IRTG 1792 Discussion Paper 2018-059

Towards the interpretation of time-varying regularization parameters in streaming penalized regression models

Lenka Zbonakova*
Ricardo Pio Monti*[2]
Wolfgang Karl Härdle*

* Humboldt-Universität zu Berlin, Germany
*[2] University College London, United Kingdom

This research was supported by the Deutsche Forschungsgemeinschaft through the International Research Training Group 1792 "High Dimensional Nonstationary Time Series".

http://irtg1792.hu-berlin.de
ISSN 2568-5619

International Research Training Group 1792

Towards the interpretation of time-varying regularization parameters in streaming penalized regression models*

Lenka Zboňáková[a], Ricardo Pio Monti[b] and Wolfgang Karl Härdle [a,c,d]

October 23, 2018

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.

JEL classification: C13, C15, C63

Keywords: Lasso, penalty parameter, stock prices, neuroimaging

*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.

[a]C.A.S.E. - Center for Applied Statistics & Economics, Humboldt-Universität zu Berlin, Spandauer Str. 1, 10178 Berlin, Germany

[b]Gatsby Computational Neuroscience Unit, UCL, 25 Howland Street, London WIT 4JG

[c]Sim Kee Boon Institute for Financial Economics, Singapore Management University, 50 Stamford Road, 178899 Singapore, Singapore

[d]W.I.S.E. - Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen, Fujian, China

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 (Hastie et al., 2015). 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 (Haykin, 2008).

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 Bottou (2010) and Duchi et al. (2011) 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. Examples of large-scale, non-stationary datasets, where the choice of the regularization parameter has been reported to be time-varying, include finance (Yu et al., 2017) and neuroscience (Monti et al., 2017a).

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. (2015) or Chernozhukov et al. (2018). 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. (2018) 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. Zboňáková et al. (2017) study the dynamics of the regularization parameter, focusing particularly on quantile regression in the context of financial data. Using sliding windows method, they demonstrate that the choice of the time-varying regularization parameter based on the adjusted Bayesian information criterion (BIC) is closely correlated with the financial volatility. The BIC was employed, as such a choice of parameter is optimal in terms of model consistency.

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, 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 the time-varying regularization parameters as well as the respective model parameters.

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 R software environment (R Core Team, 2014) and we provide code to reproduce all experiments at Quantlet platform.

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. 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 Tibshirani (1996), 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:

$$L_t(\beta, \lambda) = \sum_{t=1}^{n} w_t \left(Y_t - X_t^\top \beta\right)^2 + \lambda ||\beta||_1, \tag{1}$$

where $w_t > 0$ are weights indicating the importance given to past observations (Aggarwal, 2007) and $\| \cdot \|_1$ denotes the L_1-norm of a vector. For example, it is natural to allow w_t to decay monotonically in a manner which is proportional to the chronological proximity of the ith observation.

In the context of non-stationary data the optimal estimates of regression coefficients, $\widehat{\beta}_t$, may vary over time and several methods have been proposed in order to address this issue (Bottou, 2010; Duchi et al., 2011). However, the same argument can 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. While there exists a large range of methodologies through which to iteratively update the regression coefficients, the choice of the regularization parameter has, until recently, been largely overlooked. Lately, Monti et al. (2018) proposed a framework through which to learn a time-varying regularization parameter in a streaming scenario, named real-time adaptive penalization (RAP). The proposed algorithm is motivated by adaptive filtering theory (Haykin, 2008) and seeks to iteratively update the regularization parameter via stochastic gradient descent in the following manner

$$\lambda_{t+1} = \lambda_t - e \frac{\partial \|Y_{t+1} - X_{t+1}^\top \widehat{\beta}(\lambda_t)\|_2^2}{\partial \lambda_t}, \tag{2}$$

where e denotes the pre-specified step-size parameter of the gradient method. Note that in equation (2) and in the following we clearly denote the dependence of the estimated regression coefficients on λ. In related work, Zboňáková et al. (2017) focus on the choice of the regularization parameter in the context of a quantile regression model. They propose the use of sliding windows and information theoretic quantities to select the associated regularization parameter.

Formally, Osborne et al. (2000) 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:

$$\lambda = \frac{\{Y - X\widehat{\beta}(\lambda)\}^\top X\widehat{\beta}(\lambda)}{||\widehat{\beta}(\lambda)||_1}, \tag{3}$$

where we have ignored the weights, w_t.

As a result, we observe three main effects driving the optimal choice of the regularization parameter.

