. With a risk adjustment for Fama-French factors, these strategies do not possess significant risk-adjusted abnormal returns. Therefore, the higher returns come from the more complicated risk structures central industries endow. To conclude, we argue for the possibility of the effect 'too central to fail'.

References

- Acemoglu, D., Ozdaglar, A. & Tahbaz-Salehi, A. (2015), ‘Systemic risk and stability in financial networks’, *The American Economic Review* **105**(2), 564–608.
- Baur, D. G. (2012), ‘Financial contagion and the real economy’, *Journal of Banking & Finance* **36**, 2680–2692.
- Baveias, A. (1950), ‘Communication patterns in task-oriented groups’, *Journal of the Acoustical Society of America* **22**, 725–730.
- Baxamusa, M., Javaid, S. & Harery, K. (2015), ‘Network centrality and mergers’, *Review of Quantitative Finance and Accounting* **44**, 393–423.
- Billio, M., Getmansky, M., Lo, A. & Pelizzon, L. (2011), ‘Econometric measures of connectedness and systemic risk in the finance and insurance sectors’, *Journal of Financial Economics* **17**(1), 1–23.
- Breckling, J. & Chambers, R. (1988), ‘M-quantiles’, *Biometrika* **75**, 761–771.
- Chan-Lau, J. A., Chuang, C., Duan, J.-c. & Sun, W. (2016), Banking network and systematic risk via forward-looking partial default correlations, Technical report, Working papers//International Monetary Fund and National Singapore University.
- Chiang, T. C., Lao, L. & Xue, Q. (2016), ‘Comovement between chinese and global stock markets: Evidence from aggregate and sectoral data’, *Review of Quantitative Finance and Accounting* **47**, 1003–1042.
- Chiu, W.-C., Peña, J. I. & Wang, C.-W. (2015), ‘Industry characteristics and financial risk contagion’, *Journal of Banking & Finance* **75**, 411–427.
- Claessens, S., Tong, H. & Wei, S.-J. (2012), ‘From the financial crisis to the real economy: Using firm-level data to identify transmission channels’, *Journal of International Economics* **88**, 375–387.

- Diebold, F. X. & Yilmaz, K. (2014), ‘On the network topology of variance decompositions: Measuring the connectedness of financial firms’, *Journal of Econometrics* **182**, 119–134.
- Fagiolo, G. (2007), ‘Clustering in complex directed networks’, *Physical Review E* **76**, 026107–1–8.
- Freeman, L. C. (1977), ‘A set of measures of centrality based on betweenness’, *Sociometry* **40**(1), 35–41.
- Gençay, R., Signori, D., Xue, Y., Yu, X. & Zhang, K. (2015), ‘Economic links and credit spreads’, *Journal of Banking & Finance* **55**, 157–169.
- Georg, C.-P. (2013), ‘The effect of the interbank network structure on contagion and common shocks’, *Journal of Banking & Finance* **37**, 2216–2228.
- Härdle, W. K. & Simar, L. (2015), *Applied Multivariate Statistical Analysis*, Springer.
- Härdle, W. K., Wang, W. & Yu, L. (2016), ‘Tenet: Tail-event driven network risk’, *Journal of Econometrics* **192**, 499–513.
- Hasman, A. (2013), ‘A critical review of contagion risk in banking’, *Journal of Economic Surveys* **27**, 978–995.
- Li, Y. & Zhu, J. (2008), ‘ l_1 -norm quantile regression’, *Journal of Computational and Graphical Statistics* **17**(1), 1–23.
- Markowitz, H. (1952), ‘Portfolio selection’, *Journal of Finance* **7**(1), 77–91.
- May, R. M. & Arinaminpathy, N. (2010), ‘Systemic risk: the dynamics of model banking systems’, *Journal of The Royal Society Interface* **7**, 823–838.
- Poshakwale, S. S. & Mandal, A. (2016), ‘Sources of time varying return comovements during different economic regimes: Evidence from the emerging indian equity market’, *Review of Quantitative Finance and Accounting* pp. 1–34.

- Rapach, D. E., Strauss, J. K., Tu, J. & Zhou, G. (2016), Industry interdependencies and cross-industry return predictability, Technical report, Working papers//Washington University in St. Louis, Olin School of Business.
- Rodriguez, J. C. (2007), ‘Measuring financial contagion: A copula approach’, *Journal of Empirical Finance* **14**(3), 401–423.
- Schweitzer, F., Fagiolo, G., Sornette, D., Vega-Redondo, F., Vespignani, A. & White, D. R. (2009), ‘Economic networks: The new challenges’, *Science* **325**, 422–425.
- Tibshirani, R. (1996), ‘Regression shrinkage and selection via the lasso’, *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1), 267–288.
- Zhu, X., Pan, R., Li, G., Liu, Y. & Wang, H. (2016), ‘Network vector autoregression’, *Annals of Statistics* .
- Zhu, X., Wang, W., Wang, H. & Härdle, W. K. (2016), Network quantile autoregression, SFE 649 Discussion Paper 2016-050, Sonderforschungsbereich 649, Humboldt Universität zu Berlin, Germany. available at <http://sfb649.wiwi.hu-berlin.de/papers/SFE649DP2016-050.pdf>.


6 Appendices

6.1 Tables

Code	Full name	Code	Full name
Agric	Agriculture	Drugs	Pharmaceutical Products
Food	Food Products	Chems	Chemicals
Soda	Candy & Soda	Rubbr	Rubber & Plastic Products
Beer	Beer & Liquor	Txtls	Textiles
Smoke	Tobacco Products	BldMt	Construction Materials
Toys	Recreation	Cnstr	Construction
Fun	Entertainment	Steel	Steel Works Etc
Books	Printing and Publishing	FabPr	Fabricated Products
Hshld	Consumer Goods	Mach	Machinery
Clths	Apparel	ElcEq	Electrical Equipment
Hlth	Healthcare	Autos	Automobiles and Trucks
MedEq	Medical Equipment	Aero	Aircraft
Ships	Shipbuilding, Railroad Equipment	LabEq	Measuring and Control Equipment
Guns	Defense	Paper	Business Supplies
Gold	Precious Metals	Boxes	Shipping Containers
Mines	Non-Metallic & Industrial Metal	Mining Trans	Transportation
Coal	Coal	Whsl	Wholesale
Oil	Petroleum and Natural Gas	Meals	Restaurants, Hotels, Motels
Util	Utilities	Banks	Banking
Telcm	Communication	Insur	Insurance
PerSv	Personal Services	REst	Real Estate
BusSv	Business Services	Fin	Trading
Hardw	Computers	Other	Almost Nothing
Softw	Computer Software	Rtail	Retail
Chips	Electronic Equipment		


Table 1: 49 industry portfolios from French's data library

τ	<i>cycle</i>	<i>middleman</i>	<i>in</i>	<i>out</i>
0.05	0.013	0.016	0.016	0.013
0.50	0.003	0.006	0.006	0.003
0.95	0.015	0.019	0.020	0.014

Table 2: Four connectedness measures of whole sample industry network
 INET-connectedness


τ	Leading industries	
	receivers	emitters
0.05	Gold(44)(0.293), Softw(3)(0.284), Smoke(26)(0.279), Hardw(10)(0.262), Fin(12)(0.248)	Insur(8)(0.378), Books(32)(0.355), Autos(21)(0.287), Bank(1)(0.252), Toys(42)(0.241), Coal(48)(0.228), Agric(41)(0.209), Aero(24)(0.209), Hlth(34)(0.203)
0.50	Gold(44)(0.612), Smoke(26)(0.238), Meals(25)(0.234), Coal(48)(0.216)	Coal(48)(0.624), Fin(12)(0.342), Ships(46)(0.289), Clths(31)(0.246), Hlth(34)(0.246)
0.95	Softw(3)(0.356), Guns(39)(0.281), Gold(44)(0.249), Steel(33)(0.224), Agric(41)(0.222), Hlth(34)(0.219)	Ships(46)(0.506), Drugs(2)(0.229), Other(11)(0.225), Coal(48)(0.214), Boxes(40)(0.213)

Table 3: Top central industries under various stress situations (the number in the first (second) parentheses is the rank of industry size (eigenvector centrality score))

 INET-eigcentr


τ	No. of Coefficients	No. of Nonzeros	No. of Negatives	No. of Positives	Max	Min	Average
0.05	2401	498	158	340	0.517	-0.477	0.007
0.50	2401	218	86	132	0.228	-0.108	0.001
0.95	2401	658	348	310	0.367	-0.729	-0.003

Table 4: Summary predictability magnitude and directions - whole sample

 INET-totaldire

τ	No. of Coefficients	No. of Nonzeros	No. of Negatives	No. of Positives	Max	Min	Average
0.05	2401	579	218	361	0.535	-0.618	0.007
0.50	2401	255	99	156	0.199	-0.150	0.001
0.95	2401	718	396	321	0.575	-0.590	-0.003

Table 5: Summary predictability magnitude and directions - rolling window

 INET-rollingdire

Strategies	$\tau = 0.05$				$\tau = 0.50$			
	HI	HO	TI	TO	HI	HO	TI	TO
Excess Returns	0.094**	0.088**	0.074*	0.079**	0.059	0.089***	0.084*	0.065*
<i>t-stat</i>	2.157	2.050	1.678	2.068	0.929	2.761	1.855	1.756


Strategies	$\tau = 0.95$				MKT
	HI	HO	TI	TO	
Excess Returns	0.088*	0.095***	0.065*	0.072	0.073
<i>t-stat</i>	1.911	2.742	1.853	1.577	1.328

Table 6: Excess returns of centrality-based trading strategy and market portfolio. *, **, *** denoting the 10%, 5% and 1% significance level respectively.

 INET-nettrading

Strategies	intercept	Mkt-Rf	SMB	HML	R^2	Adjusted R^2
HI-lowertail	0.002	0.978***	0.125***	0.147***	0.791	0.788
HO-lowertail	-0.000	1.020***	0.085**	0.467***	0.835	0.833
TI-lowertail	-0.002	1.045***	0.228**	0.564***	0.875	0.874
TO-lowertail	0.000	0.970***	0.095***	0.331***	0.890	0.888
HI-median	-0.002*	1.148***	0.155***	0.153***	0.886	0.885
HO-median	0.000	1.048***	0.173***	0.359***	0.810	0.808
TI-median	0.001	0.983***	0.147***	0.321***	0.797	0.795
TO-median	0.000	0.968***	0.126***	0.279***	0.885	0.883
HI-uppertail	0.001	0.968***	0.251***	0.285***	0.768	0.766
HO-uppertail	0.001	1.001***	0.169***	0.339***	0.853	0.851
TI-uppertail	0.001	0.870***	-0.042	0.334***	0.842	0.840
TO-uppertail	0.000	0.979***	0.070**	0.284***	0.888	0.886

Table 7: Coefficients of excess portfolio returns regressing on Fama-French risk factors. *, **, *** denoting the 10%, 5% and 1% significance level respectively.

 INET-regression

6.2 Figures

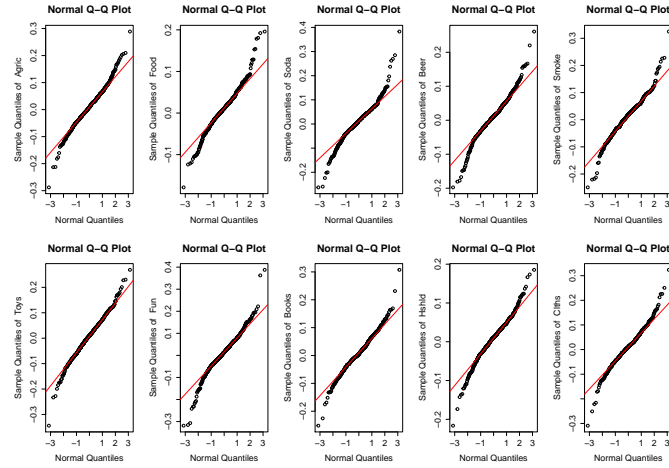


Figure 1: QQ plots of the industries

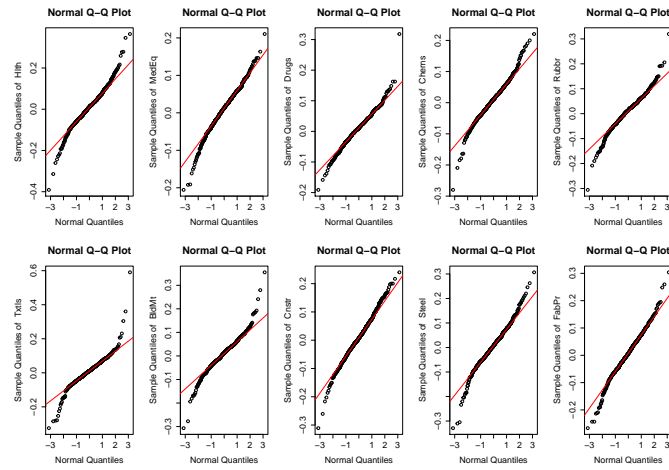
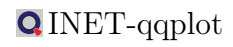


Figure 2: QQ plots of the industries

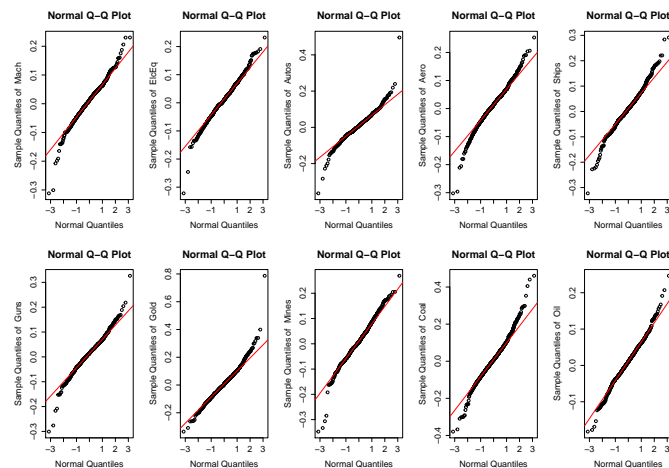
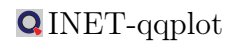
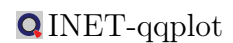


Figure 3: QQ plots of the industries



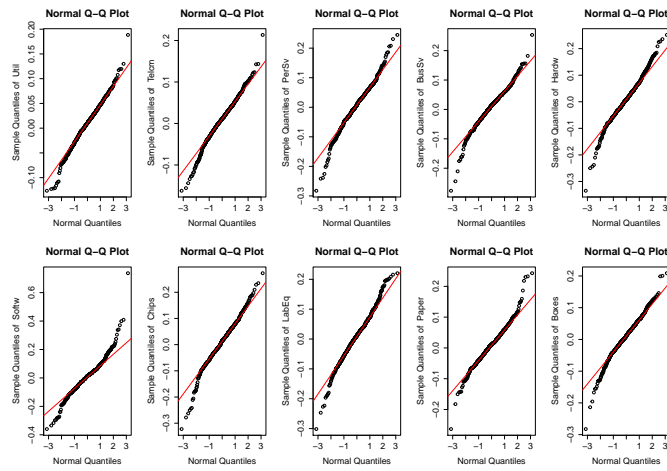


Figure 4: QQ plots of the industries

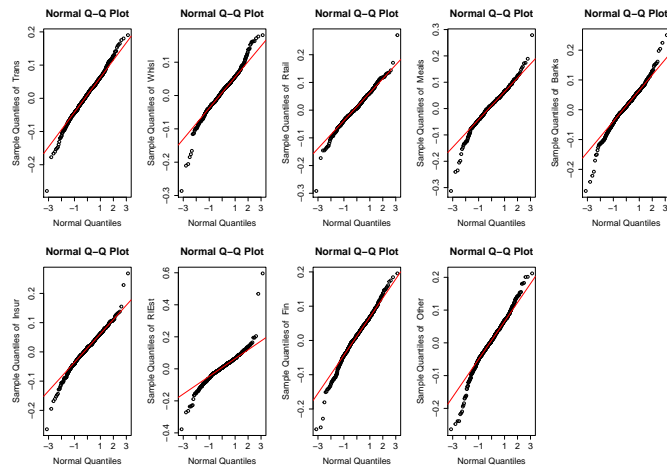
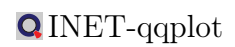
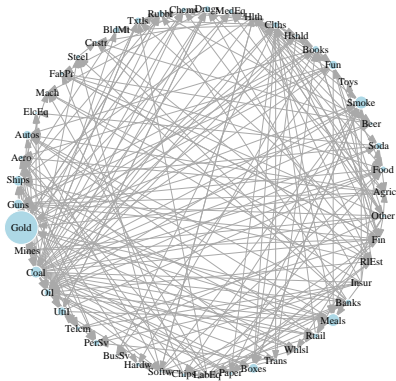


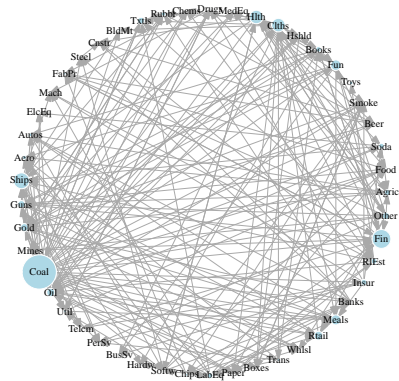
Figure 5: QQ plots of the industries



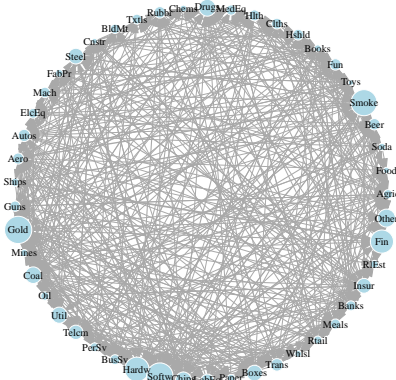
Whole sample industry network_ tau = 0.50 (197001–201701)



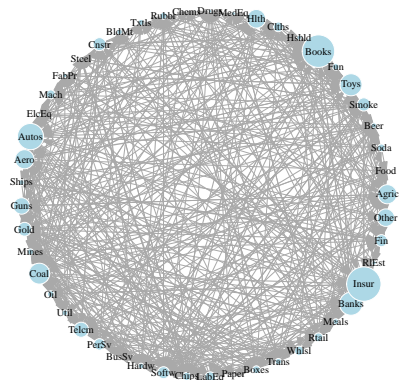
Whole sample industry network_ tau = 0.50 (197001–201701)



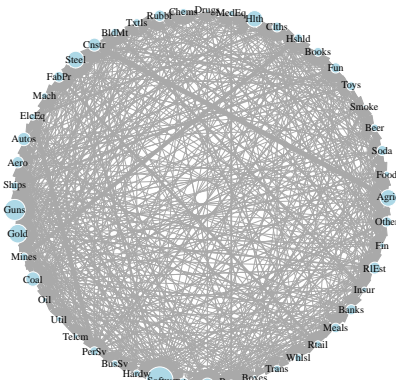
Whole sample industry network_ tau = 0.05 (197001–201701)



Whole sample industry network_ tau = 0.05 (197001–201701)



Whole sample industry network_tau=0.95(197001–201701)



Whole sample industry network_tau=0.95(197001–201701)

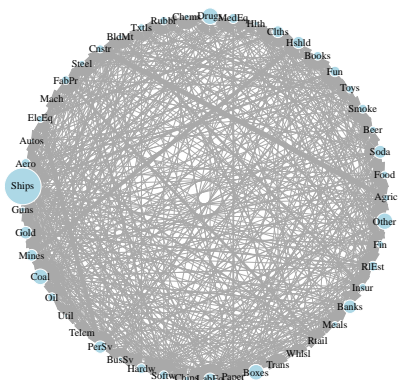


Figure 6: Whole sample network of industry portfolios with larger size denoting the eigenvector centrality ('in'-left, 'out'-right) - $\tau = 0.50$ (top), $\tau = 0.05$ (middle), $\tau = 0.95$ (bottom)

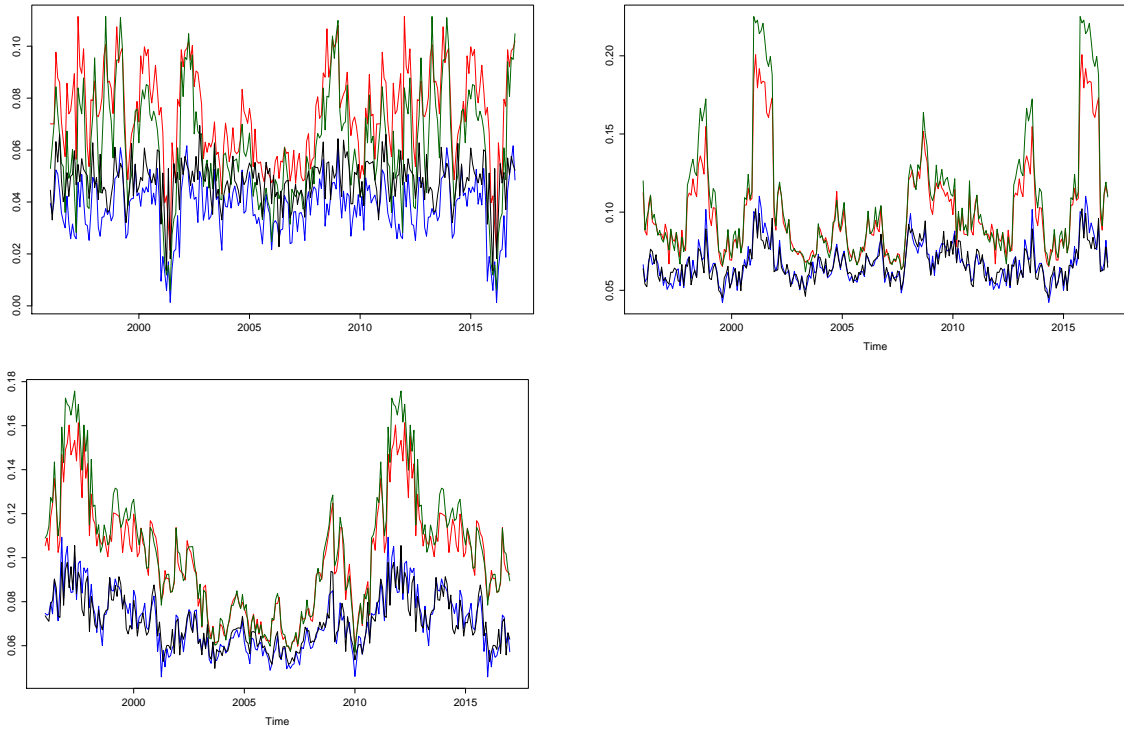


Figure 7: The **cycle**, **middleman**, **in** and out connectedness of industry networks: upperleft($\tau = 0.50$), upperright($\tau = 0.05$), lower($\tau = 0.95$)

 INET-connectedness

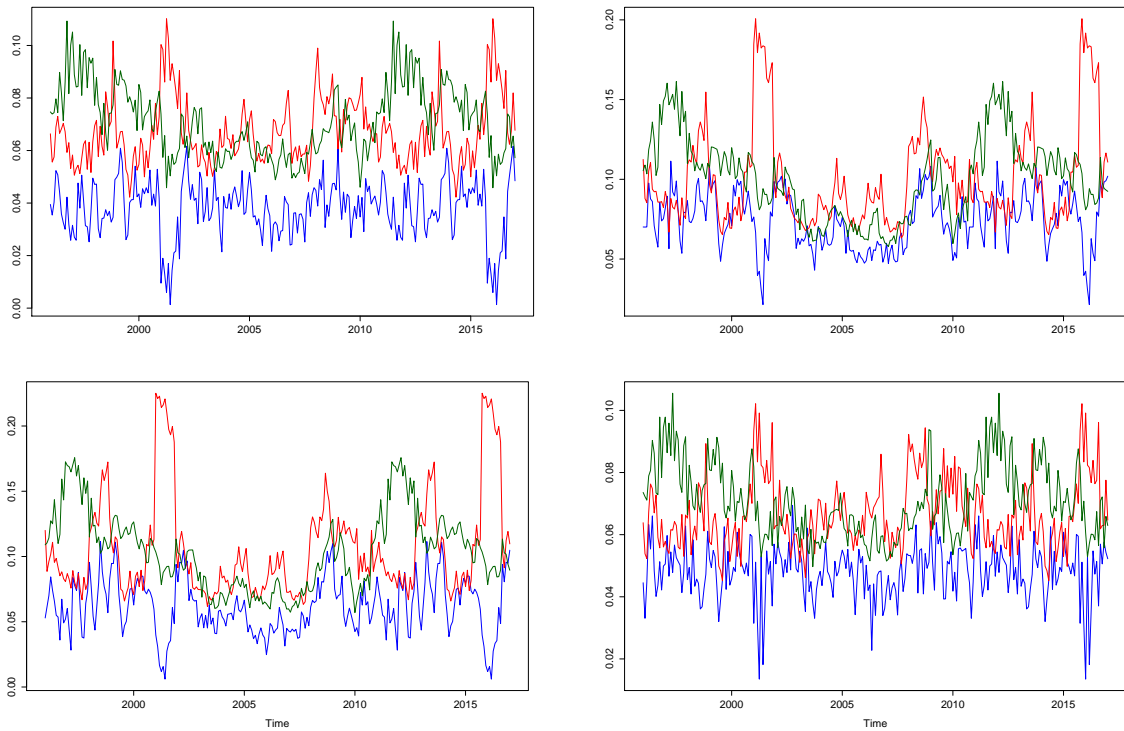



Figure 8: The comparison of industry network connectedness under $\tau = 0.50$, $\tau = 0.05$ and $\tau = 0.95$: upperleft(cycle), upperright(middleman), lowerleft(in), lower(right(out)