1. Variance or magnitude of the residuals, $Y - X\widehat{\beta}(\lambda)$. As the variance of residuals increases so does the associated regularization parameter, leading to an increase in sparsity of $\widehat{\beta}(\lambda)$. This is natural as an increase of the variance of residuals is indicative of a drop in the signal-to-noise ratio of the data.

2. The L_1- or L_0-norm of the model coefficients, $||\widehat{\beta}(\lambda)||_1$. As this term appears in the denominator of equation (3), 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.

3. Covariance structure of the design matrix, X. The term related to the covariance structure of the design matrix, $X^\top X$, can be extracted from the elements in the numerator of equation (3). 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, equation (3) 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.

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 the 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 $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$, such that $\varepsilon_t \sim (0, \sigma_t^2)$. The p-dimensional vector of predictor variables X_t was generated from the normal distribution $\mathrm{N}_p(0, \Sigma)$, 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 $i, j = 1, \ldots, p$, with a correlation parameter ρ. 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 $p = 20$. The optimal choice of the regularization parameter (together with associated regression coefficients) was estimated using three distinct methods. We consider the use of the sliding window method in combination with both the Bayesian information criterion (BIC) and the generalized cross-validation (GCV) to select the associated regularization parameter. This means setting some of the weights from (1) to 1 and rest to 0, depending on the window size. Finally, the RAP method proposed by Monti et al. (2018), with the regularization parameter as in (2) is also considered. In the latter, we employed a fixed forgetting factor, r, of size 0.95 and thereby adjusted the weights w_t from the objective function (1) to $w_t = \sum_{i=1}^{t} 0.95^{t-i}$. A burn-in period of 50 observations was employed to obtain an initial estimate for 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 the residual variance on the choice of the regularization parameter λ. The regression coefficients were set to $\beta_t = (1, 1, 1, 1, 1, 0, \ldots, 0)^\top$, 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

$$\varepsilon_t \sim \begin{cases} \mathrm{N}(0, \sigma_1^2), & \text{for} \quad t < 200; \\ \mathrm{N}(0, \sigma_2^2), & \qquad\quad t \geq 200, \end{cases}$$

resulting in a significant change in the variance of the residuals at the 200th observation. Throughout these experiments we set $\sigma_1 = 1$ and allowed σ_2 to vary from $\sigma_2 \in \{1.1, \ldots, 2\}$.

In Figure 1, one can see the effect of the changes in the standard deviation on the Lasso parameter λ. As expected when looking at the formula (3), there is a linear dependence visible. In the case of the BIC and GCV as selection criteria for the values of λ, the line is almost identical. For the RAP algorithm, λ changes slower, but the effect can be clearly seen.

In order to illustrate how the series of values of the Lasso parameter changes over time

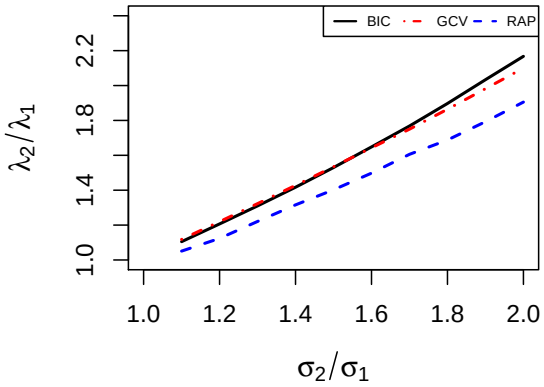

Figure 1: Relative changes of λ in dependence on relative changes of the standard deviation, σ, for BIC (solid), GCV (dot-dashed) and RAP (dashed) method.

TVRPchangeSQR

and how long it takes to adjust for the new settings of the model, we depict the average λ over the 100 scenarios in Figure 2, where $\sigma_1 = 1$ and $\sigma_2 = 1.5$. Since the BIC and GCV yield very similar results, we omit the GCV in this case and we normalize the BIC and RAP values of λ to fit into the interval [0, 1].

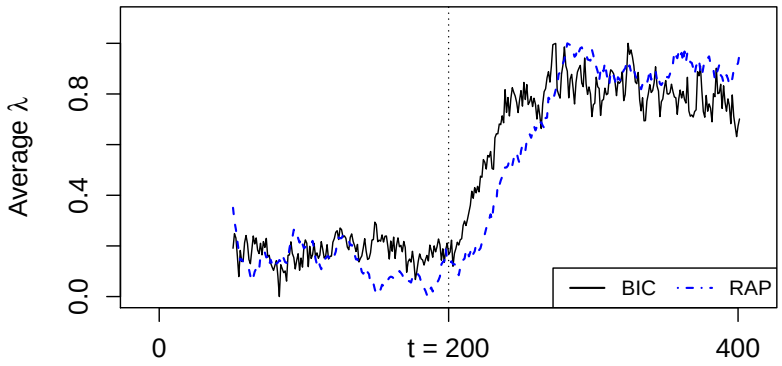

Figure 2: Standardized series of average λ over 100 scenarios with a change point at $t_0 = 200$ and $\sigma_1 = 1$ and $\sigma_2 = 1.5$, for BIC (solid) and RAP (dashed) method.

TVRPchangeSQR

From Figure 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 the BIC is implemented and for the RAP algorithm the adjustment is dependent on the size of the fixed forgetting factor, r. Nevertheless, the changes are obvious and confirm the drawback of using a pre-specified value of λ for the whole data sample.

3.1.2 Change of the L_1- and L_0-norm of β

In the case of changing either L_1- or L_0-norm of the parameter vector β, we put $\sigma_1 = \sigma_2 = 1$ and $\rho = 0.5$. For the first example, the change of L_1-norm, we generated β_t as

$$\beta_t = \begin{cases} (1,1,1,1,1,0,\ldots,0)^\top, & \text{for} \quad t < 200; \\ (1,0.8,0.6,0.4,0.2,0,\ldots,0)^\top, & t \geq 200. \end{cases} \tag{4}$$

The time series of estimated λ values is presented in Figure 3.

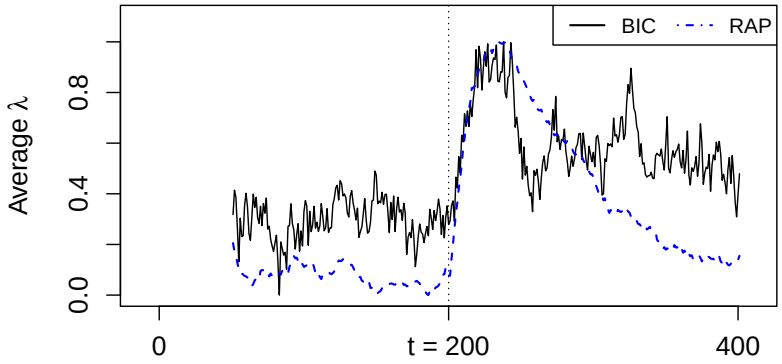

Figure 3: Standardized series of average λ over 100 scenarios with a change point at $t_0 = 200$ and $\beta_{1,2}$ defined by (4), for BIC (solid) and RAP (dashed) method.
TVRPchangeB

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 parameter choice visible in the long run. For the short period after the change, exactly the period of 50 observations in 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

$$\|\beta_t\|_0 = \begin{cases} q_1, & \text{for} \quad t < 200; \\ q_2, & t \geq 200, \end{cases}$$

with $q_1 = 5$ and $q_2 \in \{6,\ldots,10,15\}$.

From Figure 4, where the relative changes of λ in dependence on the relative changes of the size of the active set q can be found, there is a decay of the values of λ visible. This figure provides an empirical validation of the inverse relationship between the magnitude of the active set and the estimated regularization parameter.

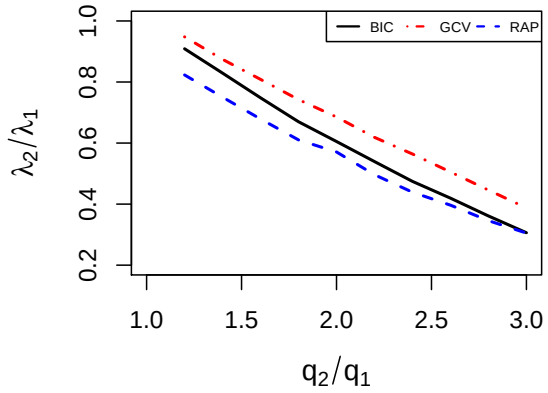

Figure 4: Relative changes of λ in dependence on relative changes of the size of the active set q, for BIC (solid), GCV (dot-dashed) and RAP (dashed) method.
TVRPchangeSQR

The time series of the estimated regularization parameter, as inferred either via the BIC or using the RAP algorithm, are visualized in Figure 5. It is interesting to observe, that for the BIC case there is a visible upward turn of λ values, which is of size exactly as long as the moving window length, as it was in the case of the change in the L_1-norm. This can be explained by the model misspecification when the observation with the change point is a part of the window. More specifically, within this period of time, the sliding window contains the data from both distributions, implying that a correctly specified model would need to account for this mixture. The misspecification of the proposed model leads to an increase in the magnitude of the residuals, which in turn drives the increase in λ. Interestingly, the RAP method deals with the misspecification differently and there is no upward 'bump-shape' visible. In both cases, however, the final values of λ decrease as the size of the active set increases.