 INET-connectedness

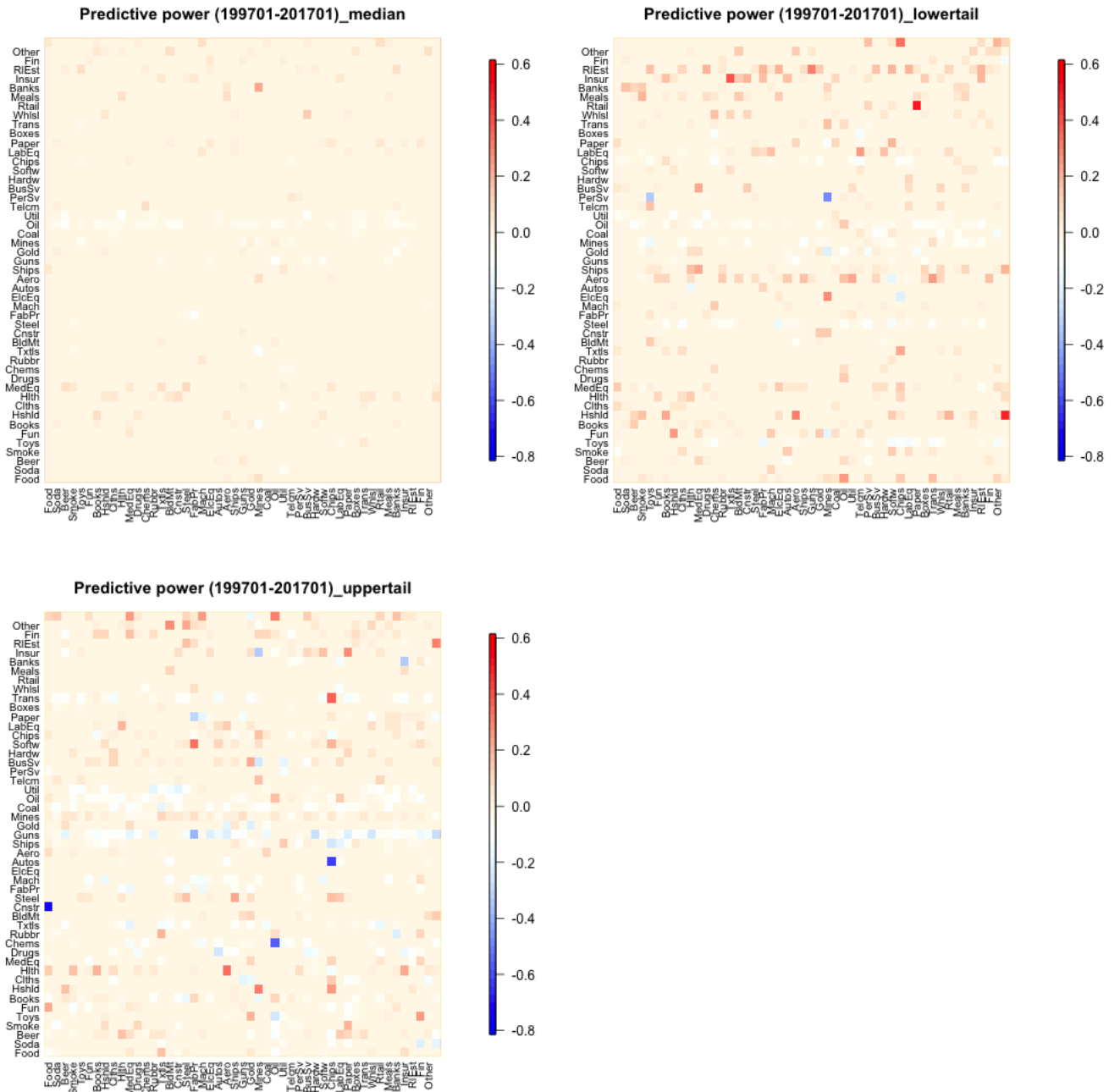


Figure 9: Connectedness across 49 industries - $\tau = 0.50$ (upperleft), $\tau = 0.05$ (upper-right), $\tau = 0.95$ (lower)

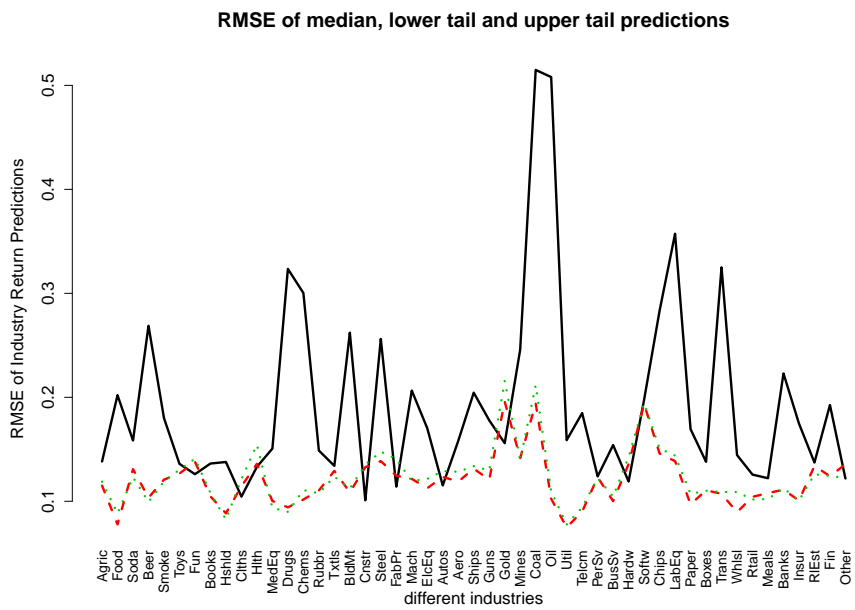



Figure 10: RMSE of industry return prediction over 253 months for $\tau = 0.05$, $\tau = 0.50$ and $\tau = 0.95$

 INET-RMSE



Risk-Constrained Kelly Portfolios Under Alpha-Stable Laws

Niels Wesselhöfft¹ · Wolfgang K. Härdle^{1,2,3}

Accepted: 2 August 2019 / Published online: 19 August 2019
© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract

This paper provides a detailed framework for modeling portfolios, achieving the highest growth rate under risk constraints such as value at risk (VaR) and expected shortfall (ES) in the presence of α -stable laws. Although the maximization of the expected logarithm of wealth induces outperforming any other significantly different strategy, the Kelly criterion implies larger bets than a risk-averse investor would accept. Restricting the Kelly optimization by spectral risk measures, the authors provide a generalized mapping for different measures of growth and risk. Analyzing over 30 years of S&P 500 returns for different sampling frequencies, the authors find evidence for leptokurtic behavior for all respective sampling frequencies. Given that lower sampling frequencies imply a smaller number of data points, this paper argues in favor of α -stable laws and its scaling behavior to model financial market returns for a given horizon in an i.i.d. world. Instead of simulating from the class of elliptically α -stable distributions, a semiparametric scaling approximation, based on hourly NASDAQ data, is proposed. Our paper also uncovers that including long put options into the portfolio optimization, improves portfolio growth for a given level of VaR or ES, leading to a new Kelly portfolio providing the highest geometric mean.

Keywords Growth-optimal · Kelly criterion · Protective put · Portfolio optimization · Stable distribution · Value at risk · Expected shortfall

JEL Classification C13 · C46 · C61 · C73 · G11

Financial support from the Deutsche Forschungsgemeinschaft via International Research Training Group 1792 “High Dimensional Nonstationary Time Series”, Humboldt-Universität zu Berlin, is gratefully acknowledged.

✉ Niels Wesselhöfft
wesselhn@hu-berlin.de

¹ Humboldt-Universität zu Berlin, IRTG 1792, Dorotheenstr. 1, 10117 Berlin, Germany

² School of Business, Singapore Management University, 50 Stamford Road, Singapore 178899, Singapore

³ Faculty of Mathematics and Physics, Charles University, Ke Karlovu 3, 121 16 Prague, Czech Republic

1 Introduction

Given a set of investment opportunities, how should the investment weights be chosen in order to have more wealth than any other investor at the end of the investment period? The Kelly growth-optimum strategy is a betting scheme for an investor, who seeks to asymptotically maximize his growth rate of capital. This strategy outperforms any other significantly different strategy, given knowledge of the true underlying process (Breiman 1961). But, the sole use of the Kelly criterion implies larger bets than a representative, risk-averse investor would accept in terms of risk (Clark and Ziemba 1987; Hausch and Ziemba 1985). Thus, the Kelly optimization needs to be restricted by a risk measure. We use α -stable laws and its scaling behavior in order to model the underlying financial market returns. Upon the Generalized Central Limit Theorem (GCLT), the horizon distribution is modelled in an discrete i.i.d. framework.

The aim is to maximize the geometric portfolio return, i.e. Kelly Criterion and restrict the objective to a subjective risk constraint, formulated as spectral risk measure, including quantile (VaR) or expected shortfall as special cases. The formulated trade-off introduces a mapping over growth and risk in order to evaluate the investment decision. The contribution of this paper is three-fold: The first contribution represents the application of multidimensional α -stable laws, in the form of elliptically α -stable distributions, to the constrained Kelly portfolio. Second, instead of simulating from the class of elliptically α -stable distributions, a semiparametric scaling approximation, based on the data set itself, is proposed. Third, assets with non-linear payoff structure, long put-options, are incorporated into the nonlinear optimization to allow for asymmetric payoffs, which lead to a higher growth criterion, given a fixed risk constraint.

The Kelly criterion originates from Kelly (1956), dealing with, from the point of information theory, an optimal investment strategy in a binary channel. Breiman (1961) formally proves the asymptotic outperformance of the Kelly strategy for arbitrary distributions in an i.i.d. world. For arbitrarily distributed, possibly non-stationary processes, those results have been extended by Algeot and Cover (1988). Incorporating risk measures into the Kelly optimization, MacLean et al. (1992) discuss the growth-risk trade-off in terms of efficiency. Roll (1973) compares the Markowitz arithmetic mean maximization with the Kelly geometric mean maximization. In contrast to Constant Proportion Portfolio Insurance (CPPI), the investment strategy remains fixed fraction, given stationarity. More recently, Busseti et al. (2016) introduce an alternative risk constraint, limiting the probability of a drawdown of wealth to a given undesirable level.

The distribution of financial market returns for a chosen horizon is modelled as the sum of hourly random variables. As the distribution in some horizon is presumed to be non-Gaussian, the classical Central Limit Theorem (CLT) does not apply as second and higher moments may not exist. Thus, the generalized central limit theorem (GCLT) of Gnedenko and Kolmogorov (1954) is applied for the sum of random variables, whose second and higher moments may not be bounded. For the financial application this implies the use of α -stable laws (Fama 1965; Lévy 1925; Mandelbrot 1963). As multidimensional α -stable random variables are difficult to evaluate for larger dimen-

sions, elliptical α -stable distributions are employed, allowing for efficient portfolio estimation for dimensions $k \leq 40$ (Nolan 2013) in the presence of linear dependence.

Price data, both for assets with linear and non-linear payoff structure, were gathered from Lobster and Bloomberg. For computation, Matlab 2016a was utilized. In order to solve the formulated nonlinear optimization problem the sequential quadratic algorithm in *fmincon* was employed.

The paper is organized as follows: In chapter one the portfolio allocation problem is stated. The financial model is formulated by using generalized measures for growth and risk. Chapter two, the estimation, starts with a case for non-Gaussianity of financial log-returns of different sampling frequencies, reasoning the utilization of α -stable laws. For the multidimensional case, elliptically α -stable distributions are introduced in order to have an analytically tractable class of distributions. As the semiparametric scaling approximation is introduced, the estimation of location and scale is illustrated. An application is given in chapter three, the implementation. For a representative investor with a planning horizon of one year, the optimally VaR/ES-constrained Kelly portfolios are found, benefitting from the protective put strategy.

2 Model

2.1 Portfolio Allocation Problem

Given initial wealth of the investor $W_0 \in \mathbb{R}^+$, there are $j = 1, \dots, k$ investment opportunities with fractions $f_t = [f_{1,t}, \dots, f_{k,t}]^T \in \mathbb{R}^k$ in period $t = 1, 2, \dots, T$. $T \in \mathbb{N}^+$ represents the planning horizon. Assessing solely self-financing strategies, the budget constraint is given by $\sum_{j=1}^k f_{j,t} \leq 1$. Given a statistical model for continuous returns $X_t \in \mathbb{R}^k$, discrete returns are calculated by $\tilde{X}_t = \exp \{X_t\} - 1$. Given outcomes in $t = 1, \dots, T$ the wealth in T is given by

$$\begin{aligned}
 W_T(f_t) &= W_0 \prod_{t=1}^T \left\{ 1 + \sum_{j=1}^k f_{j,t} \tilde{X}_{j,t} \right\} \\
 &= W_0 \prod_{t=1}^T \left\{ 1 + f_t^T \tilde{X}_t \right\}.
 \end{aligned}
 \tag{1}$$

Given the stochastic wealth process, measures for growth and risk are formulated in order to choose investment fractions f_t , which suit investor preferences.

For a cdf $F_{W_T}(x)$ the spectral risk/growth measure with weight function $\phi(x)$ is defined through the quantile function $F_{W_T}^{-1}(x) \stackrel{\text{def}}{=} \{x : P(W_T(f_t) \leq x) = \alpha\}$, $\alpha \in (0, 1)$.

$$M_\phi \{W_T(f_t)\} = \int_0^1 \phi(x) F_{W_T}^{-1}(x) dx
 \tag{2}$$

Within the context of spectral risk measures, the measure will be coherent iff the weight function is positive $\phi(x) \geq 0$, increasing $\phi'(x) \geq 0$ and normalized $\int_0^1 \phi(x) =$

1 (Acerbi 2002). For the discrete framework (1) with $n \in \mathbb{N}^+$ wealth trajectories, the measure is defined as

$$M_\phi \{W_T(f_t)\} = \sum_{i=1}^n \phi_i W_{T,i}(f_t), \tag{3}$$

where $W_{T,i}$ denotes element i out of n wealth paths with weight ϕ_i .

Growth measures

Following Roll (1973), there are two main strands dealing with the accumulation of wealth and thus, the allocation of wealth into a portfolio. On the one hand, the Markowitz optimization aims to maximize the expected portfolio return (Lintner 1965; Markowitz 1952; Sharpe 1964; Tobin 1958). On the other hand, the Kelly growth-optimum approach by Kelly (1956), Breiman (1961) and Thorp (1971), aims to maximize the expected logarithm of wealth, which is equivalent to maximizing the geometric portfolio return. Within the framework of spectral growth/risk measures, the growth measures for the Markowitz and the Kelly optimization are evaluated:

- G_1 : For the expected wealth, the growth criterion from the Markowitz optimization, the weight function is

$$\phi_E(x) = 1,$$

giving

$$G_{\phi_E} \{W_T(f_t)\} = \int_0^1 F_{W_T}^{-1}(x) dx = E \{W_T(f_t)\}. \tag{4}$$

- G_2 : The expected logarithm of wealth, representing the optimization criterion for the Kelly strategy, is obtained for the weight function

$$\phi_{E\log}(x) = \log(x),$$

giving

$$G_{\phi_{E\log}} \{W_T(f_t)\} = \int_0^1 \log F_{W_T}^{-1}(x) dx = E \{\log W_T(f_t)\}. \tag{5}$$

The growth measure will be denoted by $G_\phi \{W_T(f_t)\}$ and the optimization for horizon T without risk constraints is formulated as

$$\max_{f_t \in \mathbb{R}^k} \left[G_\phi \{W_T(f_t)\} \mid \sum_{j=1}^k f_{j,t} \leq 1 \right]. \tag{6}$$

This paper focusses on the Kelly growth criterion as it represents a betting scheme for an investor, who seeks to asymptotically maximize his growth rate of capital. The betting strategy outperforms any other significantly different strategy asymptotically and minimizes the expected time to reach a goal (Algeot and Cover 1988; Breiman 1961). For a comprehensive treatment of the Kelly criterion, see MacLean et al. (2011). Whereas the maximization of the expected wealth in the Markowitz optimization, given favorable investment possibilities, always implies betting the entire

fortune, the maximization of the expected logarithm of wealth leads to one growth-optimal portfolio, which is not necessarily optimal in terms of the Markowitz portfolio (Thorp 1971). Accordingly Markowitz (1976) considers the Kelly portfolio to be the upper limit for a conservative investor. Furthermore, the log-optimal strategy is fixed fraction, independent of time (MacLean et al. 1992).

Risk measures

The sole use of the Kelly criterion implies larger bets than a representative, risk-averse investor would accept in terms of risk (Clark and Ziemba 1987; Hausch and Ziemba 1985). In order to formulate individual risk measures for different investors, the spectral risk measure from (2), denoted by $S_\phi \{W_T(f_t)\}$, will be used. Two specific risk measures to include the degree of risk-aversion into the portfolio optimization are quantile (Value at Risk) and conditional tail expectation (expected shortfall) constraints:

- S_1 : The quantile constraint (VaR) is a special case of the spectral risk measure from (2)

$$\phi_{Q_\alpha}(x) = \delta(x = \alpha), \quad \alpha \in (0, 1), \quad (7)$$

where $\delta(x = \alpha)$ is the Dirac delta function, well known to be a non-coherent risk measure. Further drawbacks of the quantile constraint are treated in Basak and Shapiro (2001). However, the quantile restriction allows to ask the investor specifically to name a fraction of his wealth he can accept to lose with probability $1 - \alpha$.

- S_2 : In contrast, Conditional Tail Expectation (ES) is a coherent risk measure representing the average loss beyond a given quantile constraint. Being a special case of the spectral measure, the weight function is given as

$$\phi_{\text{CTE}_\alpha}(x) = \alpha^{-1} \mathbf{1}(x < \alpha). \quad (8)$$

Growth-risk frontier

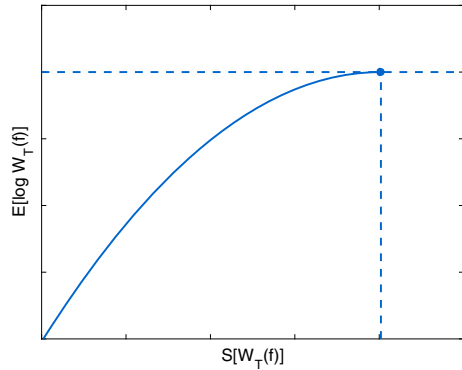
Following MacLean et al. (1992), the possible combinations of growth and risk measures are given by the set

$$U = [G_\phi \{W_T(f_t)\}, S_\phi \{W_T(f_t)\}], \quad f_t \text{ feasible}. \quad (9)$$

The growth-risk frontier is accordingly formulated as

$$U_t^* = [G_\phi \{W_T(f_t^*)\}, S_\phi \{W_T(f_t^*)\}], \quad f_t^* \text{ feasible}, \quad (10)$$

Fig. 1 Kelly-risk frontier with unconstrained Kelly portfolio exhibiting the highest geometrical mean



where the $f_t^* \in \mathbb{R}^k$ is the investment fraction maximizing the growth measure under risk restriction.

$$\begin{aligned}
 f_t^* &= \arg \max_{f_t \in \mathbb{R}^k} G_\phi \{W_T(f_t)\} \\
 \text{s.t. } &S_\phi \{W_T(f_t)\} \leq b, \quad b \in \mathbb{R}, \\
 &\sum_{j=1}^k f_{j,t} \leq 1
 \end{aligned}
 \tag{11}$$

For the Kelly criterion with a risk constraint as proposed, the frontier is illustratively visualized in Fig. 1. In contrast to the Markowitz maximization, implying a steady tradeoff between mean and risk, the geometric mean maximization implies one specific portfolio—the Kelly portfolio—exhibiting the highest geometric mean possible (horizontal dotted line). From this viewpoint, portfolios exhibiting a larger risk constraint than the Kelly portfolio (to the right of the vertical dotted line) are not efficient. If the investor prefers a smaller risk constraint than the full Kelly investor, restricted Kelly portfolios (solid line) constitute the Kelly-risk frontier. These are portfolio strategies with the highest growth criterion given risk constraint.

2.2 Tail Constraints and Non-linear Instruments

The introduction of assets as nonlinear functions of the underlyings, derivatives, allows for controlling the asymmetry of the wealth distribution in such a way, that it will be skewed to the left. Albeit the distribution of the risk measure, the loss of the portfolio is limited by construction for high confidence levels. The instruments to achieve the asymmetric payoff profile are long put options. By construction, corridor options, as argued in the context of quantile constraints, are circumvented (Basak and Shapiro 2001). A simplified representation of the protective put strategy is given in Fig. 2, consisting of one stock (blue) and one long put option (green) with chosen strike (dotted black). The result is the protective put strategy (red). The difference in payoff above the strike level is due to the put price, which the option holder has to pay. For

Fig. 2 Protective put strategy (red) consisting of long stock (blue) and long put (green) with chosen strike (dotted black). (Color figure online)

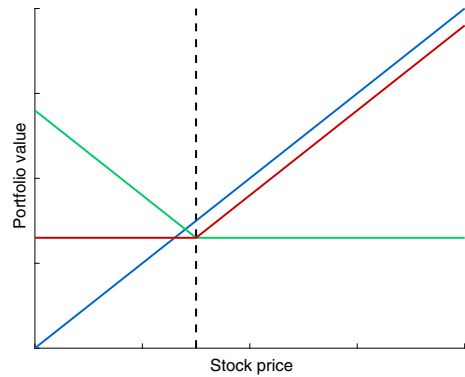


Table 1 Log-return descriptives for the different sampling frequencies, S&P 500 1985–2015 Frequency weekly* omits one week in the financial crisis 2009

Descriptives	S&P (daily)	S&P (weekly)	S&P (weekly*)	S&P (monthly)	S&P (yearly)
Data points	7564	1513	1512	360	30
Mean (p.a. in %)	8.37	8.37	10.41	8.37	8.37
Std (p.a. in %)	18.35	20.93	17.96	16	16.61
Skewness	-1.29	-6.7	-1.27	-1.98	-1.78
Kurtosis	31.26	131.09	13.67	12.48	7.13

$k \in \mathbb{N}^+$ linear assets with multiple put options each, given a pre-specified horizon, the choice of the fraction of linear and nonlinear assets is not obvious.

3 Estimation

3.1 A Case for Non-Gaussianity

Although Fama (1965) finds evidence for α -stable characteristics for all returns of the Dow Jones Index, it can be observed that financial (log-)returns tend to the Gaussian distribution as the sampling frequency decreases, see also McFarland et al. (1982), Boothe and Glassman (1987), and Dacorogna et al. (2001). The subsequent textbook example for the Standard and Poor's 500 reads as Table 1. Due to the 2009 financial crisis, an outlier week of -60% increases (decreases) the sample kurtosis (skewness) for the weekly frequency significantly from 13.67 (-1.27) to 131.09 (-6.7). If the outlier week is omitted, see column S&P (weekly*) of Table 1, the general observation of decreasing kurtosis and increasing negative skewness is supported for *different* sample sizes. Still, including the outlier week of 2009, erratic behavior of sample moments definitely appears for this reference data series.

The empirical observation of Gaussian convergence for lowering sampling frequencies cannot be shown explicitly by existing data, as data-records capture only 7564

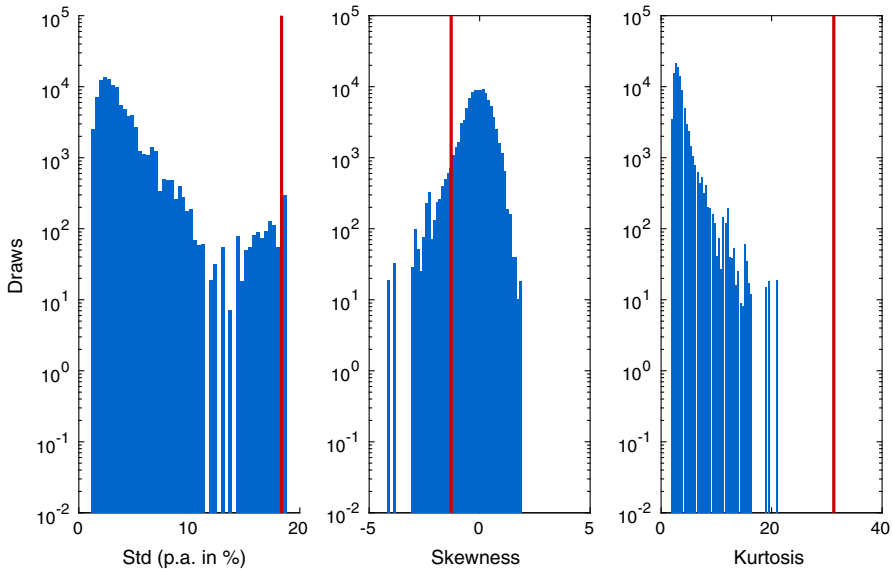


Fig. 3 Whole sample (red) and block-bootstrapped standard deviations (p.a.), skewness and kurtosis for 10^5 draws of 30 subsequent daily returns (blue) from the S&P 500, 1985 to 2015. (Color figure online)

trading days, representing 30 years of data. The empirical verification would require an appropriately large number of weeks, months and years.

In order to show that the annual return distribution, consisting of 30 data points, is with large probability not Gaussian, we randomly sample 10^5 30 blocks of daily returns from the S&P 500 and calculate second, third and fourth moments in order to evaluate dispersion, skewness and leptokurtic behavior (Bootstrap). Hence, for the three moments, the block-bootstrap estimators are plotted as histogram in Fig. 3. The vertical red lines represent the moment estimators for the whole daily data series. In essence, dispersion, skewness and especially leptokurtic behavior of the bootstrap estimators are significantly biased, compared to the estimator of the whole series. Fewer sampled data-points imply less probability of sampling data in the tails of the return distribution. There was not one out of 10^5 30 day sub-samples, which resulted in a comparable kurtosis of the complete data-series. The result holds for sampling 30 separate days randomly under the i.i.d. assumption. Moments of order larger than one behave erratically over an increasing data sample, as first analyzed for commodity prices in Mandelbrot (1963). Figure 4 plots standard deviation (in %), skewness and kurtosis as function of the used data-points of the series. The red lines represents the empirical moment behavior with increasing daily data points. The blue lines represent 100 trajectories of Gaussian moments with increasing data points. The observation of erratic moment behavior stands in contrast to Gaussian behavior. The observation holds over sampling frequencies daily, weekly, monthly and annually. This specific sample-size problem is crucial in risk management, especially for estimating quantiles of high confidence of the wealth distribution as in the constrained portfolio optimization in (11). As the confidence level tends to one, having only a limited amount of data,

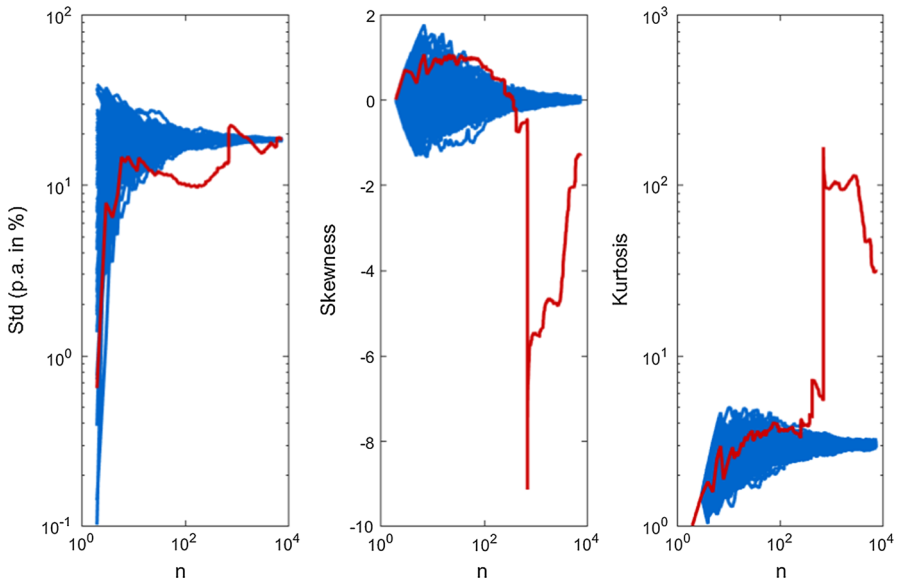


Fig. 4 (Log-)Log plots of standard deviation (in %), skewness and kurtosis with increasing data points, S&P 500 from 1985 to 2015 (red) and 100 Gaussian simulations with S&P 500 moments (blue). (Color figure online)

the quantile estimate is systematically biased as the quantile is overestimated. The portfolio analyst has to evaluate if the estimated quantile given the chosen confidence level still has an acceptable distribution.

Consequently, for investors with longer investment horizons, such as a year, the sum of daily random variables, constituting the yearly distribution, should not converge to the Gaussian, but to a heavy-tailed distribution, which will turn out to be the class of α -stable distributions. For financial markets, this assumption will imply infinite variance, skewness and kurtosis, leading to non-converging moments, i.e. the observed erratic behavior. The model of Sect. 2 will be estimated within a stationary framework for elliptically α -stable distributions, striving for scale invariance. Although daily and higher frequency returns exhibit non-stationary characteristics, the horizon distribution, i.e. yearly, cannot be shown to exhibit significant volatility clustering.

3.2 Scale Invariance

Let $X_t \in \mathbb{R}^k$ be a multidimensional, i.i.d. random variable from distribution P^t , where t indicates the scale e.g. days. Given the investment horizon of the investor, T days, the wealth equation of (1)

$$W_T(f_t) = W_0 \prod_{t=1}^T \left\{ 1 + f_t^\top \tilde{X}_t \right\} = W_0 \prod_{t=1}^T \left\{ f_t^\top \exp(X_t) \right\} \tag{12}$$

can be simplified, given $f_t = f \forall t = 0, \dots, T$.

$$W_T(f) = W_0 \left\{ f^\top \exp \left(\sum_{t=1}^T X_t \right) \right\} = W_0 \left\{ f^\top \exp(X) \right\}, \quad X \stackrel{\text{def}}{=} \sum_{t=1}^T X_t \quad (13)$$

As the horizon T grows, the sum of the random variables X_t tends to the Gaussian as long as the first two moments of the underlying distribution are finite. Formally, let random variable X_t have expectation vector $\mu_t = E(X_t)$ and covariance matrix $\sigma_t = E[\{X_t - E(X_t)\}\{X_t - E(X_t)\}^\top]$. Then

$$\begin{aligned} \sum_{t=1}^T X_t &\xrightarrow{\mathcal{L}} N \left(\sum_{t=1}^T \mu_t, \sum_{t=1}^T \sigma_t \right) \\ X &\xrightarrow{\mathcal{L}} N(\mu, \sigma) \\ T^{-\frac{1}{2}} \sum_{t=1}^T (X_t - \mu_t) &\xrightarrow{\mathcal{L}} N(0, \sigma). \end{aligned} \quad (14)$$

If the distribution in horizon T is modelled as the sum of higher frequency distributions, the multidimensional process of returns, which may not be Gaussian, but of finite variance, converges to the Gaussian. In contrast, as argued in Sect. 3.1, returns of horizons beyond the sampled frequency, are presumed to be heavy-tailed. Hence, the standard Central Limit Theorem (CLT) does not apply.

Except for the Gaussian itself, finite variance distributions change their shape under aggregation. In contrast, the class of α -stable distributions is scale invariant (Mandelbrot 1963). Scale invariance of distribution P is defined via a continuous function g , such that for all x

$$g(\lambda)P(x) = P(\lambda x), \quad (15)$$

with $\lambda x \geq x_0$ and $x_0 > 0$. Equivalently, distribution P has a power-law tail, implying that for $x \geq x_0 \geq 0, c \geq 0$ and $\alpha > 0$

$$P(x) = cx^{-\alpha}. \quad (16)$$

In that respect, a one-dimensional random variable $X \sim S(\alpha, \beta, \gamma, \delta)$ will be α -stable distributed with parameters $0 < \alpha \leq 2, -1 \leq \beta \leq 1, \gamma \geq 0$ and $\delta \in \mathbb{R}$ (Cizek et al. 2011; Nolan 2017), if

$$X \stackrel{\mathcal{L}}{=} \begin{cases} \gamma Z + \delta, & \alpha \neq 1 \\ \gamma Z + (\delta + \beta \frac{2}{\pi} \gamma \log \gamma), & \alpha = 1. \end{cases} \quad (17)$$

$S(Z \mid \alpha, \beta, 1, 0)$ represents the standard α -stable form. Only special cases of α -stable distribution are available as real-valued densities (e.g. Gaussian, Cauchy and Lévy).

Scale invariance under addition implies that for the sum of α -stable variables $X_t \sim S(\alpha, \beta, \gamma, \delta_t)$, $t = 1, \dots, T$

$$X_1 + X_2 + \dots + X_T = \sum_{t=1}^T X_t = X \sim S\left(\alpha, \beta, T^{\frac{1}{\alpha}}\gamma, \delta\right), \tag{18}$$

where $\delta = T\delta_t$.

According to Gnedenko and Kolmogorov (1954), the limiting distribution of T i.i.d. α -stable random variables, $0 < \alpha \leq 2$ is

$$a_T \left(\sum_{t=1}^T X_t \right) - b_T \xrightarrow{\mathcal{L}} S(\alpha, \beta, 1, 0), \tag{19}$$

where $a_T > 0$ and $b_T \in \mathbb{R}$. The special case of the Generalized Central Limit Theorem (GCLT) is the CLT of Eq. (14) for $\alpha = 2$, $\beta = 0$, $\gamma = \frac{\sigma}{\sqrt{2}}$ and $\delta_t = \mu_t$, given $a_T = \frac{1}{\sigma\sqrt{T}}$ and $b_T = \frac{\sqrt{T}\mu}{\sigma}$. In general, for $0 < \alpha \leq 2$,

$$T^{-\frac{1}{\alpha}} \sum_{t=1}^T (X_t - \delta_t) \xrightarrow{\mathcal{L}} S(\alpha, 0, \gamma, 0). \tag{20}$$

3.3 Elliptically Contoured α -Stable Distributions

For the multidimensional estimation, α -stable laws are not extensively accessible as closed-form densities are only available for special cases. One computationally tractable exception are elliptically contoured α -stable laws, which can be efficiently estimated for dimensions $k \leq 40$ (Nolan 2013). This class of distributions enables the modeling of heavy tails while preserving its shape under aggregation in the presence of linear dependence.

Random vector $Y = [Y_1, \dots, Y_k]^\top$ has a spherical distribution iff the characteristic function $\varphi_Y(u)$ satisfies for all $u \in \mathbb{R}^k$

$$\varphi_Y(u) = \mathbb{E} \left\{ \exp \left(iu^\top Y \right) \right\} = \psi(u^\top u) = \psi(u_1^2 + \dots + u_k^2), \tag{21}$$

where ψ is the characteristic generator of the spherical distribution.

Random vector $X \sim S_k(\delta, \Gamma, \psi)$ is elliptically distributed with positive definite scaling matrix $\Gamma = AA^\top$, $A \in \mathbb{R}^{k \times k}$ and location vector $\delta \in \mathbb{R}^k$ when

$$X \stackrel{\mathcal{L}}{=} \delta + AY, \tag{22}$$

where Y is spherical with characteristic generator ψ . The characteristic function is given by

$$\varphi_X(u) = \mathbb{E} \left\{ \exp \left(iu^\top X \right) \right\} = \exp \left(iu^\top \delta \right) \psi \left(u^\top \Gamma u \right). \tag{23}$$

A subclass of elliptical distributions are normal variance mixtures $X = [X_1, \dots, X_k]^T$ for

$$X \stackrel{\mathcal{L}}{=} W^{1/2}AZ + \delta, \tag{24}$$

with $Z \sim N(0, I_k)$ and $W \geq 0$ being a non-negative one-dimensional random variable, independent of Z (Kring et al. 2009).

A further subclass of normal variance mixtures are α -stable sub-Gaussian $X = [X_1, \dots, X_k]^T$ for $W \sim S(\alpha/2, (\cos \pi\alpha/4)^{2/\alpha}, 1, 0)$, $0 < \alpha < 2$, being one-dimensionally α -stable distributed, parameterized following Nolan (2017). $G \sim N(0, \Gamma)$ is multidimensional Gaussian with scaling matrix $\Gamma = AA^T$. Then $X \sim S_k(\alpha, \beta, \Gamma, \delta, \psi)$, $\beta = 0$ is α -stable sub-Gaussian if

$$\begin{aligned} X &\stackrel{\mathcal{L}}{=} W^{1/2}G + \delta \\ &\stackrel{\mathcal{L}}{=} W^{1/2}AZ + \delta, \quad Z \sim N(0, I_k) \\ &\stackrel{\text{def}}{=} AY + \delta, \end{aligned} \tag{25}$$

while $Y \sim S_k(\alpha, 0, I_k, 0)$ is radially symmetric α -stable. The according characteristic function of X is

$$\begin{aligned} \varphi_X(u) &= \int_{-\infty}^{\infty} f_X(x) \exp(iu^T X) dx = E\left(iu^T X\right) \\ &= \exp\left\{-\left(\frac{1}{2}u^T \Gamma u\right)^{\alpha/2} + iu^T \delta\right\}, \end{aligned} \tag{26}$$

$f_X(x)$ as probability density function. $\Gamma \in \mathbb{R}^{k \times k}$ is the positive definite scale matrix and $\delta \in \mathbb{R}^k$ the location vector. The characteristic generator is therefore given by

$$\psi(s, \alpha) = \exp\left\{-\left(\frac{1}{2}s\right)^{2/\alpha}\right\}. \tag{27}$$

This implies that α -stable sub-Gaussian distributions are scale mixtures of multivariate normal distributions (Samorodnitsky and Taqqu 1994). Note, for $\alpha = 2$, the characteristic function collapses to the Gaussian. For $G \sim N(0, I_k)$, the characteristic function of Y in Eq. (25) simplifies to

$$\varphi_Y(u) = E\left(iu^T Y\right) = \exp\left(-\gamma^\alpha |u|^\alpha\right). \tag{28}$$

For the horizon of the investor, T , the estimated higher sampling frequency log-returns are summed to the chosen frequency:

$$\tilde{X} = TX \sim S_k(\alpha, 0, T\Gamma, T\delta, \psi) \tag{29}$$

For subsequent estimation, stability parameter $0 < \alpha \leq 2$, scale Γ and location δ need to be estimated, given that $-1 \leq \beta \leq 1$ can be assumed to be not significantly different from zero.

3.4 Parameter Estimation

The utilization of α -stable laws implies that fractional moments of random variable X

$$E|X|^p = \int_{-\infty}^{+\infty} |X|^p f(X) dX \quad (30)$$

are finite for $0 < p < \alpha$, $p \in \mathbb{R}$ and infinite for $p \geq \alpha$. This implies that for the α -stable Paretian case, representing a slower decay than under the Gaussian, $0 < \alpha < 2$, the second moment $E|X|^2 = \infty$ and higher moments such as skewness and kurtosis are infinite. For the empirical financial market returns $1 < \alpha < 2$ (see Sect. 4), the first moment remains finite. For elliptically α -stable random variable $X \sim S_k(\alpha, 0, \Gamma, \delta, \psi)$ the expectation is

$$E X = \delta < \infty. \quad (31)$$

In general, for univariate α -stable laws the mean is undefined for $\alpha \leq 1$ and $E X = \delta - \beta\gamma \tan\left(\frac{\pi\alpha}{2}\right) < \infty$ for $\alpha > 1$. From the perspective of a data scientist, analyzing the sample, empirical moments are always finite. But under the assumptions of being α -stable distributed, fractional moments with $p \geq \alpha$ have no intrinsic meaning. As shown in Sect. 3.1 higher moments behave erratic with increasing data points, contrary to moment convergence under Gaussianity.

For portfolio allocation the estimation of location and scale are crucial. Founding on the analysis of Chopra and Ziemba (1993), the mean represents the largest source of error for estimating the portfolio fraction. Their final implication is straightforward: "[...] the bulk of resources should be spent on obtaining the best estimates of expected returns of the asset classes under consideration".

Simulating from the class of elliptically α -stable distributions implies to estimate the stability parameter α , scaling matrix Γ and location δ , given that the skewness parameter β is zero. For the characteristic exponent α the method of Rachev and Mittnik (2000) is used:

- i. Simulate U_1, \dots, U_n uniformly i.i.d. random variables on the unit hypersphere S^{k-1} .
- ii. Estimate the MLE for the index of stability $\hat{\alpha}_i$ (Nolan 2001) for each i from 1 to n , $U_i^\top X_1, \dots, U_i^\top X_n$.
- iii. Calculate the index of stability by $\hat{\alpha} = n^{-1} \sum_i^n \hat{\alpha}_i$.

By utilizing the MLE for the characteristic exponent α , severe estimation biases from e.g. the Hill estimator (Hill 1975) are circumvented, see also McCulloch (1997) and Kearns and Pagan (1997). For the proposed semiparametric scaling approximation in Sect. 3.5, the estimation of stability α will not be necessary.

Estimating the location vector $\delta \in \mathbb{R}^k$ of multidimensional variable $X \sim S_k(\alpha, 0, \Gamma, \delta, \psi)$ is of crucial importance for portfolio allocation, representing the driver for asset growth.

From the perspective of information theory, we aim to chose the parameter vector, which maximizes the probability of coming from the empirical data-set. From the perspective of decision theory, this method coincides with the minimization of expected loss under the 0–1 loss function:

$$L(\delta, \hat{\delta}) = 1(\delta \neq \hat{\delta}). \tag{32}$$

The according risk function is

$$R(\delta, \hat{\delta}) = E \left\{ L(\delta, \hat{\delta}) \right\} = E \left\{ 1(\delta \neq \hat{\delta}) \right\} \tag{33}$$

Consequently the optimization

$$\delta^* = \arg \min_{\delta \in \mathbb{R}^k} \left[E \left\{ 1(\delta \neq \hat{\delta}) \right\} \right] \tag{34}$$

leads to the common Maximum Likelihood Estimate (MLE). If the loss function is not presumed to be 0–1 loss, e.g. quadratic, the usual ML estimator may not be suitable. The inadmissibility of the sample mean under the Gaussian for dimensions $k > 2$ has been first shown by Stein (1955), leading to the class of shrinkage estimators, starting with James and Stein (1961). An overview over the class of shrinkage estimators is given in Hansen (2015). To our knowledge, those results have not been extended to α -stable laws.

Following Nolan (2013), there are two methods to estimate the scale matrix Γ :

- i. Given that X is elliptically α -stable,

$$\forall u, u^\top X \sim S_k \left(\alpha, 0, (u^\top \Gamma u)^{\frac{1}{2}}, u^\top \delta, \psi \right). \tag{35}$$

The $k(k + 1)/2$ parameters of the scale matrix Γ are estimated by

$$\begin{aligned} \hat{\Gamma}_{j,j} &= \hat{\gamma}_j^2 \\ \hat{\Gamma}_{j,i} &= \frac{1}{2} \left\{ \hat{\gamma}^2(1, 1) - \hat{\gamma}_i^2 - \hat{\gamma}_j^2 \right\}, \end{aligned} \tag{36}$$

where $\hat{\gamma}^2(1, 1) = (1, 1)^\top (X_j, X_i) = X_j + X_i$ and $\hat{\gamma}_j$ is the univariate scale ML estimate of asset j . Note that $\hat{\Gamma}_{j,i}$ depends solely on directions $(1, 1), (1, 0)$ and $(0, 1)$.

- ii. As $E \left\{ \exp(iu^\top X) \right\} = \exp \{-\gamma(u)^\alpha\}$

$$\left\{ -\log E \exp(iu^\top X) \right\}^{\frac{2}{\alpha}} = u^\top \Gamma u = \sum_i u_i^2 \Gamma_{i,i} + 2 \sum_{i < j} u_i u_j \Gamma_{i,j}, \tag{37}$$

so $\Gamma_{i,j}$ can be estimated as linear function via regression, taking more directions into account than the first method.

For the remainder of the paper, the first method is utilized due to its analytical tractability.

3.5 Semiparametric scaling approximation

Instead of simulating from the estimated elliptically contoured α -stable distribution, a semiparametric scaling approximation based on higher sampling frequency data is proposed. Simulating from the elliptically α -stable distribution implies that

$$\begin{aligned} \forall j = 1, \dots, k \quad \alpha_j &= \alpha, \quad 0 < \alpha < 2, \\ \forall j = 1, \dots, k \quad \beta_j &= 0. \end{aligned} \tag{38}$$

We deal with this drawback by using α -stable properties of the empirical data-set. Assume that the higher sampling frequency data-set $X_t \sim S_k(\alpha, 0, \Gamma_t, \delta_t, \psi)$ is elliptically α -stable distributed. Then,

- i. estimate location δ_t and scale $\Gamma_t = A_t A_t^\top$ of higher frequency returns X_t as proposed in section 3.4.
- ii. Normalize X_t to radially symmetric $Y \sim S_k(\alpha, 0, I_k, 0, \psi)$

$$Y = A_t^{-1} X_t - \delta_t. \tag{39}$$

- iii. Rescale radially symmetric Y to distribution $X \sim S_k(\alpha, 0, \Gamma, \delta, \psi)$, $\Gamma = A A^\top$ with investment horizon T ,

$$X = AY + \delta \tag{40}$$

with $\Gamma = T \Gamma_t$ and $\delta = T \delta_t$.

The resulting distribution for horizon T , represented by convoluted higher frequency distributions, is simply an affine transformation of its radially symmetric analogue, given its scaling nature. Given that $\beta = 0$, we can use the potentially different stabilities α_j of the marginals, having no effect on location δ and scale Γ .

As the horizon distribution represents a limited number of data points (see section 3.1), empirical quantiles Q_α , $\alpha < 0.02$ are overestimated, implying that risk measures for large confidence levels are underestimated. Vice versa, quantiles Q_α , $\alpha > 0.98$ are consequently underestimated, see Fig. 5. By using empirical higher sampling frequency data, we can scale high-frequency events to a manifold of large scale events, which never happened in the original data history of the lower sampling frequency, enriching the tails of the horizon distribution.

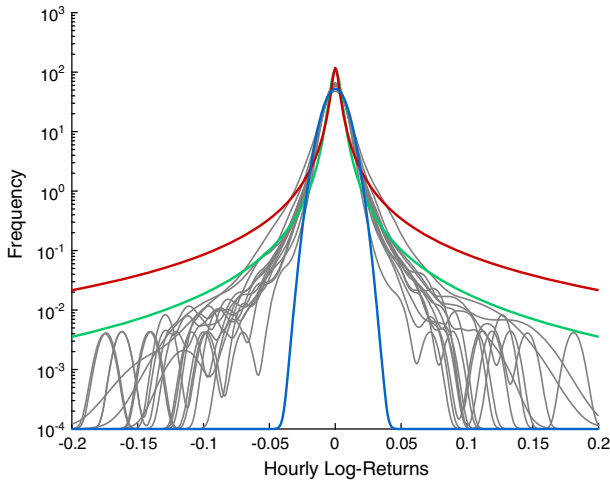


Fig. 5 Semi-log densities for hourly Apple log-returns for Gaussian, $\alpha = 2$ (blue), Stable, $\alpha = 1.33$ (green), Cauchy, $\alpha = 1$ (red) and financial assets (gray). (Color figure online)

4 Implementation

4.1 Data

The hourly financial stock prices come from Lobster and cover the time span from 2007-06-27 to 2018-05-25, representing 17862 hourly prices per asset. The $k = 14$ assets with a linear payoff structure (stocks) are the stocks with the biggest market capitalization in the NASDAQ 100, representing a technology driven portfolio. A risk-free asset, which can be bought with annual rate $r = 0.01$ is included into the optimization. Relevant asset statistics including Maximum Likelihood Estimates (MLE) under α -stability (Nolan 2001) are given in Table 2.

The assets with a non-linear payoff structure are represented as long put options, written on the stock market index NASDAQ 100. As will be assumed for the representative investor in section 4.4, the maturity, and hence the investment horizon T , is chosen to be one year. The prices coming from the ask implied volatilities of the long put options determine the price of the hedge and accordingly the reduction in wealth if the stocks close above the chosen strike levels. For the distribution of wealth in T , the put option price O_T at maturity is given by the inner value

$$O_T = \max \{0; K - S_T\}. \quad (41)$$

Solely for evaluating the price of the non-linear assets between $t = 1$ and horizon T , a pricing model is needed.

Table 2 Log-return descriptive statistics with Maximum Likelihood Estimates under α -stability, 2007-06-27 to 2018-05-25

	$\mu \times T$	$\sigma \times T^{1/2}$	Skewness	Kurtosis	α	β	$\gamma \times T^{1/\alpha}$	$\delta \times T$
Apple	0.22	0.31	-0.75	41.09	1.33	0.03	0.69	0.08
Adobe	0.16	0.32	-0.97	68.52	1.42	-0.05	0.55	0.33
Amgen	0.11	0.26	1.12	44.33	1.49	0.00	0.39	0.06
Amazon	0.29	0.38	1.79	62.00	1.40	0.03	0.67	0.14
Comcast	0.08	0.30	-0.16	29.31	1.43	-0.02	0.51	0.14
Costco	0.11	0.23	-0.64	32.68	1.44	0.03	0.38	0.04
Cisco	0.04	0.29	-0.94	62.65	1.44	-0.02	0.47	0.09
Gilead	0.11	0.30	-0.94	48.60	1.47	0.02	0.47	-0.01
Intel	0.08	0.28	-0.24	25.55	1.43	-0.01	0.51	0.07
Microsoft	0.11	0.27	-0.05	37.09	1.42	-0.01	0.47	0.09
Nvidia	0.16	0.47	-2.95	134.48	1.40	-0.01	0.86	0.20
Pepsi	0.04	0.18	-0.74	38.56	1.44	0.00	0.30	0.08
Qualcomm	0.03	0.31	-0.31	59.96	1.39	-0.01	0.58	0.07
Texas Instruments	0.10	0.28	-0.71	28.31	1.44	-0.04	0.49	0.25

4.2 Stable tests

In order to verify if the class of elliptically α -stable distributions is suitable for the financial assets, the following prerequisites have to be met:

- heavy tails beyond the Gaussian (Leptokurtic behavior),
- linear dependence structure between the margins,
- comparable range of α_j (for simulation),
- skewness parameter β not coherently different from zero.

As examined descriptively in Table 2, empirical financial market returns are significantly non-Gaussian. In Fig. 5, the densities of the normalized log-returns on log-scale are plotted for Gaussian ($\alpha = 2$), Stable ($\alpha = 1.33$), Cauchy ($\alpha = 1$) and the individual assets using Kernel Density Estimates. Within the α -stable framework all examined assets lie between Gaussian and Cauchy, $1 < \alpha < 2$. The α -stable fit for $\alpha = 1.33$ captures the tails adequately, although events are captured, which never took place in the data history. The range of characteristic exponents stands in line with results of Westerfield (1977), McCulloch (1997) or Nolan (2013). The elliptical behavior is assessed by using two dimensional scatter matrices of the empirical log-returns. The significance of the skewness parameters β_j is verified by the utilization of the Fisher information from the MLE. The respective confidence intervals for the individual parameters show that β_j are not consistently different from zero, given a confidence level of 99%. For larger dimensions, Nolan (2013) reaches the same conclusion for the Dow Jones constituents.

Making use of the semiparametric scaling approximation implies that there is no need to estimate one specific α for the elliptical α -stable distribution. As $\forall j 1 < \alpha_j < 2$ we can deny the null of Gaussianity coherently for the 99% confidence level, speaking in favour of the α -stable hypothesis. As we are interested in the horizon distribution, constituted by the sum of hourly random variables, the generalized CLT is utilized.

4.3 Stable estimation

Following section 3.4, the parameter estimates for the hourly distribution $X_t \sim E_k(\alpha, \beta, \Gamma_t, \delta_t, \psi)$, $\beta = 0$ are scaled to the chosen horizon of one year. Exemplary, the semi-log densities for yearly Apple log-returns under Gaussian, Cauchy, Stable and the semiparametric scaling are plotted in Fig. 6. Additionally, Gaussian scaling, representing the scaling of the hourly distribution utilizing the square root of time rule under Gaussianity whilst neglecting the CLT, is displayed. In comparison, the semiparametric scaling distribution exhibits heavier tails than under Gaussianity, implying stock market events, which never occurred in the history of the original sampling frequency. The utilized scaling approximation provides the horizon distribution $X_T \sim E_k(\alpha, 0, \Gamma_T, \delta_T, \psi)$ with location vector $\delta_T = T\delta_t$ and scaling matrix $\Gamma_T = T\Gamma_t$, given that $\beta = 0$.

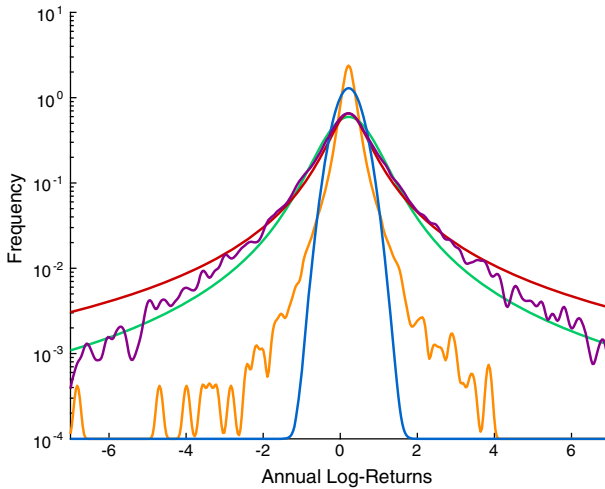


Fig. 6 Semi-log densities for yearly Apple log-returns for Gaussian, $\alpha = 2$ (blue), Stable, $\alpha = 1.33$ (green), Cauchy, $\alpha = 1$ (red), Gaussian scaling (orange) and semiparametric scaling (violet). (Color figure online)

4.4 Portfolio implementation

Exemplary, the representative investor has an investment horizon of one year. According to his client, no more than 20% ($b = 0.20$) of his wealth should be lost given probability $1 - \alpha = 99.5\%$ (Value at Risk). This implies that only $\alpha = 0.5\%$ of the wealth return paths should end below -20% . The investor is able to buy risk-free bonds with risk free rate $r_f = 1\%$ per year, representing the 15th asset. The maximization problems, without ($k = 15$) and with options ($k = 101$), as a special case of optimization in (11), are formulated within the framework of spectral measures. Subject to the VaR($1 - \alpha$), $\alpha = 0.5\%$ constraint, the Kelly criterion $G_{\phi_{\text{Elog}}}$ is maximized to achieve the portfolio with the highest growth rate:

$$\begin{aligned}
 f^* &= \arg \max_{f^* \in \mathbb{R}^k} G_{\phi_{\text{Elog}}} \{W_T(f)\} \\
 \text{s.t. } & S_{\phi_{Q_{0.5\%}}} \left\{ 1 - \frac{W_T(f)}{W_0} \right\} \leq 0.2, \\
 & \sum_{j=1}^k f_j \leq 1.
 \end{aligned}
 \tag{42}$$

Additionally the client aims to replace the VaR(99.5%) constraint with the expected shortfall restriction $ES(1 - \alpha)$ in order to account for events beyond the VaR level.

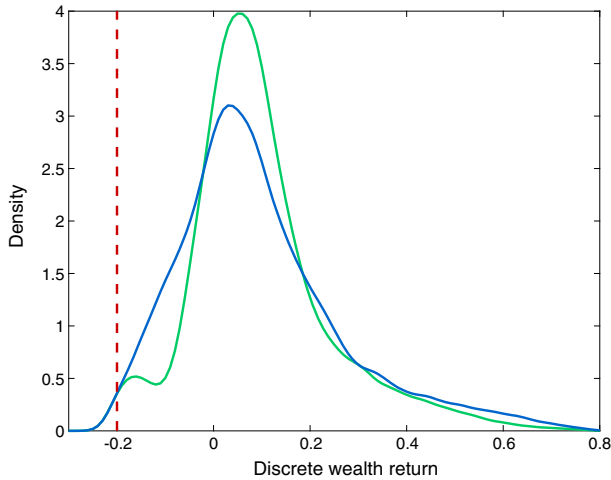


Fig. 7 Wealth return densities for VaR restricted Kelly optimization without (blue) and with (green) put options, VaR constraint (red). (Color figure online)

$$\begin{aligned}
 f^* &= \arg \max_{f^* \in \mathbb{R}^k} G_{\phi_{\text{Elog}}} \{W_T(f)\} \\
 \text{s.t. } & S_{\phi_{\text{CTE}_{0.5\%}}} \left\{ 1 - \frac{W_T(f)}{W_0} \right\} \leq 0.2, \\
 & \sum_{j=1}^k f_j \leq 1
 \end{aligned} \tag{43}$$

The resulting discrete wealth return distributions for the VaR restricted portfolios are given in Fig. 7. Including non-linear instruments into the restricted optimization proves to be beneficial for the Kelly criterion (Geometric mean), whilst preserving the VaR restriction (see Table 3). The protective put strategy allows to reduce probability mass for negative discrete wealth returns. The investment fractions of Table 4 show that the decrease in risk free bond for the portfolio with options is equivalent to the option investment.

Replacing the VaR constraint by the ES constraint, indicates that the stock investment is reduced for both cases with and without options, although not substantially. Enriching the ES restricted Kelly portfolios with put options has the same effect in terms of portfolio fractions as in the VaR restricted case (see Table 4) implying a higher geometric mean for the same ES constraint.

Extending the VaR constraint to the interval $0 \leq b \leq 1$, leads to a series of optimizations for all relevant quantile levels. $b = 0$ represents the risk free portfolio, whereas $b = 1$ implies that the investor can loose all of his fortune, given chosen confidence level.

Table 3 Discrete wealth return statistics (p.a. in %) for (restricted) Kelly portfolios with and without put options and Sharpe maximizing portfolio

Wealth Statistics	Kelly s.t. VaR $\leq 20\%$	Kelly s.t. ES $\leq 20\%$	Kelly s.t. ES $\leq 20\%$ with puts	Kelly	Kelly with puts	Sharpe
Geometric mean	8.04	7.9	8.19	18.25	20.92	$-\infty$
Arithmetic mean	9.64	9.51	9.45	34.05	32.45	62.08
Standard deviation	17.35	17.51	14.73	58.42	48.1	124.7
Skewness	1.09	1.12	1.21	0.65	0.89	1.11
Kurtosis	4.48	4.53	5.29	3.66	4.77	4.65
Minimum (in %)	-21.86	-20.66	-20.79	-95.86	-87.39	-100.00
CTE _{0.5%}	-20.9	-20.00	-20.00	-89.23	-82.41	-100.00
Q _{0.5%}	-20.00	-19.37	-19.29	-84.72	-78.64	-100.00
Q _{1%}	-19.11	-18.55	-18.39	-78.87	-73.43	-100.00
Q _{10%}	-9.65	-9.8	-4.54	-35.1	-9.79	-86.13
Q _{50%}	6.59	6.13	6.6	27.17	21.79	40.41
Q _{90%}	33.25	33.41	29.03	113.15	97.35	229.96
Q _{99%}	64.22	64.76	57.99	198.52	179.02	453.77
Maximum	94.29	91.78	84.2	330.45	294.65	849.39

Table 4 Portfolio fractions (in %) for (restricted) Kelly portfolios with and without put options and Sharpe maximizing portfolio

Investment Fractions	Kelly s.t. VaR $\leq 20\%$	Kelly s.t. VaR $\leq 20\%$ with puts	Kelly s.t. ES $\leq 20\%$	Kelly s.t. ES $\leq 20\%$ with puts	Kelly	Kelly with puts	Sharpe
Apple	4.31	4.65	5.24	4.55	19.57	18.76	72.03
Adobe	1.64	3.73	2.16	2.71	10.91	12.15	30.54
Amgen	0.03	1.66	0.65	0.63	9.35	7.05	21.76
Amazon	12.16	10.57	12.43	11.55	24.84	26.30	61.16
Comcast	0.27	0.17	0	0	1.12	0.43	-18
Costco	2.75	1.01	0.27	0.73	11.75	7.34	45.61
Cisco	0	0	0	0	0	0	-53.12
Gilead	0.56	0.39	0.5	0.67	7.97	6.54	14.63
Intel	0.05	0	0	0	0	0	-24.24
Microsoft	0.16	0	0	0	1.57	1.05	5.49
Nvidia	0.31	0.08	0.14	0.54	4.79	5.9	7
Pepsi	0.22	0	0	0	2.96	0	-19.98
Qualcomm	0	0	0	0	0	0	-51.19
Texas Instruments	0.11	0	0.01	0.34	1.05	1.84	8.32
Risk free	77.35	71.96	78.55	73.81	4.06	0	0
Put options	0	5.39	0	4.7	0	12.6	0

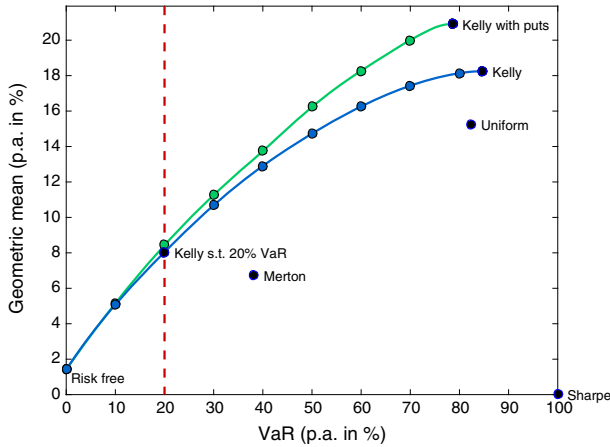


Fig. 8 Kelly-VaR (99.5%) frontier without (blue) and with (green) options, VaR constraint (red), benchmark portfolios for Uniform, Merton, Sharpe and Kelly (black). (Color figure online)

$$\begin{aligned}
 f^* &= \arg \max_{f^* \in \mathbb{R}^k} G_{\phi_{\text{Elog}}} \{W_T(f)\} \\
 \text{s.t. } & S_{\phi_{Q_{0.5\%}}} \left\{ 1 - \frac{W_T(f)}{W_0} \right\} \leq b, \quad 0 \leq b \leq 1, \\
 & \sum_{j=1}^k f_j \leq 1.
 \end{aligned} \tag{44}$$

This series of restricted optimizations constitutes the Kelly-VaR frontier (Fig. 8), in which each point represents a growth-optimal portfolio given quantile (VaR) constraint. The portfolio with (without) options, which can lose at most 20% with 99.5% probability is the portfolio where the green (blue) frontier crosses the quantile constraint (red). Except for the risk-free portfolio, $0 < b \leq 1$, every restricted portfolio with put options outperforms the portfolio without options in terms of the geometric mean. The unrestricted Kelly portfolio exhibits the highest geometric mean possible (18.25%), for a given VaR of 84.72%. Including put options into the unrestricted Kelly optimization increases the geometric mean (20.92%) and reduces the VaR to 78.64% at the same time. Relevant benchmark portfolios such as equally distributed (uniform),

$$f_{\text{Uniform}} = \mathbf{1} \frac{1}{k} \tag{45}$$

the closed-form Merton solution under log-utility and Gaussianity (Merton 1992),

$$f_{\text{Merton}} = \Sigma^{-1}(\mu - \mathbf{1}r_f) \tag{46}$$

and the Sharpe portfolio (Merton 1972),

$$f_{\text{Sharpe}} = \frac{\Sigma^{-1}(\mu - \mathbf{1}r_f)}{\mathbf{1}^\top \Sigma^{-1}(\mu - \mathbf{1}r_f)} \quad (47)$$

are not close to the Kelly-VaR frontier. Specifically the Sharpe investor, who allows for a larger risk constraint b than the unrestricted Kelly solution, should still invest into the growth-optimal portfolio, as the geometric mean of the unrestricted Kelly portfolio cannot be surpassed. Although the one period Sharpe maximizer obtains a larger arithmetic return, the Kelly (geometric mean) optimization rests on a multiperiod investment process. The multiperiod investor cannot sustain substantial draw-downs in one period as for the Sharpe portfolio. Given the α -stable process, see Table 4, the Sharpe investor goes bankrupt every one-hundred years.

5 Conclusion

Whereas the unrestricted Kelly portfolio ensures the asymptotic outperformance of the investor's wealth towards significantly different strategies, the presented model ensures growth-optimal investment subject to personal risk. The constrained optimization is formulated within the framework of spectral measures, inducing quantile (VaR) and Conditional tail expectation (expected shortfall) as special cases. In order to allow for an asymmetric wealth distribution, long put options are included into the optimization.

Financial market returns are with large probability non-Gaussian. Founding on the work of Mandelbrot (1963), it can be observed that the stability parameter α is significantly smaller than two, speaking in favor of the class of α -stable distributions. Given a chosen investment horizon, the distribution of financial market returns is modelled as the sum of hourly random variables. For α -stable laws with $\alpha < 2$, the variance of those random variables is infinite. Hence, the standard CLT does not apply and the generalized CLT of Gnedenko and Kolmogorov (1954) is applied. For the multidimensional estimation elliptical α -stable distributions, implying a linear dependence structure, are used. Instead of simulating from this class of distributions, a semiparametric scaling approximation is proposed. The resulting annual distribution, represented by convoluted hourly distributions, is simply an affine transformation of its normalized hourly analogue, given its scaling nature.

Heavy tails beyond the Gaussian, linear dependence between the marginals and nonsignificant skewness are empirically supported. Correspondingly, the joint distribution of financial market returns for a specified horizon is estimated by elliptical α -stable distributions utilizing a semiparametric scaling approximation. The portfolio model is implemented for a representative investor with quantile (VaR) constraint. The resulting growth-optimum strategy maximizes the geometric mean, given his risk constraint. Including put options into the optimization levers the portfolio by a suitable protective put strategy, leading to an increased geometric mean for the same risk. For the Kelly-quantile frontier, except for the risk-free portfolio, every restricted portfolio with options outperforms the portfolio without options in terms of the geometric mean.

References

- Acerbi, C. (2002). Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking and Finance*, 26, 1505–1518.
- Algeot, P., & Cover, T. (1988). Asymptotic optimality and asymptotic equipartition properties of log-optimum investment. *Annals of Probability*, 16, 876–898.
- Basak, S., & Shapiro, A. (2001). Value-at-risk-based risk management: Optimal policies and asset prices. *Review of Financial Studies*, 14(2), 371–405.
- Boothe, P., & Glassman, D. (1987). The statistical distribution of exchange rates. *Journal of International Economics*, 22, 297–319.
- Breiman, L. (1961). Optimal gambling system for favorable games. *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*, 1, 63–68.
- Busseti, E., Ryu, E. K., & Boyd, S. (2016). Risk-constrained Kelly gambling. *The Journal of Investing*, 25(3), 118–134.
- Chopra, V., & Ziemba, W. T. (1993). The effect of errors in the mean, variance, and covariance estimates on optimal portfolio choice. *Journal of Portfolio Management*, winter, 6–11.
- Cizek, P., Härdle, W., & Weron, R. (2011). *Statistical tools for finance and insurance*. New York: Springer.
- Clark, R., & Ziemba, W. T. (1987). Playing the turn-of-the-year effect with index futures. *Operations Research*, 35(6), 799–813.
- Dacorogna, M. M., Gencay, R., Muller, U., Olsen, R. B., & Pictet, O. V. (2001). *An introduction to high frequency finance*. New York: Academic Press.
- Fama, E. F. (1965). The behavior of stock-market prices. *The Journal of Business*, 38(1), 34–105.
- Gnedenko, B. V., & Kolmogorov, A. N. (1954). *Limit distributions for sums of independent random variables*. Cambridge, MA: Addison-Wesley Mathematics Series.
- Hansen, B. E. (2015). Shrinkage efficiency bounds. *Econometric Theory*, 31, 860–879.
- Hausch, D. B., & Ziemba, W. T. (1985). Transactions costs, extent of inefficiencies, entries and multiple wagers in a racetrack betting model. *Management Science*, 31(4), 381–394.
- Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5), 1163–1174.
- James, W., & Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1, 361–379.
- Kearns, P., & Pagan, A. (1997). Estimating the density tail index for financial time series. *The Review of Economics and Statistics*, 79(2), 171–175.
- Kelly, J. (1956). A new interpretation of information rate. *Bell System Technology Journal*, 35, 917–926.
- Kring, S., Rachev, S.T., Höchstätter, M., Fabozzi, F.J. (2009). Estimation of α -stable sub-gaussian distributions for asset returns. In: G. Bol, S.T. Rachev, R. Würth (Eds.), *Risk Assessment. Contributions to Economics*. Physica-Verlag HD.
- Lévy, P. (1925). *Calcul des probabilités*. Paris: Gauthier-Villars.
- Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics*, 47(1), 13–37.
- MacLean, L. C., Thorp, E. O., & Ziemba, W. T. (2011). *The Kelly capital growth investment criterion: Theory and practice*. Singapore: World Scientific Press.
- MacLean, L. C., Ziemba, W. T., & Blazenko, G. (1992). Growth versus security in dynamic investment analysis. *Management Science*, 38(11), 1562–1585.
- Mandelbrot, B. (1963). New methods in statistical economics. *Journal of Political Economy*, 71, 421–440.
- Markowitz, H. M. (1952). Portfolio selection. *Journal of Finance*, 7, 77–91.
- Markowitz, H. M. (1976). Investment for the long run: New evidence for an old rule. *The Journal of Finance*, 31(5), 1273–1286.
- McCulloch, J. H. (1997). Measuring tail thickness to estimate the stable index α : A critique. *Journal of Business and Economic Statistics*, 15(1), 74–81.
- McFarland, J. W., Pettit, R., & Sung, S. K. (1982). The distribution of foreign exchange price changes: Trading day effects and risk measurement. *Journal of Finance*, 37(3), 693–715.
- Merton, R. C. (1972). An analytic derivation of the efficient portfolio frontier. *The Journal of Financial and Quantitative Analysis*, 7(4), 1851–1872.
- Merton, R. C. (1992). *Continuous time finance*. Malden: Blackwell Publishers Inc.
- Nolan, J. (2001). *Levy processes: Maximum likelihood estimation and diagnostics for stable distributions*. Boston: Birkhauser.

- Nolan, J. (2013). Multivariate elliptically contoured stable distributions: Theory and estimation. *Computational Statistics*, 28(5), 2067–2089.
- Nolan, J. (2017). *Stable distributions: Models for heavy tailed data*. Boston: Birkhauser.
- Rachev, S., & Mittnik, S. (2000). *Stable Paretian models in finance*. New York: Wiley.
- Roll, R. (1973). Evidence on the growth optimum model. *The Journal of Finance*, 28(3), 551–566.
- Samorodnitsky, G., & Taqqu, M. S. (1994). *Stable non-gaussian random processes: Stochastic models with infinite variance*. New York: Chapman and Hall.
- Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, 19(3), 425–442.
- Stein, C. (1955). Inadmissibility of the usual estimator for the mean of a multivariate distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1, 197–206.
- Thorp, E. O. (1971). Portfolio choice and the Kelly criterion. In *Proceedings of the business and economics section of the American Statistical Association* (pp. 215–224).
- Tobin, J. (1958). Liquidity preference as behavior towards risk. *The Review of Economic Studies*, 25(67), 65–86.
- Westerfield, R. (1977). The distribution of common stock prices changes: An application of transaction time and subordinated stochastic models. *The Journal of Financial and Quantitative Analysis*, 12(5), 743–765.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Data driven value-at-risk forecasting using a SVR-GARCH-KDE hybrid

Marius Lux¹ · Wolfgang Karl Härdle^{2,3} · Stefan Lessmann¹ 

Received: 19 June 2018 / Accepted: 30 October 2019
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract

Appropriate risk management is crucial to ensure the competitiveness of financial institutions and the stability of the economy. One widely used financial risk measure is value-at-risk (VaR). VaR estimates based on linear and parametric models can lead to biased results or even underestimation of risk due to time varying volatility, skewness and leptokurtosis of financial return series. The paper proposes a nonlinear and nonparametric framework to forecast VaR that is motivated by overcoming the disadvantages of parametric models with a purely data driven approach. Mean and volatility are modeled via support vector regression (SVR) where the volatility model is motivated by the standard generalized autoregressive conditional heteroscedasticity (GARCH) formulation. Based on this, VaR is derived by applying kernel density estimation (KDE). This approach allows for flexible tail shapes of the profit and loss distribution, adapts for a wide class of tail events and is able to capture complex structures regarding mean and volatility. The SVR-GARCH-KDE hybrid is compared to standard, exponential and threshold GARCH models coupled with different error distributions. To examine the performance in different markets, 1-day-ahead and 10-days-ahead forecasts are produced for different financial indices. Model evaluation using a likelihood ratio based test framework for interval forecasts and a test for superior predictive ability indicates that the SVR-GARCH-KDE hybrid performs competitive to benchmark models and reduces potential losses especially for 10-days-ahead forecasts significantly. Especially models that are coupled with a normal distribution are systematically outperformed.

Keywords Value-at-risk · Support vector regression · Kernel density estimation · GARCH

✉ Stefan Lessmann
stefan.lessmann@hu-berlin.de

Extended author information available on the last page of the article

1 Introduction

Events like the 2008 financial crisis or the outcome of the 2016 referendum in the UK came unexpected for many people. Yet, as these examples illustrate, unlikely events occur at times and they might have far reaching consequences. Risk management is the practice to analyze the macro-environment of an organization, identify possible adverse developments, and design suitable countermeasures.

For financial institutions and systemically important institutions in particular, a key risk management responsibility is to sustain solvency under adverse economic conditions (e.g., Silva et al. 2017; Kraus and Czado 2017). One of the most popular measures of uncertainty in financial markets is VaR (e.g., Alexander (2008)). VaR is based on the quantiles of a portfolio's profit and loss (*P&L*) distribution and can be interpreted as an upper bound on the potential loss that will not be exceeded with a given level of confidence. Its use is appealing because it summarizes the downside risk of an institution in one easily interpretable figure (e.g., Chen et al. 2012). Regulatory frameworks for the banking and insurance industry such as Basel III or Solvency II also rely on VaR for determining capital requirements. Compared to expected shortfall, an alternative risk measure with some superior mathematical properties (e.g., Kim and Lee 2016), an advantage of VaR may be seen in the fact that its estimation is more robust due to putting less weight on tail events and large losses, which may deteriorate the quality of statistical estimation routines (Sarykalin et al. 2008).

Several approaches have been proposed to estimate VaR including parametric statistical models and data-driven machine learning algorithms such as neural networks (NN) and SVR. In a seminal study, Kuester et al. (2006) review several statistical methods and compare these in a forecasting benchmark. Using more than 30 years of historical returns data, they find standard GARCH models to forecast VaR with the highest accuracy on average.

GARCH models are also employed by Chen et al. (2012) to estimate VaR for four daily series of stock market indices. More specifically, Chen et al. (2012) rely on an asymmetric Laplace distribution and model volatility using a GJR-GARCH model to introduce leverage effects. They then develop a time-varying model to allow for dynamic higher moments. These extensions allow for wider application of the model beyond forecasting.

Unlike the parametric approach of Chen et al. (2012), Franke and Diagne (2006) estimate VaR for the German stock index through fitting the mean and volatility of the return series using NNs. More specifically, they model the mean and volatility as an autoregressive (AR) and autoregressive conditionally heteroscedastic (ARCH) process, respectively. To derive VaR and expected shortfall, Franke and Diagne (2006) use the predicted mean and variance with the normal distribution. This model outperforms a standard GARCH model in terms of VaR exceedances and proofs capable of quickly adjusting volatility in case of shocks with only short impact. Dunis et al. (2010) also propose a NN-based approach towards forecasting VaR and expected shortfall.

Khan (2011) develops a VaR-model that forecasts realized volatilities using a combination of a heterogeneous AR model and SVR. VaR is then computed based on the normal, t - and skewed t -distribution. Applying this model to 5- and 15-minutes return data, Khan (2011) is able to confirm the suitability of the SVR component. O.

Radović et al. (2015) provide further evidence that SVR is a useful method for VaR forecasting. Likewise, Xu et al. (2016) introduce a multi-period VaR model using SVR in a quantile regression framework and show this approach to outperform GARCH models.

The findings of Xu et al. (2016) seem to disagree with prior results of Kuester et al. (2006) where GARCH models predict VaR with highest average accuracy and more accurately than quantile regression approaches in particular. Implementations of the quantile regression using a data-driven SVR model might explain the results of Xu et al. (2016). More specifically, the linear and parametric structure of standard GARCH models might be a limiting factor in VaR forecasting. Moreover, the parameters of GARCH-type models are usually estimated via maximum likelihood estimation (Bollerslev 1986). This necessitates distributional assumptions, which might be problematic since the distribution of financial returns is skewed and exhibits fat tails (Bali et al. 2008; Harvey and Siddique 2000).

Noting the possible limitation of the parametric framework, Schaumburg (2012) combines extreme value theory with nonparametric VaR estimation to forecast return distributions of four financial stock indices. A parametric conditional autoregressive value at risk (CAViaR) model serves as benchmark. The benchmark and the proposed model both circumvent the estimation of the mean and variance of the $P\&L$ distribution through predicting a quantile directly. In this regard, the approach of Schaumburg (2012) can be characterized as a nonparametric CAViaR model.

VaR forecasts based on CAViaR frameworks have also been considered in the benchmarking study of Kuester et al. (2006). In fact, the authors also introduce a novel CAViaR model in the paper and test it alongside various other VaR models. However, GARCH models and models relying on the t -distribution in particular emerge as most suitable for VaR modeling.

In summary, parametric GARCH models are superior to parametric quantile regression approaches for modeling VaR. To achieve better results with quantile regression approaches, it is necessary to include nonparametric parts into the model. Additionally, data driven GARCH models where the mean and variance components are modeled nonlinearly and nonparametrically lead to better results than parametric GARCH models. However, one shortcoming of the so far proposed data driven GARCH models is the use of parametric residual distributions. Particularly with regard to skewness and kurtosis, this can lead to misspecified residual distributions, resulting in wrong VaR estimates. Therefore, a novel purely nonparametric VaR model that grounds on the GARCH framework is proposed here. We apply data driven approaches to all GARCH components, i.e. mean, variance and residual distribution. Using this approach, we can overcome the possible misspecification of the residual distribution as well as the misspecifications of mean and variance. More specifically, we estimate the mean and variance of the $P\&L$ distribution using SVR and employ KDE to model the density of the standardized residuals (e.g., Härdle et al. 2004). We then integrate these components to derive a VaR forecast. In other words, we propose to start from the most effective parametric modeling approach of Kuester et al. (2006) and develop models that estimate its components in a purely data-driven manner. In contrast to other so far proposed GARCH based approaches to forecast VaR (e.g., Youssef et al. 2015; Khosravi et al. 2013; Aloui and Mabrouk 2010; Huang et al. 2009; Fan et al. 2008;

Hang Chan et al. 2007) no assumptions about process dependence structures or the distribution of residuals are made by combining SVR and KDE in a GARCH like fashion. Training the SVR-GARCH-KDE hybrid is, therefore, mainly computationally driven.

The use of SVR is motivated by the existence of a large body of research showing the effectiveness of SVR in forecasting financial time series (e.g., Chang et al. 2016; Devi et al. 2015; Tay and Cao 2001). Additionally, Sheta et al. (2015) compare the forecasting performance of SVR, ANN and traditional linear regression for the S&P 500. They find that SVR solves the task most successfully. Kazem et al. (2013) also provide evidence that SVR based models outperform ANNs in the context of financial forecasting. Moreover, Chen et al. (2010) use a SVR approach for predicting stock market volatility and use a recurrent ANN as benchmark which is outperformed. The use of nonparametric density estimation is motivated by the fact that the existence of fat tails and skewness in the distribution of financial returns can be considered as an empirically proven fact (e.g., Bali et al. 2008; Harvey and Siddique 2000). Although KDE is not a new approach in VaR forecasting (e.g., Chen et al. 2016; Schaumburg 2012; Malec and Schienle 2014), the particular combination of data-driven VaR estimation using SVR and nonparametric density estimation, which we propose in this paper, has, to the best of our knowledge, not been considered in prior work.

We assess the performance of the proposed model in comparison to GARCH-type models with different error distributions also including skewed and fat-tailed distributions. Empirical experiments using data from three major financial indices, namely the Euro STOXX 50, Nikkei 225 and Standard & Poor's 500 (S&P 500), suggest that the SVR-GARCH-KDE hybrid typically outperforms models that are coupled with a normal distribution and performs competitive to other benchmark models.

The remainder of the paper is organized as follows. In Sect. 2 VaR is defined and the methods underlying the proposed VaR modeling framework are presented. Specifically, the standard GARCH approach, nonparametric density estimation via KDE and SVR are introduced. The proposed SVR-GARCH-KDE hybrid is then developed based on these building blocks. After outlining the theoretical background, the SVR-GARCH-KDE hybrid is compared to other models on different datasets in Sect. 3. Concluding remarks and suggestions for future research are provided in the last section.

2 Methodology

2.1 Defining value-at-risk

In general, VaR can be derived from the portfolio's $P&L$ distribution. However, since today's portfolio value is usually known, it suffices to model the return distribution. For a formal description of VaR, let the portfolio returns r_t in period t have the cumulative distribution function (CDF) F_t . Then, the VaR in d trading days for a confidence level $1 - \alpha$ is defined as

$$VaR_{t+d}^\alpha = -F_{t+d}^{-1}(\alpha) = -\inf\{x \in \mathbb{R} : F_{t+d}(x) \geq \alpha\} \quad \text{with } \alpha \in (0, 1). \quad (1)$$

In the rest of the paper, VaR refers to the negative α -quantile of the next period's portfolio return distribution.

2.2 Estimating VaR using location-scale models

The proposed VaR modeling framework is based on the location-scale approach. Models of this class estimate the entire distribution of asset returns and derive VaR as a quantile of that distribution (e.g., Kuester et al. 2006). Such approaches assume the return process is described as

$$r_t = \mu_t + u_t = \mu_t + \sigma_t z_t, \quad z_t \sim (0, 1) \text{ i.i.d.} \quad (2)$$

In (2), μ_t is the location and $\sigma_t > 0$ the scale parameter. Given r_t belongs to the location-scale family and F_z is the CDF of z , we can compute VaR as

$$VaR_t^\alpha = -\left\{ \mu_t + \sigma_t F_z^{-1}(\alpha) \right\}. \quad (3)$$

Autoregressive moving average (ARMA) processes and GARCH-type models are commonly used to estimate μ_t and σ_t in (2).

2.3 Modeling volatility using GARCH models

Bollerslev (1986) introduces GARCH models by generalizing the volatility modeling approach of Engle (1982). In deriving the GARCH regression model Bollerslev (1986) starts by assuming conditional normality of the return process r_t :

$$r_t | \mathcal{F}_{t-1} \sim N(\beta^\top x_t, \sigma_t^2). \quad (4)$$

where x_t is a vector of lagged endogenous as well as exogenous variables, β an unknown parameter vector and \mathcal{F}_{t-1} the information set available at $t - 1$. Rewriting (4) as linear model with conditionally heteroscedastic and normally distributed disturbances gives:

$$r_t = \beta^\top x_t + u_t, \quad u_t | \mathcal{F}_{t-1} \sim N(0, \sigma_t^2). \quad (5)$$

Then, the GARCH(p, q) representation of the variance σ_t^2 is

$$\sigma_t^2 = \omega + \sum_{i=1}^q \delta_i u_{t-i}^2 + \sum_{j=1}^p \theta_j \sigma_{t-j}^2. \quad (6)$$

Bollerslev (1986) notes that (6) has an ARMA representation. To see this let $v_t = u_t^2 - \sigma_t^2$ and substitute σ_t^2 in (6) with $u_t^2 - v_t$ to obtain

$$u_t^2 - v_t = \omega + \sum_{i=1}^q \delta_i u_{t-i}^2 + \sum_{j=1}^p \theta_j (u_{t-j}^2 - v_{t-j}). \quad (7)$$

Rearranging (7) yields an ARMA representation for u_t^2 :

$$u_t^2 = \omega + \sum_{i=1}^q \delta_i u_{t-i}^2 + \sum_{j=1}^p \theta_j (u_{t-j}^2 - v_{t-j}) + v_t \quad (8)$$

$$= \omega + \sum_{i=1}^{\max(p,q)} (\delta_i + \theta_i) u_{t-i}^2 - \sum_{j=1}^p \theta_j v_{t-j} + v_t. \quad (9)$$

Based on (9) nonlinear and nonparametric volatility models can be introduced. This can be seen by noting that the conditional expectation of u_t^2 is equal to σ^2 . Consequently, the variance process can be modeled solely based on the observed values without making assumptions about the distributional form of the residuals or the structure of the variance process. Hence, the volatility model in the SVR-GARCH-KDE hybrid is motivated by the ARMA representation of σ^2 .

2.4 Nonparametric density estimation

The volatility of stock returns varies over time and a similar behavior has been observed for the third and fourth moment of the return distribution. For example, Bali et al. (2008) show that VaR forecasts can be improved by using past estimates of skewness and kurtosis. Given the evidence for the leptokurtic nature of stock returns (Franke et al. 2015), parametric distributional models might lack the flexibility to capture such distributional characteristics, which motivates the use of nonparametric methods such as KDE (e.g., Härdle et al. 2004).

Let X be a random variable with an absolutely continuous distribution function F . Further, denote the corresponding density function as f and let $\{x_1, \dots, x_n\}$ be a sample of i.i.d. realizations of X . Then, the kernel density estimator $\hat{f}_h(x)$ of $f(x)$ is defined as

$$\hat{f}_h(x) = \frac{1}{hn} \sum_{i=1}^n K\left(\frac{x_i - x}{h}\right) \quad (10)$$

where h is a bandwidth parameter with $h > 0$ and K is a so-called kernel function. Usually, a kernel function is assumed to be a symmetric density function, i.e.

$$\int_{-\infty}^{\infty} K(u) du = 1 \quad \text{with } K(u) \geq 0 \quad (11)$$

and

$$\int_{-\infty}^{\infty} uK(u)du = 0. \tag{12}$$

Conveniently, (11) implies that $\hat{f}_h(x)$ is also a density. Note that $\hat{f}_h(x)$ inherits all properties of K regarding continuity and differentiability.

The KDE based quantile estimator to forecast VaR can be derived as follows. First, the estimator for $F(x)$ that is based on KDE needs to be derived. Denote $\hat{F}_h(x)$ as the KDE based estimate of $F(x)$. Then, $\hat{F}_h(x)$ can be derived as follows:

$$\hat{F}_h(x) = \int_{-\infty}^x \hat{f}_h(z)dz \tag{13}$$

$$= \int_{-\infty}^x \frac{1}{nh} \sum_{i=1}^n K\left(\frac{z-x_i}{h}\right) dz \tag{14}$$

$$= \frac{1}{nh} \sum_{i=1}^n \int_{-\infty}^x K\left(\frac{z-x_i}{h}\right) dz. \tag{15}$$

Since the given kernel function K is a density, let Γ denote the corresponding CDF. Moreover, using the substitution $u = (z - x_i)/h$ one obtains

$$\hat{F}_h(x) = \frac{1}{n} \sum_{i=1}^n \int_{-\infty}^{\frac{x-x_i}{h}} K(u) du \tag{16}$$

$$= \frac{1}{n} \sum_{i=1}^n \Gamma\left(\frac{x-x_i}{h}\right). \tag{17}$$

Thus, $\hat{F}_h(x)$ is the mean of the CDF corresponding to K evaluated at $(x - x_i)/h$ for $i = 1, \dots, n$. Then, for $\alpha \in (0, 1)$ the KDE based quantile function \hat{Q}_h is obtained as

$$\hat{Q}_h(\alpha) = \hat{F}_h^{-1}(\alpha). \tag{18}$$

2.5 Support vector regression

SVR can be understood as a learning method to solve nonlinear regression tasks (e.g., Smola and Schölkopf 2004). It shares some similarities with a three-layer feed-forward NN and is able to approximate arbitrarily complex functions (Chen et al. 2010). However, NNs are based on minimizing the so-called empirical risk and tend to find only locally optimal solutions. In contrast, SVR minimizes the so-called structural risk to achieve better generalization and solves a convex optimization problem leading to a globally optimal solution. To describe the SVR model, let $\{(y_i, x_i) | i = 1, \dots, n; n \in \mathbb{N}\}$ with $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$ denote the training data set. Suppose f is a linear function

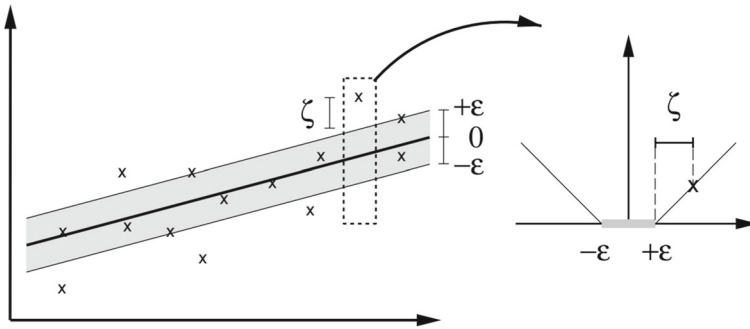


Fig. 1 The ϵ -insensitive loss function of the SVR algorithm. Slack variable ζ captures the loss above the ϵ -tube. Points within the grey shaded area have no impact on the loss. In contrast, all other observations contribute linearly to the loss.. *Source: Smola and Schölkopf (2004)*

such that

$$f(x) = \omega^\top x + b \tag{19}$$

where $\omega \in \mathbb{R}^p$ and $b \in \mathbb{R}$. Then, SVR aims to find an approximation of f that deviates at most by ϵ from the observed target y while being as flat as possible (i.e., in the sense that weights in ω are small). This translates into the following convex optimization problem:

$$\begin{aligned} &\text{minimize} && \frac{1}{2} \|\omega\|^2 \\ &\text{subject to} && \begin{cases} y_i - \omega^\top x_i - b \leq \epsilon \\ \omega^\top x_i + b - y_i \leq \epsilon. \end{cases} \end{aligned} \tag{20}$$

In view that (20) might lack a feasible solution, Vapnik (1995) introduces an ϵ -insensitive loss function:

$$L_\epsilon \{y - f(x)\} = \begin{cases} 0 & \text{if } |y - f(x)| \leq \epsilon \\ |y - f(x)| - \epsilon & \text{otherwise.} \end{cases} \tag{21}$$

To measure empirical loss (and thus model fit) using (21) Vapnik (1995) reformulates (20) using slack variables ζ and ζ^* that capture losses above and below the ϵ -tube around $f(x)$, respectively. Figure 1 depicts this approach. Only points outside the gray shaded ϵ -tube contribute linearly to the loss function.

Integrating the slack variables ζ and ζ^* into (20), the task to estimate a SVR model is equivalent to solving:

$$\begin{aligned}
 & \text{minimize} && \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^n (\zeta_i + \zeta_i^*) \\
 & \text{subject to} && \begin{cases} y_i - \omega^\top x_i - b \leq \epsilon + \zeta_i \\ \omega^\top x_i + b - y_i \leq \epsilon + \zeta_i^* \\ \zeta_i, \zeta_i^* \geq 0, \end{cases}
 \end{aligned} \tag{22}$$

where $C > 0$ is a regularization parameter to balance between model fit and complexity (e.g., Hastie et al. 2009). Larger (smaller) values of C put more (less) weight on maximizing model fit during SVR learning.

To capture nonlinear relationships between covariates and the response variable, SVR maps the input data into a higher dimensional feature space. The linear regression is then constructed in the transformed space, which corresponds to a nonlinear regression in the input space. The transformation is feasible from a computational point of view because SVR calculates the mapping by implicitly using a kernel function $k(x_i^\top x) = \phi^\top(x_i)\phi(x)$. To implement this approach, it is common practice to estimate a SVR model through solving the dual of (22), which is given as (e.g., Smola and Schölkopf 2004):

$$\begin{aligned}
 & \text{maximize} && -\frac{1}{2} \sum_{i,j=1}^n (\rho_i - \rho_i^*)(\rho_j - \rho_j^*)x_i^\top x_j - \epsilon \sum_{i=1}^n (\rho_i + \rho_i^*) + \sum_{i=1}^n y_i (\rho_i - \rho_i^*) \\
 & \text{subject to} && \begin{cases} \sum_{i=1}^n (\rho_i - \rho_i^*) = 0 \\ \rho_i, \rho_i^* \in [0, C]. \end{cases}
 \end{aligned} \tag{23}$$

The dual program (23) includes the input data only in the form of scalar products $x_i^\top x_j$. Replacing the scalar product by means of a kernel function is thus straightforward and does not affect the solver. In this work, we employ the Gaussian radial basis function (RBF) kernel (24) which is defined as

$$k(x_i^\top x) = \exp\left(-\frac{\|x - x_i\|^2}{2\gamma^2}\right) \tag{24}$$

where the meta-parameter $\gamma > 0$ governs the width of the Gaussian function and needs to be set by the modeler. The RBF kernel is used because it includes other kernels as special cases, possesses numerical advantages compared to alternatives, and often performs well in practical applications. Moreover, the RBF kernel can capture nonlinear relations. Other kernels that are usually presented as potential choices are e.g. the linear, polynomial or sigmoid kernel (e.g., Smola and Schölkopf 2004; Hastie et al. 2009). Keerthi and Lin (2003) show that the linear kernel is a special case of the RBF kernel. Moreover, the polynomial kernel has more parameters than the RBF kernel that make the tuning process more costly. Another advantage of the RBF over the polynomial kernel is that the polynomial kernel can converge to infinity which can cause numerical instability. In contrast, the domain of the RBF kernel is always

between 0 and 1. Moreover, Lin and Lin (2003) show that the sigmoid kernel behaves similar to the RBF kernel for certain parameters.

In order to construct the regression function (19), the weight vector ω is represented as a linear combination of observations in the training set:

$$\omega = \sum_{i=1}^n (\rho_i - \rho_i^*) x_i. \quad (25)$$

More specifically, for observations x_i where $f(x_i)$ is within the ϵ -tube holds that $\rho_i = \rho_i^* = 0$. Consequently, $f(x)$ depends only on the observations outside the ϵ -tube. These x_i are called support vectors. Accordingly, (25) is also called the support vector expansion of ω . Rewriting the regression function in terms of the support vector expansion gives the SVR forecasting model:

$$f(x) = \sum_{i=1}^n (\rho_i - \rho_i^*) x_i^\top x + b. \quad (26)$$

In the nonlinear case, the scalar product in (26) is once again replaced by a kernel function.

2.6 SVR-GARCH-KDE hybrid

In the following section, we introduce a nonlinear GARCH hybrid to forecast VaR based on a combination of SVR and KDE. Subsequently, we elaborate on the estimation of the corresponding forecasting model.

We assume the distribution of the return series r_t to belong to the location-scale class, such that:

$$r_t = \mu_t + u_t = \mu_t + \sigma_t z_t, \quad z_t \sim (0, 1) \text{ i.i.d.} \quad (27)$$

Consider an ARMA structure for the mean model where the only assumption about the error distribution is a zero mean and a finite variance. In addition, recall equation (9), which shows that GARCH processes can also be given an ARMA representation. This leads to the following mean and variance model:

$$r_t = c + \sum_{i=1}^s \alpha_i r_{t-i} + \sum_{j=1}^d \kappa_j u_{t-j} + u_t, \quad u_t \sim (0, \sigma_t^2) \quad (28)$$

$$u_t^2 = \omega + \sum_{i=1}^{\max(p,q)} (\delta_i + \beta_i) u_{t-i}^2 - \sum_{j=1}^p \beta_j v_{t-j} + v_t. \quad (29)$$

Let $e = \max(p, q)$, $\mathbf{r}_{t,s} = (r_{t-1}, r_{t-2}, \dots, r_{t-s})$, $\mathbf{u}_{t,k} = (u_{t-1}, u_{t-2}, \dots, u_{t-k})$ and $\mathbf{v}_{t,p} = (v_{t-1}, v_{t-2}, \dots, v_{t-p})$. Then, following Chen et al. (2010), we introduce the

nonlinear and nonparametric functions h and g such that the conditional mean and variance models of r_t are

$$r_t = h(\mathbf{r}_{t,s}, \mathbf{u}_{t,d}) + u_t \quad u_t \sim (0, \sigma_t^2) \tag{30}$$

$$u_t^2 = g(\mathbf{u}_{t,e}, \mathbf{v}_{t,p}) + v_t \quad v_t \sim WN(0, a_t^2) \tag{31}$$

where $WN(0, a_t^2)$ denotes white noise with expectation zero and variance a_t^2 . We propose to estimate $h(\cdot)$ and $g(\cdot)$ using SVR. The estimates for μ_t and σ_t in (27) are then obtained as:

$$\widehat{\mu}_t = h(\mathbf{r}_{t,s}, \mathbf{u}_{t,d}) \tag{32}$$

$$\widehat{\sigma}_t = \sqrt{g(\mathbf{u}_{t,e}, \mathbf{v}_{t,p})}. \tag{33}$$

By defining the estimated residuals as $\widehat{u}_t = r_t - \widehat{\mu}_t$, estimates of z_t are obtained as

$$\widehat{z}_t = \frac{\widehat{u}_t}{\widehat{\sigma}_t}. \tag{34}$$

Then, for $\widehat{Q}_{\widehat{z}}(\alpha)$ being the estimated quantile function of z , the VaR estimate for r_t is:

$$\widehat{VaR}_t^\alpha = - \left\{ h(\mathbf{r}_{t,s}, \mathbf{u}_{t,d}) + \sqrt{g(\mathbf{u}_{t,e}, \mathbf{v}_{t,p})} \widehat{Q}_{\widehat{z}}(\alpha) \right\}. \tag{35}$$

whereby we estimate $\widehat{Q}_{\widehat{z}}(\alpha)$ using KDE.

We now present a procedure to estimate VaR as in (35) and describe it in the context of producing 1-day-ahead VaR forecasts. A step-by-step overview is given in Algorithm 1. Let $\{r_t\}_{t=1}^T$ be the training set consisting of the daily returns from a portfolio where r_T is the most recent observation. In the first step, we model the mean process (30). To do this, we estimate an AR(s) model using SVR to obtain the estimated returns $\{\widehat{r}_t\}_{t=1+s}^T$. The set of estimated residuals $\{\widehat{u}_t\}_{t=1+s}^T$ is derived as $\widehat{u}_t = r_t - \widehat{r}_t$. Then, a moving average (MA) part can be introduced to the model such that r_t can be modeled as an ARMA(s, d) process by running SVR and including \widehat{u}_t . The sets of estimated returns and residuals from the ARMA(s, d) model are denoted as $\{\widehat{r}_t^*\}_{t=1+s+d}^T$ and $\{\widehat{u}_t^*\}_{t=1+s+d}^T$, respectively.

We also estimate the variance process in a two step approach and start by fitting the squared mean model residuals $\{\widehat{u}_t^{*2}\}_{t=1+s+d}^T$ in the way of an AR(e) process with SVR. Based on this, fitted variances $\{\widehat{\sigma}_t^{*2}\}_{t=1+s+d+e}^T$ are obtained. Then, an ARMA model for (31) is obtained in the same way as for the mean process by using the estimated model residuals $\{\widehat{v}_t\}_{t=1+s+d+e}^T$ where $\widehat{v}_t = \widehat{u}_t^{*2} - \widehat{\sigma}_t^{*2}$. Consequently, the final set of fitted variances is denoted as $\{\widehat{\sigma}_t^{*2}\}_{t=1+s+d+e+p}^T$. No assumptions are made about the starting values of the residuals. Hence, the final estimation of (31) is done using data for $T - s - d - e - p$ time points. Since SVR is applied without introducing further restrictions, it is not ensured that $\widehat{\sigma}_t^{*2}$ and $\widehat{\sigma}_t^{*2}$ are positive. Therefore, if the SVR estimate is $\widehat{\sigma}_t^{(*)2} \leq 0$, it will be replaced by the last positive estimated variance. In

case the first fitted variance is negative, it will be replaced by the first squared residual from the final mean model.

The set of estimated standardized residuals $\{\widehat{z}_t\}_{t=1+s+d+e+p}^T$ can be computed by applying (34). However, \widehat{z}_t does not necessarily have zero mean and unit variance. Hence, we perform the quantile estimation using scaled standardized residuals \widehat{z}_t^* :

$$\widehat{z}_t^* = \frac{\widehat{z}_t - \overline{\widehat{z}_t}}{\sqrt{\frac{1}{T-1} \sum_{i=1}^T (\widehat{z}_i - \overline{\widehat{z}_i})^2}} \quad (36)$$

where $\overline{\widehat{z}_t}$ denotes the empirical mean of \widehat{z}_t . The forecasted mean $\widehat{\mu}_{T+1}$ and standard deviation $\widehat{\sigma}_{T+1}$ are obtained from the mean and variance model. Finally, we use KDE to estimate the α -quantile of $\{\widehat{z}_t^*\}_{t=1+s+d+e+p}^T$. Then, the 1-day-ahead VaR forecast is:

$$\widehat{VaR}_{T+1}^\alpha = -[\widehat{\mu}_{T+1} + \widehat{\sigma}_{T+1} \widehat{Q}_{\widehat{z}^*}(\alpha)]. \quad (37)$$

An important aspect to note is that SVR and KDE depend on hyperparameters which cannot be derived analytically but must be found computationally. Therefore, the above described estimation procedure describes the estimation process only for one fixed set of hyperparameters. The hyperparameters for SVR given a RBF kernel is used are ϵ , γ and C . Additionally, if the SVR kernel is not set beforehand, it can be seen as hyperparameter itself. In the context of KDE the bandwidth h and the KDE specific kernel function K are hyperparameters. Since the goal is to train an efficient model on a purely data driven basis it is advisable to derive the hyperparameters computationally. However, for some hyperparameters exist theoretical results that support fixing them beforehand. As stated in Sect. 2.5 the RBF kernel is a reasonable choice in SVR. Moreover, for KDE the kernel choice has only a low practical relevance and there exist rule-of-thumb estimators for h when a Gaussian kernel is used which are computationally inexpensive (Härdle et al. 2004; Silverman 1986). Hence, the most relevant aspect in hyperparameter tuning of the SVR-GARCH-KDE hybrid are ϵ , γ and C for the SVR with a RBF kernel. It is important to note that the overall goal is to forecast VaR. Hence, the hyperparameters should be set with respect to the measure that is used to evaluate quantile forecasts. There exist different approaches as grid search, random search or more advanced optimization strategies that can be used to automate the process of hyperparameter optimization.

After defining the SVR-GARCH-KDE hybrid the question arises how model complexity, predictive power and the computational time for making VaR forecasts are related. Regarding the time to compute predictions the KDE part is fixed beforehand. This can be seen in Equation 10 and Algorithm 1. In order to compute an estimate the sum needs to be evaluated making the computational complexity $O(n)$ where n is number of observations. Regarding the SVR-GARCH-KDE hybrid the computational complexity of the KDE part is reduced depending on the order of the autoregressive and moving-average part of the mean and variance process to $O(n - s - d - e - p)$. With respect to SVR the complexity of making a prediction is $O(n_{SV}d)$, where d is the dimension of the feature space and n_{SV} the number of SVs. For the proposed

Algorithm 1 SVR-GRACH-KDE Estimation Algorithm for Forecasting VaR

- 1: AR(s) model for $\{r_t\}_{t=1}^T$ using SVR
- 2: Get errors from Step 1 $\{\widehat{u}\}_{t=1+s}^T$
- 3: ARMA(s, d) model for $\{r_t\}_{t=1+s}^T$ with results from Step 2 using SVR
- 4: Get errors from Step 3 $\{\widehat{u}^*\}_{t=1+s+d}^T$
- 5: AR(e) model for $\{\widehat{u}^{*2}\}_{t=1+s+d}^T$ using SVR
- 6: Get errors from Step 5 $\{\widehat{v}\}_{t=1+s+d+e}^T$
- 7: ARMA(e, p) model for $\{\widehat{u}^{*2}\}_{t=1+s+d+e}^T$ with results from Step 6 using SVR
- 8: Obtain volatility estimates $\{\widehat{\sigma}^*\}_{t=1+s+d+e+p}^T$ from Step 7
- 9: Get standardized residuals $\widehat{z}_t = \widehat{u}_t^*/\widehat{\sigma}_t$ for $t = 1 + s + d + e + p, \dots, T$
- 10: Scale $\{\widehat{z}_t\}_{t=1+s+d+e+p}^T$ to zero mean and unit variance and obtain $\{\widehat{z}_t^*\}_{t=1+s+d+e+p}^T$
- 11: Estimate the α -quantile $\widehat{Q}_{\widehat{z}^*}(\alpha)$ with KDE
- 12: Obtain $\widehat{\mu}_{T+1}$ and $\widehat{\sigma}_{T+1}$ by using the models from Step 3 and 7
- 13: VaR forecast: $\widehat{VaR}_{T+1}^\alpha = -[\widehat{\mu}_{T+1} + \widehat{\sigma}_{T+1} \widehat{Q}_{\widehat{z}^*}(\alpha)]$

application of forecasting the mean and variance of a financial time series the number of features is usually low. Therefore, the computational time for generating predictions is mainly driven by the sample size and negligible in this application.

Regarding the predictive performance, the regression function of the mean and variance process is determined by the SVs which implies that the higher the number of SVs, the higher the complexity. Using a SVR decision function with a high number of SVs can, therefore, lead to a mediocre out-of-sample performance. However, it is not possible to make an ex ante exact statement since the Vapnik-Chervonenkis dimension of the RBF kernel is infinite (Borges 1998). Moreover, the goal of the SVR-GARCH-KDE hybrid is to forecast quantiles. Evaluating the performance ex ante does, therefore, depend on the measure that is used to evaluate forecasted quantiles rather than the statistical properties of KDE and potential error bounds of SVR. In Sect. 3 the framework of Christoffersen (1998) will be applied. For achieving a good performance it is necessary to tune the parameters of the SVR and KDE part appropriately with respect to the target measure. Consequently, in this study the focus is put on the out-of-sample performance to optimize the quality of predictions and the hyperparameter tuning is done with a separate training set.

3 Empirical study

3.1 General setting

The SVR-GARCH-KDE hybrid is tested using stock indices to evaluate the performance for different regions. We consider three indices, namely the Euro STOXX 50, S&P 500 and Nikkei 225 which represent the Euro zone, the USA and Japan, respectively. The analysis is based on the log-returns of the adjusted index closing prices P_t :

Table 1 Descriptive statistics for the log-returns of the analyzed indices in the period from July 1, 2006 to June 30, 2016

Index	1st quartile	Mean	Median	3rd quartile	Variance	Skewness	Kurtosis
EuroStoxx50	-0.74	-0.01	0.01	0.78	2.39	-0.06	5.15
S&P500	-0.46	0.02	0.07	0.59	1.74	-0.33	9.94
Nikkei225	-0.76	0.00	0.05	0.88	2.68	-0.51	7.49

Note that the log-returns were multiplied by 100 before computing the descriptive statistics

$$r_t = \log(P_t) - \log(P_{t-1}). \quad (38)$$

The descriptive statistics of the analyzed indices are given in Table 1.

We forecast VaR for the quantiles $\alpha \in \{0.01, 0.025, 0.05\}$, considering forecast horizons of one and ten trading days. Estimating VaR for a horizon of ten trading days is especially important regarding the applicability of a VaR model. Besides forecasting VaR for a confidence level of 99%, which is equivalent to $\alpha = 0.01$ in our setting, a 10 days forecast horizon is required in the regulations of the Basel Committee on Banking Supervision.

In empirical applications, the quality of SVR depends on the kernel and parameter values which need to be set manually. The prevailing approach to determine parameter settings is grid search (e.g., Lessmann and Voß 2017), which we also apply in this study. For the density estimation via KDE, the Gaussian kernel function in combination with Silverman's rule of thumb are used to reduce computational cost. The Gaussian kernel function in KDE is equivalent to the standard normal distribution:

$$K(u) = \frac{1}{2\pi} e^{-\frac{u^2}{2}}. \quad (39)$$

Silverman (1986) showed that when (39) is used, robust density estimates can be obtained with the following estimator for the bandwidth h :

$$h_{rot} = 0.9 \min\{\hat{\sigma}_e, \hat{\sigma}_{iqr}\} n^{-1/5}. \quad (40)$$

In 40 $\hat{\sigma}_e$ denotes the empirical standard deviation and $\hat{\sigma}_{iqr}$ is an estimate of the standard deviation that is based on the interquartile range R :

$$\hat{\sigma}_{iqr} = \frac{R}{1.34}. \quad (41)$$

The rule-of-thumb estimator h_{rot} is computationally inexpensive whereas other proposed approaches to bandwidth estimation are more expensive since they are usually based on cross validation (Härdle et al. 2004). Moreover, the importance of the kernel choice regarding the performance of KDE is limited. For instance, Härdle et al. (2004) conclude that the kernel choice has almost no practical relevance after deriving the asymptotic mean integrated squared error for different kernel choices.

Table 2 Hypotheses for evaluating the appropriateness of VaR forecasts with the testing framework introduced by Christoffersen (1998)

Test	H_0	H_1
Unconditional coverage	$\mathbb{E}[V_t] = \alpha$	$\mathbb{E}[V_t] \neq \alpha$
Independence of violations	$\pi_{01} = \pi_{11}$	$\pi_{01} \neq \pi_{11}$
Conditional coverage	$\mathbb{E}[V_t \mathcal{F}_{t-1}] = \alpha$	$\mathbb{E}[V_t \mathcal{F}_{t-1}] \neq \alpha$

The evaluation of the models is based on Christoffersen (1998) who proposes a likelihood ratio (LR) test framework, which assesses the unconditional and conditional coverage as well as the independence of VaR exceedances. Moreover, Christoffersen (1998) shows that the test statistic for conditional coverage can be derived as the sum of the test statistics of the test for unconditional coverage and independence of VaR exceedances. Hence, it is possible to test whether the performance of a VaR model in terms of conditional coverage is determined by its ability to achieve correct unconditional coverage or adjust for changing volatility. This is useful in situations where the model has relatively bad conditional coverage but only one of the test statistics for unconditional coverage or independence is small. The hypotheses for the three tests are shown in Table 2 where α is the target quantile and \mathcal{F}_{t-1} the information set available at $t - 1$. V_t is a series of VaR violations with $V_t = \mathbb{I}(r_t < -VaR_t^\alpha)$ where $\mathbb{I}(x < c)$ denotes the indicator function:

$$\mathbb{I}(x < c) = \begin{cases} 1 & \text{if } x < c \\ 0 & \text{if } x \geq c. \end{cases} \tag{42}$$

For the test of independence of violations V_t is assumed to be a binary first-order Markov chain. The corresponding transition probability matrix of V_t is

$$\Pi_1 = \begin{bmatrix} 1 - \pi_{01} & \pi_{01} \\ 1 - \pi_{11} & \pi_{11} \end{bmatrix} \tag{43}$$

where

$$\pi_{ij} = \mathbb{P}(V_t = j | V_{t-1} = i). \tag{44}$$

As criterion for selecting a model from grid search and evaluate the performance of the SVR-GARCH-KDE hybrid with respect to benchmark models, the p value of the test for conditional coverage is used. Since the null hypothesis corresponds to correct conditional coverage which is the desired property, the one with the highest p value is considered to be the best.

By using the framework of Christoffersen (1998) to evaluate and select models, the main focus is put on the statistical properties and VaR violations of the considered models. In order to measure the performance from the perspective of a loss function that also takes into account the magnitude of violations, the quadratic and asymmetric

loss function that was introduced in one of the seminal papers for evaluating VaR models by Lopez (1998) will be used:

$$L(r_t, VaR_t^\alpha) = \begin{cases} 1 + (r_t - VaR_t^\alpha)^2 & \text{if } r_t < VaR_t^\alpha \\ 0 & \text{if } r_t \geq VaR_t^\alpha. \end{cases} \quad (45)$$

The model losses will be evaluated by employing the superior predictive ability test (SPA) framework proposed by Hansen (2005). Following the notation of Hansen (2005), for a finite set of decision rules $[\delta_{k,t-h}, k = 0, 1, \dots, m]$ made h periods in advance with respect to a random variable ξ_t , the relative performance corresponding to the benchmark $\delta_{0,t-h}$ is measured as

$$d_{k,t} = L(\xi_t, \delta_{0,t-h}) - L(\xi_t, \delta_{k,t-h}). \quad (46)$$

Based on the assumption that the alternatives are superior if and only if $\mathbb{E}[d_{k,t}] > 0$, Hansen (2005) formulates the hypothesis of interest for $\mathbf{d}_t = (d_{1,t}, \dots, d_{m,t})^\top$ as

$$H_0 : \mathbb{E}[\mathbf{d}_t] \leq 0. \quad (47)$$

A high p value indicates that none of the alternatives is superior to the benchmark. For further information regarding the SPA test we refer to Hansen (2005). As for the LR test of Christoffersen (1998), the SPA test will be performed for every model. This is done by using every considered model once as benchmark in terms of Hansen (2005) and comparing it to all other models. Note that due to performing multiple statistical tests the p values should be interpreted rather as an indication of model performance than in the context of a fixed significance level.

We perform all analyses using the statistical software R. The data has been downloaded from Yahoo Finance using the `quantmod` package. For SVR, we use the package `e1071`, which is the R implementation of the LIBSVM library of Chang and Lin (2011). To reduce computational time, we employ the `doParallel` package for parallelization of computations. The benchmark methods introduced below are implemented by using the `rugarch` package. All codes are available on www.quantlet.de. For details we refer to Borke and Härdle (2018) and Borke and Härdle (2017).

3.2 Benchmark methods

To test the SVR-GARCH-KDE hybrid, we compare its performance to the standard GARCH model and two of its variations. In particular, Franke et al. (2015) state that the most important variations are the EGARCH and TGARCH model. Hence, they serve as benchmarks in the empirical comparison. The EGARCH and TGARCH models are introduced by Nelson (1991) and Zakoian (1994), respectively. In contrast to standard GARCH models, both can account for asymmetric behavior with respect to past positive or negative returns. The two main differences between EGARCH and TGARCH models are that the former has a multiplicative and the latter an additive

model structure. Moreover, TGARCH models allow for different coefficients depending on the lags whereas EGARCH models capture the asymmetric behavior for all lags with one coefficient. The GARCH-type models that Kuester et al. (2006) analyze are coupled with different error distributions, i.e. the normal distribution, t -distribution and skewed t -distribution. We adopt this approach, which implies that we compare the SVR-GARCH-KDE hybrid to nine benchmarks.

3.3 Results

3.3.1 Model setting and tuning

We assume the mean process of r_t in (32) is zero. Moreover, we assume the variance process to have one AR and one MA part. These assumptions are imposed on both the SVR-GARCH-KDE hybrid and the benchmark models. We then forecast VaR for every index trading day from 2011-07-01 until 2016-06-30. The data is scaled to zero mean and unit variance in the SVR step; as suggested in the documentation of the e1071 package.

The tuning of the hyperparameters is done in a moving window approach using 251 return observations, which corresponds to approximately one trading year. The hyperparameters are tuned for every combination of index and quantile separately. For a given set of hyperparameters the model is trained based on the returns of the last trading year to predict the next day's VaR. Then, the window is shifted by one trading day. The tuning period reaches from 2006-07-01 until 2011-06-30, where 2006-07-01 marks the date of the first VaR prediction. The considered parameter values in the grid for SVR are

- $C \in \{10^{-4}, 10^{-3}, \dots, 10^4\}$
- $\psi \in \{0, 0.1, \dots, 0.9\}$ where $\epsilon = Q_{u_{scale}^2}(\psi)$
- $\gamma \in \{10^{-4}, 10^{-3}, \dots, 10^4\}$.

Note that the second point indicates that tuning is not done over fixed values of ϵ . Instead, in every step of the estimation, ϵ is determined based on the ψ -quantile of the squared scaled disturbances of the mean model. This corresponds to the squared scaled returns because we assume zero mean of r_t . The motivation behind this is the tendency of returns to form volatility clusters. Hence, a fixed ϵ can lead to good results in one volatility regime but might have a poor performance after a regime change. For instance, in the case of a financial crisis, the right tail of the distribution of past volatilities gets thicker. Hence, an ϵ that depends on the quantile of the distribution will increase such that large volatilities have automatically a higher influence on the estimated parameters from the SVR optimization. By using squared values it is ensured that only positive values are obtained for ϵ . However, notice that the distribution of the scaled squared disturbances, which are used in the SVR training, is shifted to the left of the distribution of the squared scaled disturbances. Hence, it is possible that for high values of ψ no observations are outside the ϵ -tube such that the model cannot be estimated.

The parameter settings that resulted in the best models for the SVR-GARCH-KDE hybrid during the tuning period are shown in Table 3. It can be seen that especially for

Table 3 The best models in the tuning period according to the p value of the test for conditional coverage

C	ψ	γ	Index	Quantile	Violations	UC	ID	CC
10	0.7	0.1	S&P500	1.0	0.95	86.62	63.08	87.84
10	0.7	0.01	S&P500	2.5	2.46	93.15	96.55	96.20
10	0.6	0.001	S&P500	5.0	5.00	99.48	99.57	99.57
100	0.8	0.01	Nikkei225	1.0	0.82	50.63	68.48	73.85
100	0.7	0.01	Nikkei225	2.5	2.70	66.43	57.80	52.61
100	0.7	0.10	Nikkei225	5.0	4.49	40.84	94.38	67.06
100	0.7	0.01	EuroStoxx50	1.0	1.02	93.28	60.41	87.11
0.1	0.6	0.001	EuroStoxx50	2.5	2.52	96.42	97.76	97.66
10000	0.6	10	EuroStoxx50	5.0	4.96	94.86	99.71	99.50

UC, ID and CC indicate the p value of the corresponding LR test. All values in the columns Quantile, Violations, UC, ID and CC are given in percent

Table 4 The models are presented with their mean rank. The mean rank was computed per index and quantile using the ties method *max* in the *frankv* function of the R package *data.table*

Model	Overall	CC 1 day	SPA 1 day	CC 10 days	SPA 10 days
TGARCH-SSTD	2.3	2.2	1.4	2.3	3.2
GARCH-SSTD	2.6	2.3	2.6	2.7	2.7
SVR-GARCH-KDE	3.2	4.9	2.2	4.8	1.0
GARCH-STD	5.9	6.1	6.4	4.9	6.3
TGARCH-NORM	6.0	6.0	7.7	5.1	5.1
GARCH-NORM	6.6	6.1	6.8	6.3	7.3
TGARCH-STD	6.7	5.3	9.0	5.1	7.4
EGARCH-SSTD	6.9	5.0	5.3	9.4	7.9
EGARCH-NORM	8.6	8.7	7.3	9.3	9.0
EGARCH-STD	8.6	8.8	7.2	9.6	8.8

The lower the mean rank, the better the model

ψ only a certain range appears among the best models. Moreover, the optimal ψ tends to be higher for lower quantiles. Based on the obtained parameters, VaR forecasts are produced from 2011-07-01 until 2016-06-30. A summary of model results in terms of average ranks is available in Table 4. Subsequent sections elaborate on these results in detail.

3.3.2 Model comparison

The results are presented for each quantile separately at the end of the section. In the Tables 5, 6 and 7 are the results for the 1-day-ahead forecast. The results for the 10-days-ahead forecasts can be found in the Tables 8, 9 and 10. To clearly identify the best performing models, every table is sorted in descending order for every index according to the p value of the conditional coverage test. The p value of the conditional coverage test is used for sorting to focus rather on statistical properties than pure

Table 5 Results for 1-day-ahead VaR forecasts from July 1, 2011 to June 30, 2016 for $\alpha = 0.01$

Model	Index	Violations	SPA	UC	ID	CC
EGARCH-SSTD	S&P500	1.11	4.99	69.27	57.46	79.01
TGARCH-SSTD	S&P500	0.72	92.22	28.52	71.87	52.94
GARCH-SSTD	S&P500	1.27	20.76	35.24	43.34	28.13
TGARCH-STD	S&P500	1.43	0.20	14.91	46.97	27.20
SVR-GARCH-KDE	S&P500	0.79	29.34	44.83	18.28	13.71
GARCH-STD	S&P500	1.67	16.17	2.95	65.77	6.15
EGARCH-STD	S&P500	1.75	5.79	1.58	37.62	3.67
TGARCH-NORM	S&P500	2.15	3.39	0.04	27.64	0.10
EGARCH-NORM	S&P500	2.23	2.40	0.02	25.88	0.04
GARCH-NORM	S&P500	2.38	0.20	0.00	94.85	0.01
TGARCH-SSTD	Nikkei225	1.21	4.99	47.61	54.48	64.58
GARCH-SSTD	Nikkei225	1.21	2.79	47.61	54.48	64.58
EGARCH-SSTD	Nikkei225	1.37	4.19	21.61	49.21	36.76
GARCH-STD	Nikkei225	1.69	1.80	2.60	66.33	5.56
TGARCH-STD	Nikkei225	1.77	1.00	1.37	37.30	3.23
SVR-GARCH-KDE	Nikkei225	0.24	59.48	0.13	90.41	0.55
EGARCH-STD	Nikkei225	2.09	0.80	0.07	29.17	0.19
GARCH-NORM	Nikkei225	2.17	0.40	0.03	88.16	0.14
TGARCH-NORM	Nikkei225	2.33	0.20	0.01	93.19	0.03
EGARCH-NORM	Nikkei225	2.42	0.20	0.00	95.21	0.01
SVR-GARCH-KDE	EuroStoxx50	1.53	76.25	8.17	44.23	16.36
GARCH-SSTD	EuroStoxx50	1.69	72.85	2.60	39.54	5.84
TGARCH-SSTD	EuroStoxx50	1.69	43.51	2.60	66.33	5.56
TGARCH-STD	EuroStoxx50	1.77	1.60	1.37	70.53	3.39
GARCH-STD	EuroStoxx50	1.85	1.80	0.70	35.15	1.70
TGARCH-NORM	EuroStoxx50	2.09	0.20	0.07	85.17	0.28
GARCH-NORM	EuroStoxx50	2.17	6.19	0.03	27.33	0.08
EGARCH-STD	EuroStoxx50	2.33	9.38	0.01	93.19	0.03
EGARCH-SSTD	EuroStoxx50	2.42	4.79	0.00	95.21	0.01
EGARCH-NORM	EuroStoxx50	2.50	3.19	0.00	96.87	0.00

UC, ID, CC and SPA indicate the p value of the corresponding test. The results are in descending order with respect to CC for each index. All values in the columns Quantile, Violations, UC, ID, CC and SPA are given in percent

losses. The abbreviations NORM, STD and SSTD indicate the normal, t - and skewed t -distribution, respectively. Additionally, the column headers UC, ID, CC and SPA refer to the corresponding p value of the test for unconditional coverage, independence of violations, conditional coverage and superior predictive ability.

One-Step-Ahead Forecast Model Evaluation for $\alpha = 0.01$ The SVR-GARCH-KDE hybrid is the best model for the Euro STOXX 50 for $\alpha = 0.01$. A visualization of its performance is given in Fig. 2. Here, the SVR-GARCH-KDE hybrid estimates in

Table 6 Results for 1-day-ahead VaR forecasts from July 1, 2011 to June 30, 2016 for $\alpha = 0.025$

Model	Index	Violations	SPA	UC	ID	CC
GARCH-SSTD	S&P500	2.62	34.33	78.12	98.97	95.22
EGARCH-SSTD	S&P500	3.02	1.40	25.18	98.94	51.31
TGARCH-SSTD	S&P500	2.38	80.44	79.19	22.60	46.40
GARCH-STD	S&P500	3.50	0.80	3.24	93.45	9.47
GARCH-NORM	S&P500	3.82	1.20	0.55	99.20	2.10
SVR-GARCH-KDE	S&P500	1.43	34.53	0.83	52.45	1.62
TGARCH-STD	S&P500	3.90	0.00	0.33	75.52	1.02
TGARCH-NORM	S&P500	4.05	1.60	0.12	69.37	0.36
EGARCH-STD	S&P500	4.37	0.00	0.01	96.11	0.06
EGARCH-NORM	S&P500	4.53	0.20	0.00	92.48	0.02
GARCH-SSTD	Nikkei225	2.90	43.91	38.01	99.90	67.96
TGARCH-SSTD	Nikkei225	2.42	71.66	84.78	22.29	46.72
TGARCH-STD	Nikkei225	3.14	0.40	16.44	97.67	37.15
TGARCH-NORM	Nikkei225	3.22	1.60	11.91	96.33	28.60
EGARCH-NORM	Nikkei225	3.22	0.40	11.91	96.33	28.60
GARCH-NORM	Nikkei225	3.38	0.00	5.87	92.85	15.55
GARCH-STD	Nikkei225	3.38	0.00	5.87	89.33	14.96
EGARCH-SSTD	Nikkei225	3.06	13.97	22.21	12.14	14.30
SVR-GARCH-KDE	Nikkei225	1.69	73.65	5.26	14.27	2.18
EGARCH-STD	Nikkei225	3.86	0.20	0.43	99.42	1.68
TGARCH-SSTD	EuroStoxx50	2.98	79.84	29.36	72.90	42.00
TGARCH-NORM	EuroStoxx50	3.22	0.00	11.91	83.46	24.78
TGARCH-STD	EuroStoxx50	3.38	0.00	5.87	89.33	14.96
GARCH-SSTD	EuroStoxx50	3.46	68.86	3.99	91.84	11.13
SVR-GARCH-KDE	EuroStoxx50	3.14	29.74	16.44	11.43	4.35
GARCH-STD	EuroStoxx50	3.70	5.19	1.11	97.42	3.88
GARCH-NORM	EuroStoxx50	3.95	5.59	0.26	75.92	0.81
EGARCH-STD	EuroStoxx50	4.27	1.20	0.03	88.73	0.12
EGARCH-SSTD	EuroStoxx50	4.51	1.40	0.00	95.50	0.02
EGARCH-NORM	EuroStoxx50	4.59	1.60	0.00	97.11	0.01

UC, ID, CC and SPA indicate the p value of the corresponding test. The results are in descending order with respect to CC for each index. All values in the columns Quantile, Violations, UC, ID, CC and SPA are given in percent

general higher values for VaR than the other models and exhibits more variability. For the S&P 500 and Nikkei 225 the SVR-GARCH-KDE hybrid outperforms all models that are coupled with a normal distribution. However, all models using a skewed t -distribution perform better. Especially for the Nikkei 225 this is caused by having low unconditional coverage due to risk overestimation. However, since risk overestimation causes less losses the SPA test p value is highest. In general, the models coupled with a normal distribution perform poorly for all indices. This comes as no surprise since the distribution of asset returns is usually leptokurtic.

Table 7 Results for 1-day-ahead VaR forecasts from July 1, 2011 to June 30, 2016 for $\alpha = 0.05$

Model	Index	Violations	SPA	UC	ID	CC
GARCH-SSTD	S&P500	5.01	57.49	98.97	89.07	89.06
GARCH-NORM	S&P500	5.72	0.00	24.94	99.79	51.40
TGARCH-SSTD	S&P500	5.41	87.82	51.47	60.38	48.83
GARCH-STD	S&P500	5.80	2.20	20.21	99.23	43.99
SVR-GARCH-KDE	S&P500	4.37	24.95	29.68	30.25	17.55
EGARCH-SSTD	S&P500	6.44	1.20	2.47	83.92	6.73
TGARCH-NORM	S&P500	6.60	0.40	1.30	77.88	3.56
EGARCH-NORM	S&P500	6.76	1.00	0.65	94.35	2.33
TGARCH-STD	S&P500	6.76	0.00	0.65	71.36	1.77
EGARCH-STD	S&P500	7.00	0.80	0.21	87.46	0.78
EGARCH-SSTD	Nikkei225	5.23	15.17	70.78	97.26	90.66
SVR-GARCH-KDE	Nikkei225	4.59	27.74	50.10	97.11	77.43
TGARCH-SSTD	Nikkei225	4.67	94.81	58.95	89.41	77.30
GARCH-SSTD	Nikkei225	5.07	34.93	90.69	75.29	74.78
TGARCH-NORM	Nikkei225	5.23	0.20	70.78	68.85	64.18
TGARCH-STD	Nikkei225	5.56	0.20	37.71	89.62	60.67
GARCH-NORM	Nikkei225	5.88	4.99	16.68	78.41	30.16
EGARCH-NORM	Nikkei225	5.72	0.60	25.68	49.18	25.85
GARCH-STD	Nikkei225	6.04	0.00	10.34	71.98	19.11
EGARCH-STD	Nikkei225	6.04	0.60	10.34	71.98	19.11
TGARCH-SSTD	EuroStoxx50	5.15	89.62	80.55	65.27	63.32
GARCH-SSTD	EuroStoxx50	5.96	57.29	13.21	95.83	30.84
GARCH-NORM	EuroStoxx50	5.96	48.90	13.21	95.83	30.84
TGARCH-NORM	EuroStoxx50	6.04	2.59	10.34	78.10	20.73
SVR-GARCH-KDE	EuroStoxx50	5.48	33.53	44.90	26.86	20.17
EGARCH-SSTD	EuroStoxx50	6.12	1.40	7.99	68.65	14.82
TGARCH-STD	EuroStoxx50	6.20	0.00	6.11	59.29	10.26
EGARCH-NORM	EuroStoxx50	6.52	1.80	1.85	99.10	6.18
GARCH-STD	EuroStoxx50	6.60	3.99	1.33	96.58	4.52
EGARCH-STD	EuroStoxx50	6.84	0.80	0.46	69.51	1.27

UC, ID, CC and SPA indicate the p value of the corresponding test. The results are in descending order with respect to CC for each index. All values in the columns Quantile, Violations, UC, ID, CC and SPA are given in percent

One-Step-Ahead Forecast Model Evaluation for $\alpha = 0.025$ The performance of the SVR-GARCH-KDE hybrid at the 2.5% level is not as good as for $\alpha = 0.01$ for the Euro STOXX 50. We observe the lowest p value in the test for independence of violations but the third best regarding the test for unconditional coverage. Interestingly, although the TGARCH and EGARCH model account for asymmetries in volatility, the former is the best and the latter the worst variance model. A relatively high risk overestimation for the S&P 500 and Nikkei 225 causes the SVR-GARCH-KDE hybrid to be on the

Table 8 Results for 10-days-ahead VaR forecasts from July 1, 2011 to June 30, 2016 for $\alpha = 0.01$

Model	Index	Violations	SPA	UC	ID	CC
TGARCH-SSTD	S&P500	1.03	11.78	90.58	30.09	29.88
SVR-GARCH-KDE	S&P500	0.95	94.01	86.85	25.94	25.59
GARCH-SSTD	S&P500	1.19	10.38	50.57	3.91	3.14
TGARCH-STD	S&P500	1.99	1.80	0.19	25.26	0.21
GARCH-STD	S&P500	1.99	2.99	0.19	25.26	0.21
TGARCH-NORM	S&P500	2.31	3.59	0.01	39.72	0.01
GARCH-NORM	S&P500	2.54	3.39	0.00	51.83	0.00
EGARCH-STD	S&P500	2.78	0.00	0.00	64.14	0.00
EGARCH-SSTD	S&P500	2.70	0.80	0.00	4.50	0.00
EGARCH-NORM	S&P500	3.26	0.60	0.00	14.87	0.00
TGARCH-SSTD	Nikkei225	1.45	3.19	13.59	52.97	17.42
GARCH-SSTD	Nikkei225	1.45	2.40	13.59	52.97	17.42
SVR-GARCH-KDE	Nikkei225	0.32	57.29	0.51	87.23	1.97
GARCH-STD	Nikkei225	1.85	0.80	0.70	19.57	0.51
TGARCH-STD	Nikkei225	1.93	1.80	0.34	22.56	0.31
GARCH-NORM	Nikkei225	2.01	1.80	0.16	25.77	0.18
TGARCH-NORM	Nikkei225	2.09	2.20	0.07	5.45	0.02
EGARCH-STD	Nikkei225	3.95	0.20	0.00	75.92	0.00
EGARCH-SSTD	Nikkei225	2.50	1.20	0.00	48.53	0.00
EGARCH-NORM	Nikkei225	2.58	0.80	0.00	15.65	0.00
SVR-GARCH-KDE	EuroStoxx50	0.97	64.07	90.41	62.85	88.31
GARCH-SSTD	EuroStoxx50	2.01	8.78	0.16	25.77	0.18
TGARCH-SSTD	EuroStoxx50	2.33	5.99	0.01	40.43	0.01
TGARCH-STD	EuroStoxx50	2.58	4.99	0.00	15.65	0.00
TGARCH-NORM	EuroStoxx50	2.66	4.79	0.00	18.07	0.00
GARCH-STD	EuroStoxx50	2.33	5.79	0.01	9.64	0.00
GARCH-NORM	EuroStoxx50	2.58	0.80	0.00	15.65	0.00
EGARCH-STD	EuroStoxx50	4.03	1.40	0.00	17.32	0.00
EGARCH-SSTD	EuroStoxx50	3.70	1.80	0.00	2.49	0.00
EGARCH-NORM	EuroStoxx50	3.30	0.20	0.00	44.47	0.00

UC, ID, CC and SPA indicate the p value of the corresponding test. The results are in descending order with respect to CC for each index. All values in the columns Quantile, Violations, UC, ID, CC and SPA are given in percent

sixth and ninth rank, respectively. As seen for $\alpha = 0.01$ this leads, however, to high p values in the SPA test due to lower losses. The performance for the S&P 500 can be seen in Fig. 3.

One-Step-Ahead Forecast Model Evaluation for $\alpha = 0.05$ The best performance of the SVR-GARCH-KDE hybrid for $\alpha = 0.05$ is rank two for the Nikkei 225. Here, it is only beaten by the EGARCH-SSTD model. A visualization is given in Fig. 4. For the other indices the SVR-GARCH-KDE hybrid ranks on place five. This is caused by

Table 9 Results for 10-days-ahead VaR forecasts from July 1, 2011 to June 30, 2016 for $\alpha = 0.025$

Model	Index	Violations	SPA	UC	ID	CC
TGARCH-SSTD	S&P500	2.62	13.57	78.12	55.96	53.84
GARCH-SSTD	S&P500	2.78	6.99	52.88	22.94	18.81
GARCH-STD	S&P500	3.18	3.39	13.82	39.68	13.22
TGARCH-STD	S&P500	3.42	0.20	4.80	51.32	7.27
GARCH-NORM	S&P500	3.58	2.79	2.14	59.33	4.20
SVR-GARCH-KDE	S&P500	1.75	95.21	7.15	16.45	3.24
TGARCH-NORM	S&P500	3.58	0.40	2.14	25.03	1.77
EGARCH-STD	S&P500	4.37	0.00	0.01	30.25	0.02
EGARCH-SSTD	S&P500	4.77	0.20	0.00	22.74	0.00
EGARCH-NORM	S&P500	4.61	0.20	0.00	17.75	0.00
GARCH-STD	Nikkei225	2.74	0.20	59.74	60.95	53.02
GARCH-SSTD	Nikkei225	2.25	8.58	57.31	36.53	31.17
GARCH-NORM	Nikkei225	3.30	0.00	8.45	86.53	19.54
TGARCH-SSTD	Nikkei225	2.74	0.00	59.74	20.72	18.02
TGARCH-STD	Nikkei225	3.06	0.00	22.21	33.30	15.80
TGARCH-NORM	Nikkei225	3.46	0.00	3.99	52.40	6.35
EGARCH-SSTD	Nikkei225	4.03	0.40	0.15	43.03	0.28
EGARCH-NORM	Nikkei225	4.11	0.00	0.09	46.87	0.19
SVR-GARCH-KDE	Nikkei225	1.29	48.10	0.26	5.17	0.06
EGARCH-STD	Nikkei225	5.31	0.20	0.00	96.40	0.00
TGARCH-SSTD	EuroStoxx50	3.22	50.30	11.91	40.62	12.06
GARCH-SSTD	EuroStoxx50	3.30	50.10	8.45	44.47	10.04
TGARCH-NORM	EuroStoxx50	3.38	9.38	5.87	48.40	8.11
GARCH-STD	EuroStoxx50	3.46	8.78	3.99	52.40	6.35
SVR-GARCH-KDE	EuroStoxx50	3.06	85.03	22.21	9.73	4.61
TGARCH-STD	EuroStoxx50	3.46	13.77	3.99	20.23	2.45
GARCH-NORM	EuroStoxx50	3.78	0.20	0.70	68.40	1.80
EGARCH-STD	EuroStoxx50	5.88	0.40	0.00	43.78	0.00
EGARCH-SSTD	EuroStoxx50	4.91	0.20	0.00	3.67	0.00
EGARCH-NORM	EuroStoxx50	4.83	0.00	0.00	23.77	0.00

UC, ID, CC and SPA indicate the p value of the corresponding test. The results are in descending order with respect to CC for each index. All values in the columns Quantile, Violations, UC, ID, CC and SPA are given in percent

having relatively low p values for the ID test. In terms of UC, the SVR-GARCH-KDE hybrid is the second and third best model for the S&P 500 and Nikkei 225, respectively. In comparison to the results for $\alpha = 0.01$, the models with a normal distribution show a better performance for $\alpha \in \{0.025, 0.05\}$. However, using the skewed t -distribution leads also for $\alpha \in \{0.025, 0.05\}$ to the best rankings.

Ten-Steps-Ahead Forecast Model Evaluation for $\alpha = 0.01$ As for the one-step-ahead forecasts, the SVR-GARCH-KDE hybrid is also the best model for the Euro STOXX

Table 10 Results for 10-days-ahead VaR forecasts from July 1, 2011 to June 30, 2016 for $\alpha = 0.05$

Model	Index	Violations	SPA	UC	ID	CC
TGARCH-NORM	S&P500	5.09	1.20	88.72	34.98	34.63
TGARCH-STD	S&P500	5.01	0.00	98.97	31.66	31.66
GARCH-SSTD	S&P500	4.29	2.20	23.85	57.38	28.64
GARCH-STD	S&P500	5.17	0.00	78.70	17.88	17.24
TGARCH-SSTD	S&P500	4.21	4.39	18.86	24.22	10.20
GARCH-NORM	S&P500	5.48	0.00	43.69	4.89	3.62
EGARCH-NORM	S&P500	6.52	0.00	1.80	29.69	1.81
EGARCH-STD	S&P500	6.60	0.00	1.30	32.87	1.50
SVR-GARCH-KDE	S&P500	3.34	56.89	0.41	17.11	0.28
EGARCH-SSTD	S&P500	7.23	0.00	0.06	62.62	0.18
TGARCH-SSTD	Nikkei225	5.23	10.18	70.78	94.63	88.21
GARCH-SSTD	Nikkei225	4.99	21.16	98.96	87.41	87.40
TGARCH-STD	Nikkei225	5.31	8.18	61.51	72.91	64.25
TGARCH-NORM	Nikkei225	5.31	12.18	61.51	72.91	64.25
GARCH-NORM	Nikkei225	5.72	5.99	25.68	99.95	52.55
GARCH-STD	Nikkei225	5.72	8.18	25.68	89.19	46.89
SVR-GARCH-KDE	Nikkei225	3.95	88.82	7.70	75.92	15.90
EGARCH-SSTD	Nikkei225	6.36	0.60	3.44	90.35	9.65
EGARCH-NORM	Nikkei225	6.12	3.59	7.99	31.60	6.82
EGARCH-STD	Nikkei225	7.97	0.00	0.00	91.62	0.00
SVR-GARCH-KDE	EuroStoxx50	5.07	50.10	90.69	61.34	60.92
TGARCH-NORM	EuroStoxx50	6.12	2.59	7.99	6.55	1.41
TGARCH-SSTD	EuroStoxx50	6.12	4.99	7.99	2.43	0.53
TGARCH-STD	EuroStoxx50	6.36	0.00	3.44	4.28	0.46
GARCH-NORM	EuroStoxx50	6.36	2.20	3.44	4.28	0.46
GARCH-SSTD	EuroStoxx50	6.52	2.99	1.85	6.06	0.38
GARCH-STD	EuroStoxx50	6.76	0.40	0.67	9.76	0.25
EGARCH-STD	EuroStoxx50	9.10	0.20	0.00	1.03	0.00
EGARCH-SSTD	EuroStoxx50	7.41	0.20	0.03	0.13	0.00
EGARCH-NORM	EuroStoxx50	7.17	0.60	0.10	0.18	0.00

UC, ID, CC and SPA indicate the p value of the corresponding test. The results are in descending order with respect to CC for each index. All values in the columns Quantile, Violations, UC, ID, CC and SPA are given in percent

50. For the S&P 500 and Nikkei 225 the SVR-GARCH-KDE hybrid is the second and third best model, respectively. For all indices, it is the only model not underestimating the risk. Especially the EGARCH models perform poorly. A visualization is given in Fig. 5. In contrast to the one-step-ahead forecasts, the models using a normal distribution are not shown but the EGARCH models due to their bad performance for 10-days-ahead forecasts. With respect to the SPA test, the SVR-GARCH-KDE

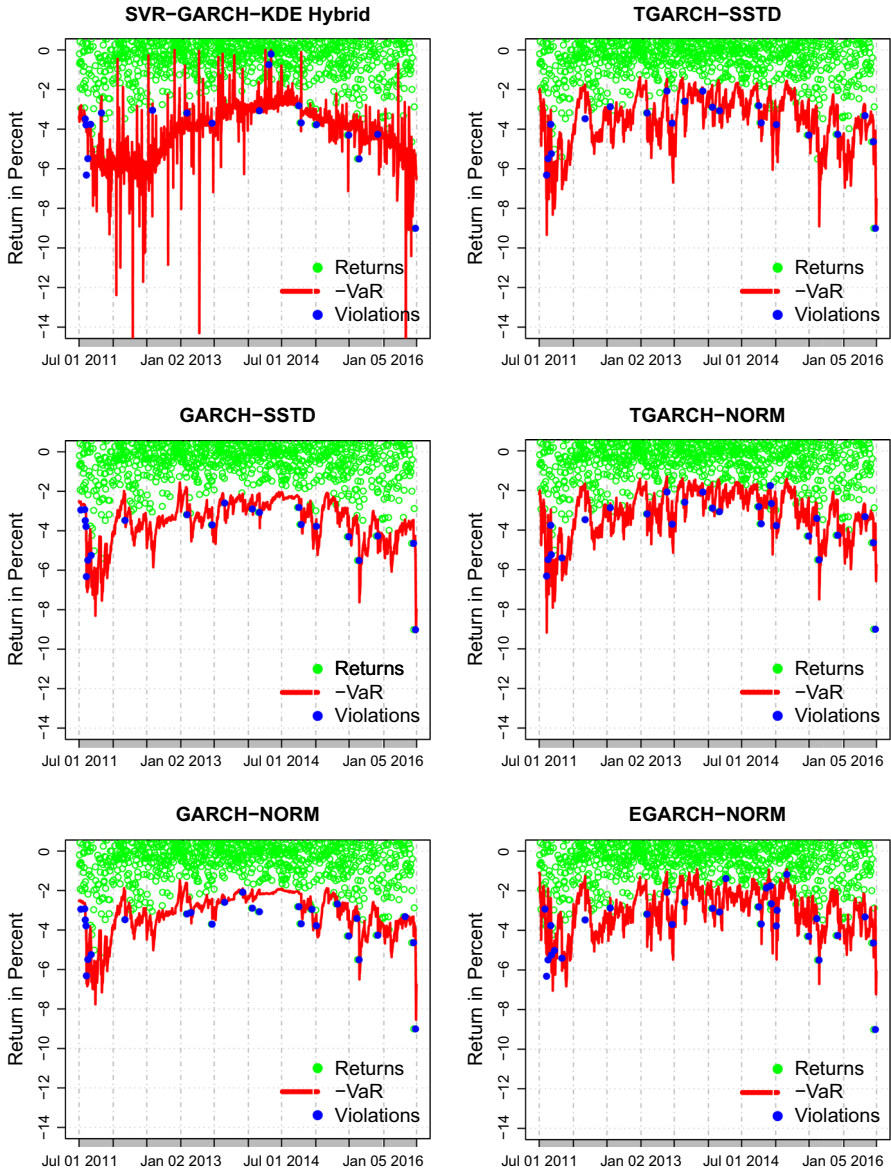


Fig. 2 VaR 1-day-ahead forecast model comparison for the Euro Stoxx 50 at $\alpha = 0.01$ in the period from July 1, 2011 to June 30, 2016. The SVR-GARCH-KDE hybrid is compared to models that are overall on average better and to the models using a normal distribution

hybrid causes less severe losses than any other model for all levels of the ten-steps-ahead forecast horizon.

Ten-Steps-Ahead Forecast Model Evaluation for $\alpha = 0.025$ Overall, the SVR-GARCH-KDE hybrid performs worse at the 2.5% than for the 1% level. In case of the

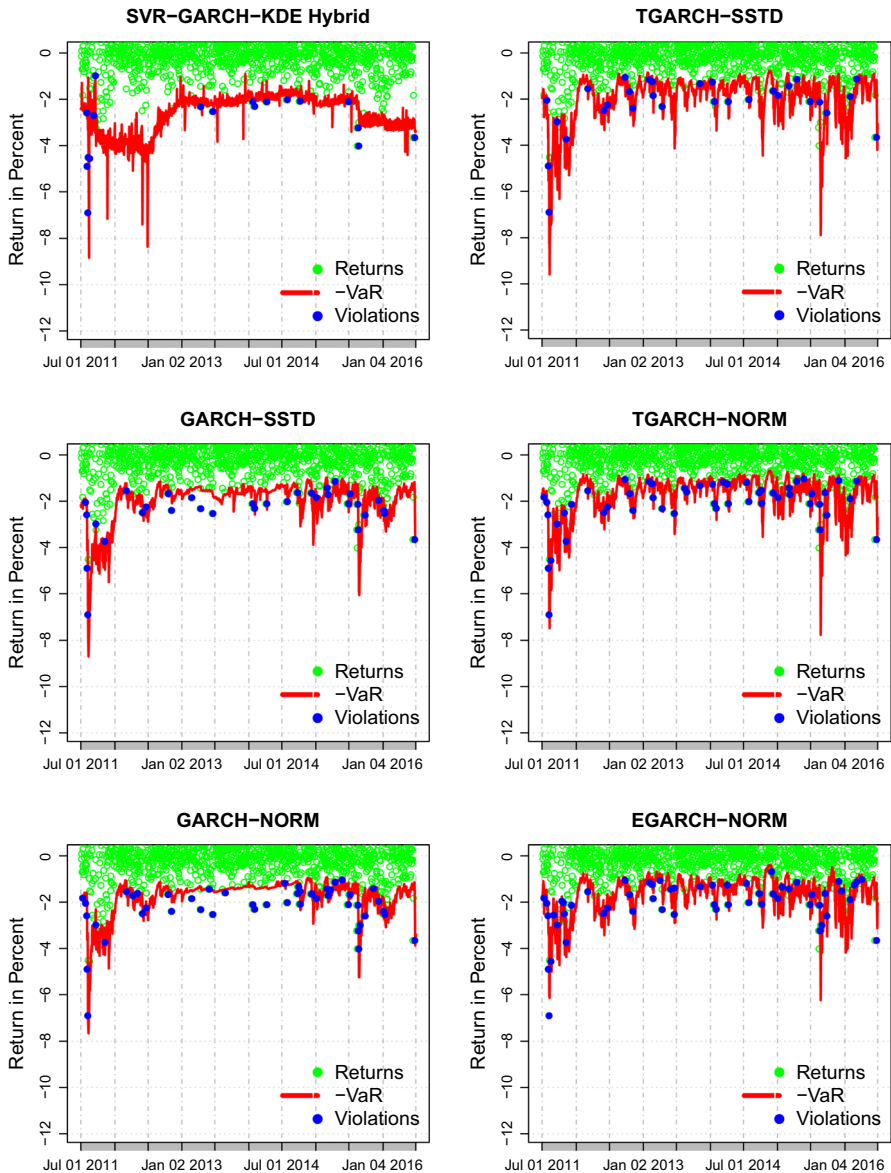


Fig. 3 VaR 1-day-ahead forecast model comparison for the S&P 500 at $\alpha = 0.025$ in the period from July 1, 2011 to June 30, 2016. The SVR-GARCH-KDE hybrid is compared to models that are overall on average better and to the models using a normal distribution

the S&P 500 and Nikkei 225 this is mainly driven by overestimating the risk leading to low p values in the UC test. In contrast, even though the SVR-GARCH-KDE hybrid has the highest p value in the UC test for the EURO STOXX 50, it is relatively low for the ID test. An exemplary visualization is given for the EURO STOXX 50 in Fig. 6.

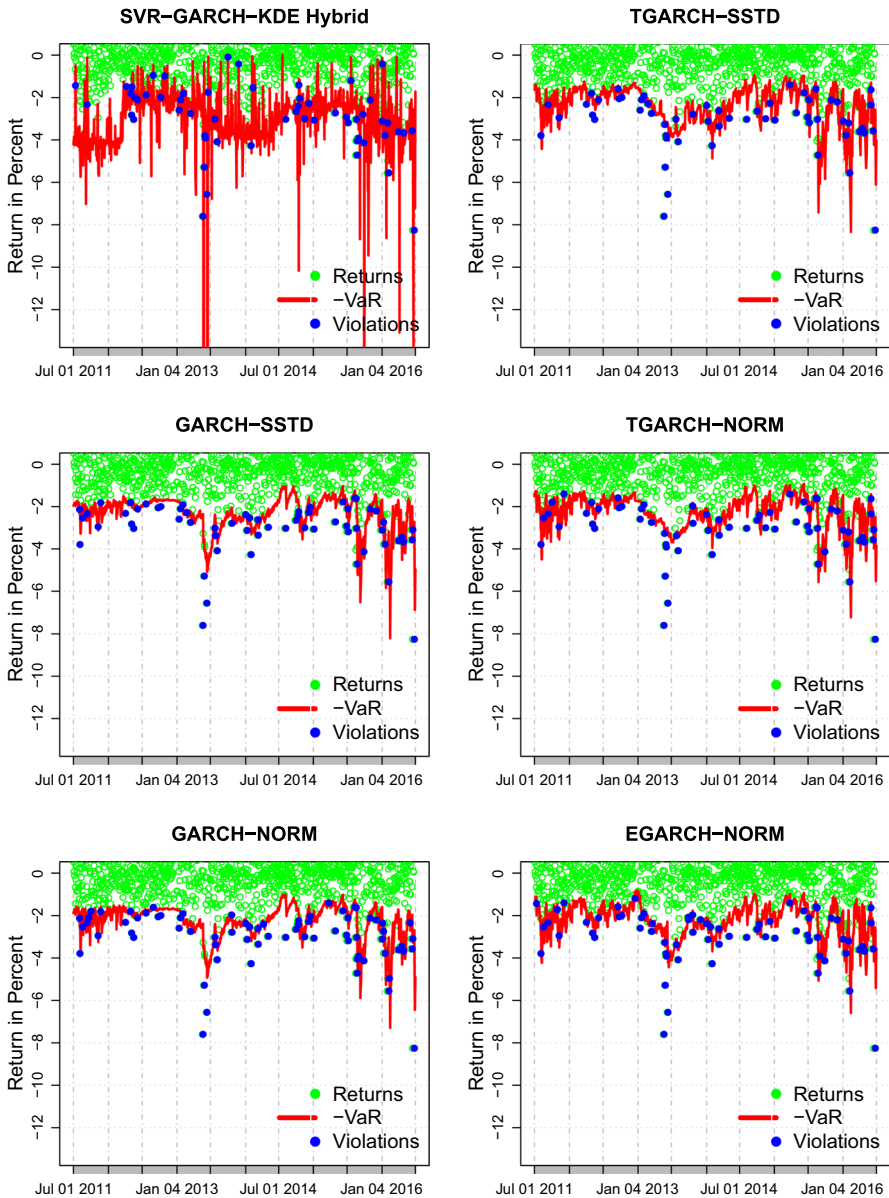


Fig. 4 VaR 1-day-ahead forecast model comparison for the Nikkei 225 at $\alpha = 0.05$ in the period from July 1, 2011 to June 30, 2016. The SVR-GARCH-KDE hybrid is compared to models that are overall on average better and to the models using a normal distribution

Ten-Steps-Ahead Forecast Model Evaluation for $\alpha = 0.05$ The performance of the SVR-GARCH-KDE hybrid for $\alpha = 0.05$ is mixed. While it is by far the best model for the EURO STOXX 50, it achieves only rank seven and nine for the Nikkei 225 and S&P 500, respectively. As for $\alpha = 0.025$ this is mainly driven by risk overestimation

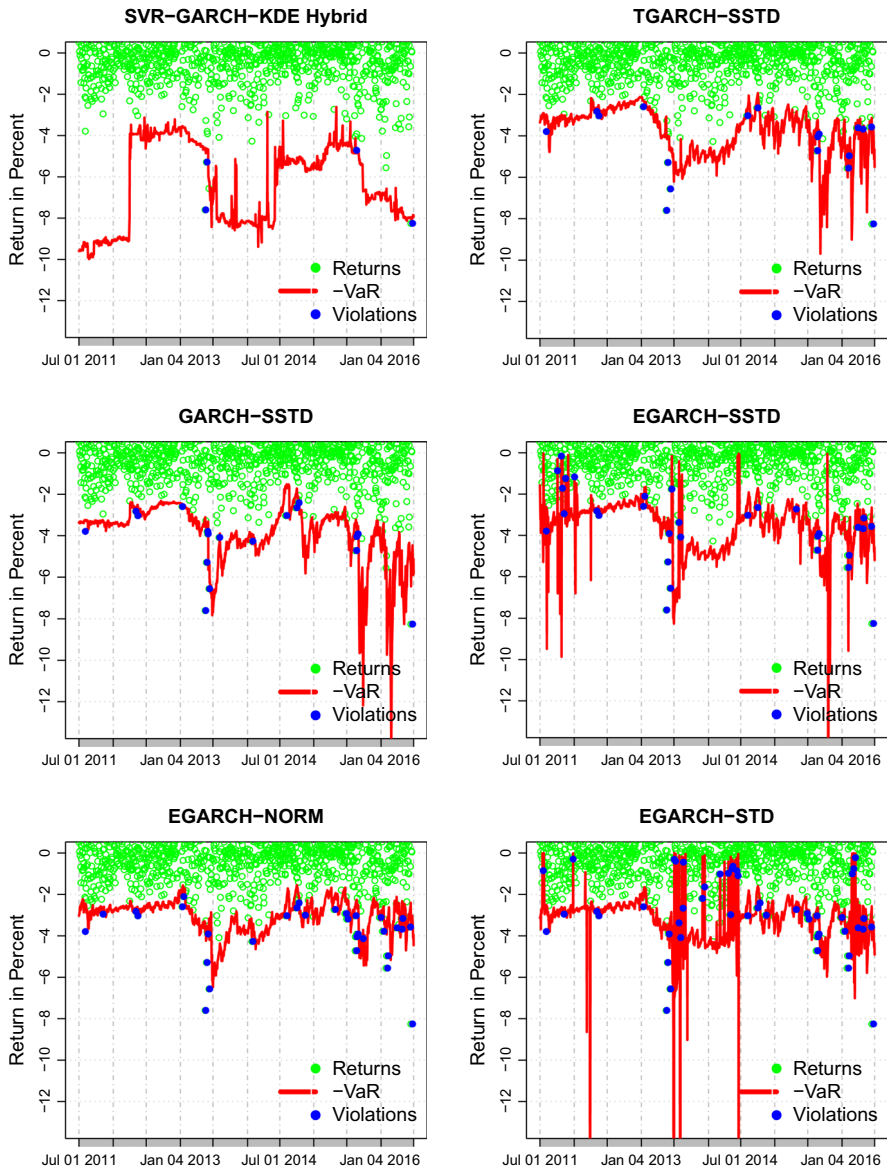


Fig. 5 VaR 10-days-ahead forecast model comparison for the Nikkei 225 at $\alpha = 0.01$ in the period from July 1, 2011 to June 30, 2016. The SVR-GARCH-KDE hybrid is compared to models that are overall on average better and to the EGARCH models

and low p values for the UC test. Similar to the 1-day-ahead forecasts, the models using a normal distribution perform better for $\alpha = 0.05$ than for the other quantiles. The TGARCH-NORM model is even the best for the S&P 500. A visualization is shown in Fig. 7.

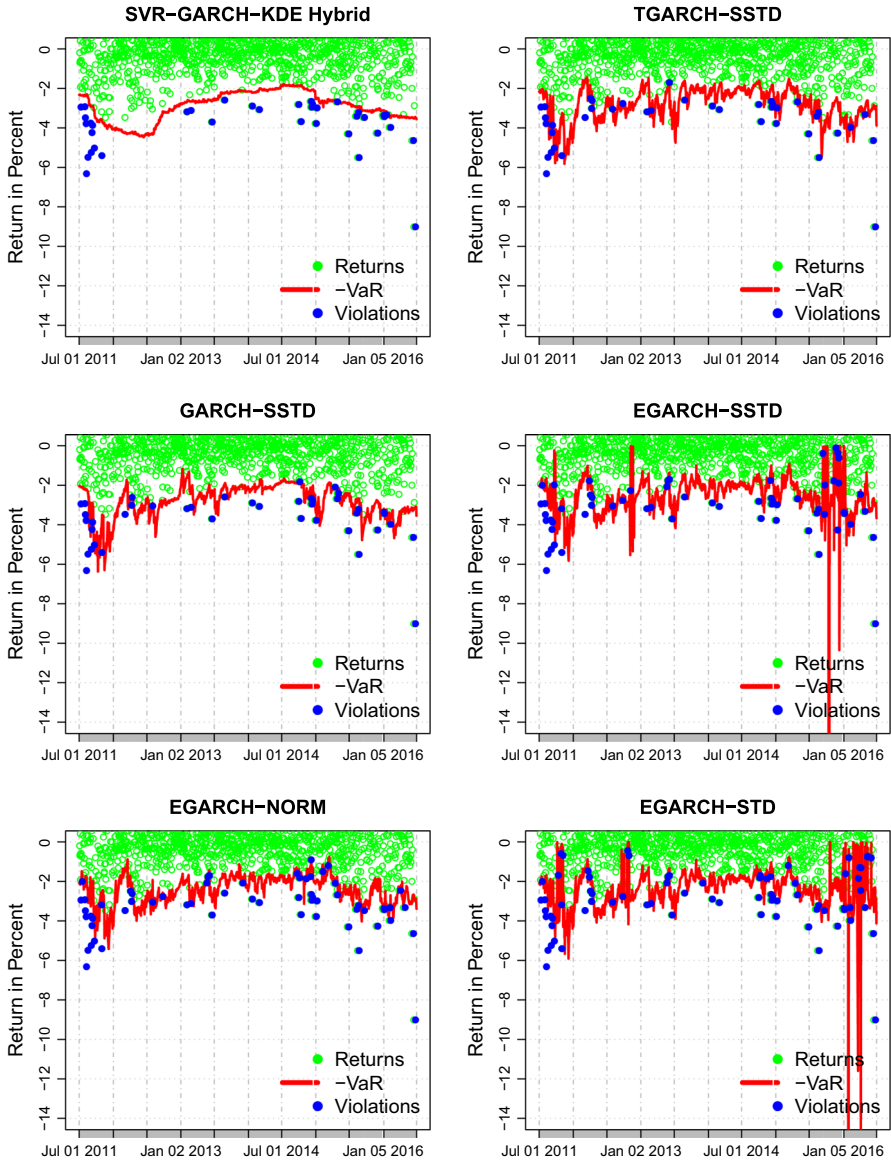


Fig. 6 VaR 10-days-ahead forecast model comparison for the Euro Stoxx 50 at $\alpha = 0.025$ in the period from July 1, 2011 to June 30, 2016. The SVR-GARCH-KDE hybrid is compared to models that are overall on average better and to the EGARCH models

Evaluation Summary Summarizing the results observed across all indices, quantiles and the two forecast horizons, we conclude that the SVR-GARCH-KDE hybrid displays a competitive performance. This can be seen in Table 4 where the mean ranks per index and quantile are presented for each model with respect to the CC and SPA test as well as the two forecast horizons. The SVR-GARCH-KDE hybrid is over-

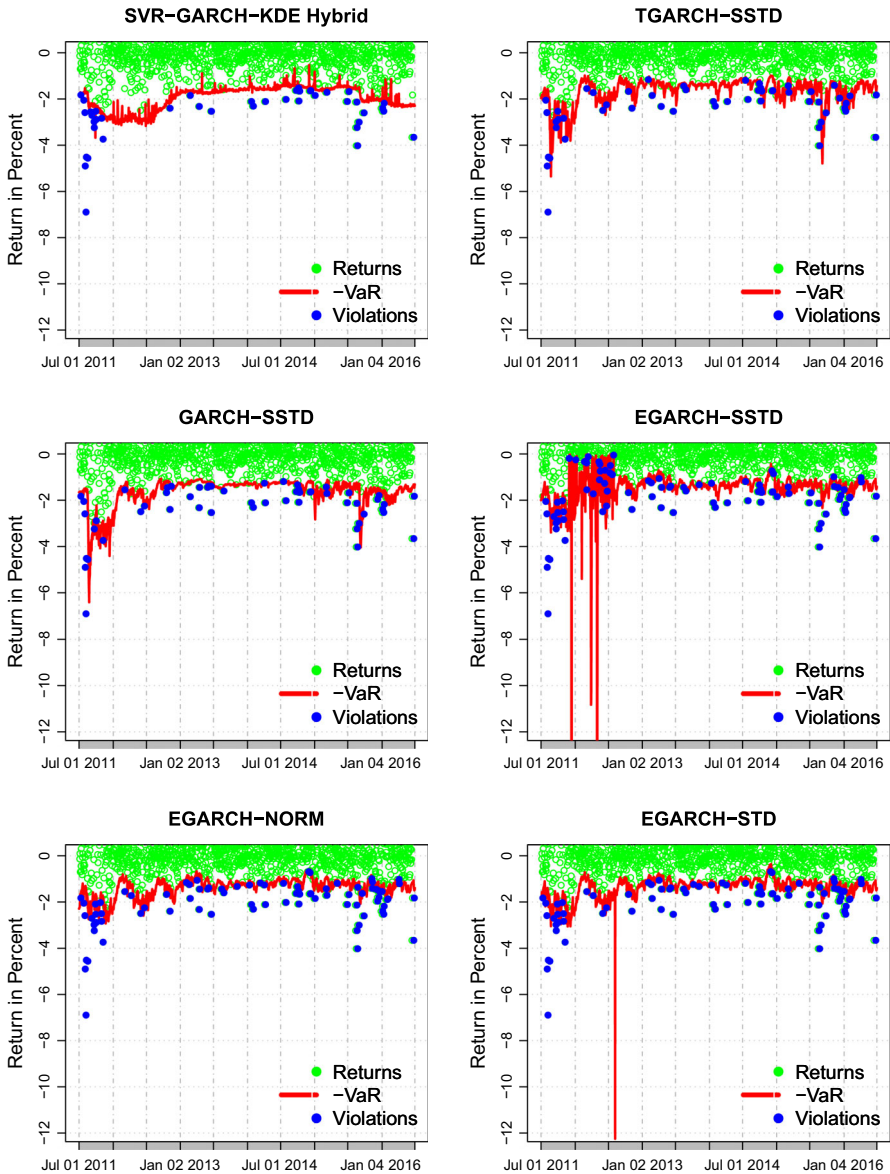


Fig. 7 VaR 10-days-ahead forecast model comparison for the S&P 500 at $\alpha = 0.05$ in the period from July 1, 2011 to June 30, 2016. The SVR-GARCH-KDE hybrid is compared to models that are overall on average better and to the EGARCH models

all and for the CC test p values the third best model for both forecast horizons. Regarding the SPA test it has rank two for the 1-day-ahead forecast horizon and is the best model for the 10-days-ahead forecast horizon. Benchmark models coupled with a normal distribution are usually outperformed by the SVR-GARCH-KDE hybrid. Consequently, the SVR-GARCH-KDE hybrid has a relatively high accuracy

especially compared to models with a normal distribution. The importance of this result stems from the fact that GARCH models with a normal distribution can be seen as a popular standard approach for modeling VaR or volatility and are, therefore, a natural benchmark. Additionally, excepting the case of the 10-days-ahead S&P 500 forecast for $\alpha = 0.05$ there is no setting where the models using a normal distribution perform best. This provides further evidence that using the normal distribution to measure market risk is inappropriate. However, models with a skewed t -distribution show in many cases the best performance. For instance, the TGARCH model coupled with the skewed t -distribution is with two exceptions always among the top three in terms of the CC test. This confirms the results of previous research showing usually skewed return distributions. Unlike the benchmarks, the SVR-GARCH-KDE hybrid tends to overestimate market risk. This might come from the choice of time interval for SVR parameter tuning. In particular, the tuning period covers the financial crisis of 2008 where market risk was extremely high. However, in the context of risk management, risk underestimation is more critical than risk overestimation because it can lead to bankruptcy in the short term. For instance, assume a hypothetical situation with the goal to forecast the 5% VaR, where the SVR-GARCH-KDE hybrid and a benchmark model have the same p value regarding the independence test. Additionally, assume the p value of the benchmark in the test of conditional coverage is higher because it has an unconditional coverage of 5.5% whereas that of the SVR-GARCH-KDE hybrid is 4%. The 1% overestimation of the SVR-GARCH-KDE hybrid works like a buffer for model risk since all estimation techniques exhibit statistical uncertainty. Hence, the use of the SVR-GARCH-KDE hybrid may be still more appealing than the use of benchmark models that tend to underestimate risk.

4 Conclusion

In a large-scale empirical comparison Kuester et al. (2006) find VaR models belonging to the location-scale class superior to alternative approaches. However, the location-scale models considered in their study are parametric and based on distributional assumptions. Motivated by the potential shortcomings of a parametric approach, the paper introduces a nonparametric and nonlinear VaR forecasting framework based on the location-scale class. The mean and volatility model are modeled with SVR in an ARMA and GARCH like fashion, respectively. In addition, the VaR forecast is obtained by estimating the distribution function of the standardized residuals via KDE.

To evaluate the performance of the SVR-GARCH-KDE hybrid, VaR is forecasted for three indices: Euro STOXX 50, Nikkei 225 and S&P 500, considering the different quantiles of $\alpha \in \{0.01, 0.025, 0.05\}$ for forecast horizons of 1 and 10 days. GARCH, EGARCH and TGARCH models coupled with the normal, t - and skewed t -distribution serve as benchmarks and are compared to the proposed model using a LR testing framework for interval forecasts (Christoffersen 1998) and the SPA test of Hansen (2005). Grid search with the goal to maximize the p value of the LR test for conditional coverage is used to set the SVR parameters. The SVR-GARCH-KDE hybrid delivers competitive results. For instance, in case of the 1-day-ahead forecast horizon it is the best model for the Euro STOXX 50 for $\alpha = 0.01$ and the third best model overall. The

TGARCH and GARCH model in combination with the skewed t -distribution show the best results. In contrast to the benchmark models, which usually underestimate risk, the SVR-GARCH-KDE hybrid tends to overestimate risk. This can lead to situations where the SVR-GARCH-KDE hybrid has an average performance regarding the p value of the test for conditional coverage. However, with respect to risk management, the use of the SVR-GARCH-KDE might be still favorable over using the benchmarks since all approaches exhibit statistical uncertainty. The tendency to overestimate risk can, therefore, serve as a model risk buffer. This is supported by the results of the SPA test that evaluates the loss function which penalizes especially large VaR violations. For instance, the proposed model has the best performance for the 10-days-ahead forecast horizon.

In general, the competitive results indicate that the proposed SVR-GARCH-KDE hybrid is a promising alternative. Despite fixing and not retuning the hyperparameters for five years, it is among the top three models with respect to conditional coverage and loss minimization measured by the SPA test for a forecast horizon of one as well as 10 days. Additionally, it is the best model in minimizing the loss function for 10-days-ahead forecasts. This indicates the proposed SVR-GARCH-KDE hybrid is a robust VaR modeling approach that is capable of capturing complex nonlinear structures in the volatility process and has the flexibility to model a wide class of tail events. Moreover, further improvements can be expected by refining the tuning routine. However, there exist several ways that can lead to an improved performance. First, tuning could be done for more parameters. For instance, in the KDE part of the estimation procedure, the kernel function and bandwidth estimator are set without tuning. Hence, considering different kernel functions and more flexible bandwidth estimators are potential ways to improve the performance further. Moreover, the kernel in the SVR part is also fixed and could be varied. Second, more recent information could be used in the parameter selection by re-tuning the model. Here, tuning is done for a block of five years of data. Then, based on the optimal parameters found for this data block, forecasts for five years are made and the parameters are held fixed. Thus, annual or even shorter re-tuning periods could result in parameters that are more appropriate for the existing market risk. Additionally, refining the grid can also result in better parameter choices.

In addition to modifying the tuning routine, the SVR-GARCH-KDE hybrid could be improved by changing the model specification. Overall, the TGARCH model with the skewed t -distribution achieves very good results. Hence, the SVR-GARCH-KDE hybrid could be modified such that it also accounts for asymmetric reactions of the volatility to past returns in a TGARCH like manner. Moreover, the proposed procedure does not ensure that the estimated variances are positive. Here, this problem is handled by replacing non-positive estimates with the last positive. However, positive variance estimates could be ensured by modeling the logarithm of the squared mean model residuals instead.

All above mentioned adjustments are potential starting points for future research to further improve the proposed framework. However, although the suggested modifications of the tuning procedure are reasonable approaches to improve the model performance, they also increase the computational complexity. After all, this is a slight drawback of the SVR-GARCH-KDE hybrid in comparison to standard models. It is, however offset by potential performance gains.

Acknowledgements Funding was provided by Deutsche Forschungsgemeinschaft (Grant No. IRTG 1792).

References


- Alexander C (2008) Value-at-risk models, market risk analysis, vol 4. Wiley, Chichester
- Aloui C, Mabrouk S (2010) Value-at-risk estimations of energy commodities via long-memory, asymmetry and fat-tailed GARCH models. *Energy Policy* 5:2326–2339. <https://doi.org/10.1016/j.enpol.2009.12.020>
- Bali TG, Mo H, Tang Y (2008) The role of autoregressive conditional skewness and kurtosis in the estimation of conditional VaR. *J Bank Finance* 32(2):269–282. <https://doi.org/10.1016/j.jbankfin.2007.03.009>
- Bollerslev T (1986) Generalized autoregressive conditional heteroskedasticity. *J Econ* 31(3):307–327. [https://doi.org/10.1016/0304-4076\(86\)90063-1](https://doi.org/10.1016/0304-4076(86)90063-1)
- Borke L, Härdle WK (2017) GitHub API based QuantNet mining infrastructure in R (March 6, 2017). SFB 649 discussion paper (2017-008). <https://doi.org/10.2139/ssrn.2927901>
- Borke L, Härdle WK (2018) Q3-D3-LSA. In: Hardle W, Lu HH-S, Shen X (eds) *Handbook of big data analytics*. Springer, New York. ISBN: 978-3-319-18284-1. <https://doi.org/10.1007/978-3-319-18284-1>
- Burges CJ (1998) A tutorial on support vector machines for pattern recognition. *Data Min Knowl Discov* 2(2):121–167. <https://doi.org/10.1023/A:1009715923555>
- Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol* 2(3):27:1–27:27
- Chang PC, Wu JL, Lin JJ (2016) A takagi sugeno fuzzy model combined with a support vector regression for stock trading forecasting. *Appl Soft Comput* 38:831–842. <https://doi.org/10.1016/j.asoc.2015.10.030>
- Chen S, Härdle WK, Jeong K (2010) Forecasting volatility with support vector machine-based GARCH model. *J Forecast* 29:406–433. <https://doi.org/10.1002/for.1134>
- Chen Q, Gerlach R, Lu Z (2012) Bayesian value-at-risk and expected shortfall forecasting via the asymmetric Laplace distribution. *Comput Stat Data Anal* 56(11):3498–3516. <https://doi.org/10.1016/j.csda.2010.06.018>, 1st issue of the *Annals of Computational and Financial Econometrics Sixth Special Issue on Computational Econometrics*
- Chen CYH, Chiang TC, Härdle WK (2016) Downside risk and stock returns. *J Bank Finance*. <https://doi.org/10.18452/4612>
- Christoffersen PF (1998) Evaluating interval forecasts. *Int Econ Rev* 39(4):841–862. <https://doi.org/10.2307/2527341>
- Devi KN, Bhaskaran VM, Kumar GP (2015) Cuckoo optimized SVM for stock market prediction. In: IEEE sponsored 2nd international conference on innovations in information, embedded and communication systems (ICJIECS). <https://doi.org/10.1109/ICIECS.2015.7192906>
- Dunis CL, Laws J, Sermpinis G (2010) Modelling commodity value at risk with higher order neural networks. *Appl Financ Econ* 20(7):585–600. <https://doi.org/10.1080/09603100903459873>
- Engle RF (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50(4):987–1007. <https://doi.org/10.2307/1912773>
- Fan Y, Zhang YJ, Tsai HT, Wei YM (2008) Estimating value at risk of crude oil price and its spillover effect using the GED-GARCH approach. *Energy Econ* 30(6):3156–3171. <https://doi.org/10.1016/j.eneco.2008.04.002>
- Franke J, Diagne M (2006) Estimating market risk with neural networks. *Stat Decis* 24(2):233–253. <https://doi.org/10.1524/std.2006.24.2.233>
- Franke J, Härdle WK, Hafner CM (2015) *Statistics of financial markets: an introduction*. Universitext, 4th edn. Springer, Berlin. <https://doi.org/10.1007/978-3-642-16521-4>
- Hang Chan N, Deng SJ, Peng L, Xia Z (2007) Interval estimation of value-at-risk based on GARCH models with heavy-tailed innovations. *J Econ* 137(2):556–576. <https://doi.org/10.1016/j.jeconom.2005.08.008>
- Hansen PR (2005) A test for superior predictive ability. *J Bus Econ Stat* 23(4):365–380. <https://doi.org/10.1198/073500105000000063>
- Härdle W, Müller M, Sperlich S, Werwatz A (2004) *Nonparametric and semiparametric models*. Springer series in statistics. Springer, Berlin

- Harvey CR, Siddique A (2000) Conditional skewness in asset pricing tests. *J Finance* 55(3):1263–1295. <https://doi.org/10.1111/0022-1082.00247>
- Hastie T, Tibshirani R, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction. Springer series in statistics, 2nd edn. Springer, New York
- Huang JJ, Lee KJ, Liang H, Lin WF (2009) Estimating value at risk of portfolio by conditional copula-GARCH method. *Insur Math Econ* 45(3):315–324. <https://doi.org/10.1016/j.insmatheco.2009.09.009>
- Kazem A, Sharifi E, Hussain FK, Saberi M, Hussain OK (2013) Support vector regression with chaos-based firefly algorithm for stock market price forecasting. *Appl Soft Comput* 13(2):947–958. <https://doi.org/10.1016/j.asoc.2012.09.024>
- Keerthi SS, Lin CJ (2003) Asymptotic behaviors of support vector machines with gaussian kernel. *Neural Comput* 15(7):1667–1689
- Khan AI (2011) Modelling daily value-at-risk using realized volatility, non-linear support vector machine and ARCH type models. *J Econ Int Finance* 3(5):305–321
- Khosravi A, Nahavandi S, Creighton D (2013) A neural network-garch-based method for construction of prediction intervals. *Electr Power Syst Res* 96:185–193. <https://doi.org/10.1016/j.epsr.2012.11.007>
- Kim M, Lee S (2016) Nonlinear expectile regression with application to value-at-risk and expected shortfall estimation. *Comput Stat Data Anal* 94(Supplement C):1–19. <https://doi.org/10.1016/j.csda.2015.07.011>
- Kraus D, Czado C (2017) D-vine copula based quantile regression. *Comput Stat Data Anal* 110(Supplement C):1–18. <https://doi.org/10.1016/j.csda.2016.12.009>
- Kuester K, Mittnik S, Paolella MS (2006) Value-at-risk prediction: a comparison of alternative strategies. *J Financ Econ* 4(1):53–89. <https://doi.org/10.1093/jjfinec/nbj002>
- Lessmann S, Voß S (2017) Car resale price forecasting: the impact of regression method, private information, and heterogeneity on forecast accuracy. *Int J Forecast* 33(4):864–877. <https://doi.org/10.1016/j.ijforecast.2017.04.003>
- Lin HT, Lin CJ (2003) A study on sigmoid kernels for svm and the training of non-PSD kernels by SMO-type methods. Technical report, Department of Computer Science, National Taiwan University. <https://www.csie.ntu.edu.tw/~cjlin/papers/tanh.pdf>
- Lopez JA (1998) Methods for evaluating value-at-risk estimates. *Econ Policy Rev*. <https://doi.org/10.2139/ssrn.1029673>
- Malec P, Schienle M (2014) Nonparametric kernel density estimation near the boundary. *Comput Stat Data Anal* 72(Supplement C):57–76. <https://doi.org/10.1016/j.csda.2013.10.023>
- Nelson DB (1991) Conditional heteroskedasticity in asset returns: a new approach. *Econometrica* 59(2):347–370
- Radović O, Stanković J, Stanković J (2015) Tail risk assessment using support vector machine. *J Eng Sci Technol Rev* 8(1):61–64
- Sarykalin S, Serraino G, Uryasev S (2008) Value-at-risk vs. conditional value-at-risk in risk management and optimization, pp 270–294. <https://doi.org/10.1287/educ.1080.0052>
- Schaumburg J (2012) Predicting extreme value at risk: nonparametric quantile regression with refinements from extreme value theory. *Comput Stat Data Anal* 56(12):4081–4096. <https://doi.org/10.1016/j.csda.2012.03.016>
- Sheta AF, Ahmed SEM, Faris H (2015) A comparison between regression, artificial neural networks and support vector machines for predicting stock market index. *Int J Adv Res Artif Intell* 4(7):4081–4096. <https://doi.org/10.14569/IJARAI.2015.040710>
- Silva W, Kimura H, Sobreiro VA (2017) An analysis of the literature on systemic financial risk: a survey. *J Financ Stab* 28(Supplement C):91–114. <https://doi.org/10.1016/j.jfs.2016.12.004>
- Silverman BW (1986) Density estimation for statistics and data analysis. Monographs on statistics and applied probability. Chapman & Hall, London. <https://doi.org/10.1007/978-1-4899-3324-9>
- Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. *Stat Comput* 14(3):199–222. <https://doi.org/10.1023/B:STCO.0000035301.49549.88>
- Tay FEH, Cao L (2001) Application of support vector machines in financial time series forecasting. *Omega* 29(4):309–317. [https://doi.org/10.1016/S0305-0483\(01\)00026-3](https://doi.org/10.1016/S0305-0483(01)00026-3)
- Vapnik VN (1995) The nature of statistical learning theory. Springer, New York. <https://doi.org/10.1007/978-1-4757-2440-0>
- Xu Q, Jiang C, He Y (2016) An exponentially weighted quantile regression via SVM with application to estimating multiperiod VaR. *Stat Methods Appl* 25(2):285–320. <https://doi.org/10.1007/s10260-015-0332-9>

Youssef M, Belkacem L, Mokni K (2015) Value-at-risk estimation of energy commodities: a long-memory GARCH-EVT approach. *Energy Econ* 51:99–110. <https://doi.org/10.1016/j.eneco.2015.06.010>
Zakoian JM (1994) Threshold heteroskedastic models. *J Econ Dyn Control* 18(5):931–955. [https://doi.org/10.1016/0165-1889\(94\)90039-6](https://doi.org/10.1016/0165-1889(94)90039-6)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Marius Lux¹ · Wolfgang Karl Härdle^{2,3} · Stefan Lessmann¹ 

Marius Lux
lux.marius@gmail.com

Wolfgang Karl Härdle
haerdle@hu-berlin.de

¹ School of Business and Economics, Humboldt-University Berlin, Unter den Linden 6, 10099 Berlin, Germany

² SKBI School of Business, Singapore Management University, 50 Stamford Road, Singapore 178899, Singapore

³ C.A.S.E.-Center of Applied Statistics and Economics, Humboldt-University Berlin, Unter den Linden 6, 10099 Berlin, Germany