3.1.3 Change of the covariance parameter ρ

Finally, we study the effect of changes in the covariance structure on the regularization parameter. We note that while it is possible to vary the covariance structure in many ways, we consider a simple model of covariance structure, where $\Sigma = (\sigma_{ij})_{i,j=1}^{p}$ and set

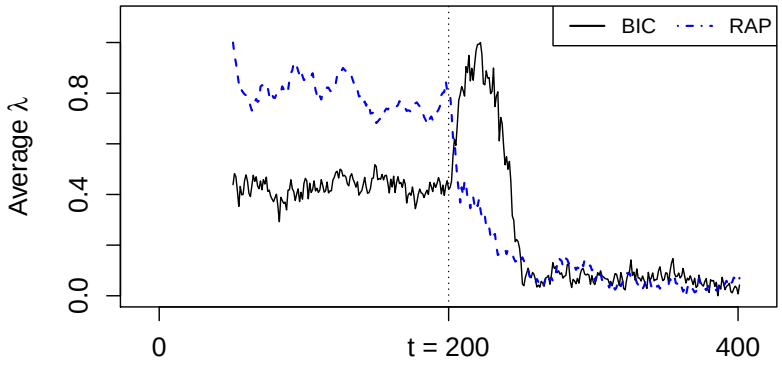

Figure 5: Standardized series of average λ over 100 scenarios with a change point at $t_0 = 200$ and $q_1 = 5$ and $q_2 = 10$, for BIC (solid) and RAP (dashed) method.
TVRPchangeSQR

$\sigma_{ij} = \rho^{|i-j|}$. The benefit of such a model is that it only depends on a single parameter, ρ. As such, we consider changes in the covariance parameter ρ, while fixing $\sigma = 1$ and $q = 5$. The data is generated as follows

$$\rho_t = \begin{cases} \rho_1, & \text{for} \quad t < 200; \\ \rho_2, & \qquad t \geq 200, \end{cases}$$

where $\rho_1 = 0.1$ and $\rho_2 \in \{0.2, 0.3, \ldots, 0.9\}$.

As for the previous experiments, we visualize the relative changes of λ with respect to the relative changes in ρ in Figure 6 and the time series of the estimated values of λ over the whole sample size in Figure 7.

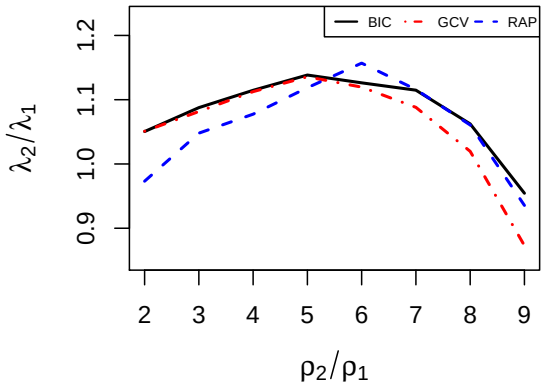

Figure 6: Relative changes of λ in dependence on relative changes of the covariance parameter ρ, for BIC (solid), GCV (dot-dashed) and RAP (dashed) method.
TVRPchangeSQR

From Figure 6, it is important to note that changes of λ no longer show a linear

dependence. For $\rho_2 = 0.2, \ldots, 0.8$ the values of λ tend to rise with a rising covariance of the predictors and the biggest change occurs for $\rho_2 = 0.5$ in the case of the BIC and GCV. In the RAP method example, the values of λ decrease for $\rho_2 = 0.2$ and 0.9 and the biggest change is visible in case that ρ changes to the value of $\rho_2 = 0.6$.

A potential explanation for the non-linear nature of the relationship demonstrated in Figure 6 is due to the selection properties of the Lasso. It is widely acknowledged that in the presence of strongly correlated variables, corresponding to large ρ values, the Lasso tends to choose only a single variable form the group of strongly correlated covariates (indeed this phenomenon is the inspiration for the elastic net (Zou and Hastie, 2005)). Hence, as ρ increases, the term $X^\top X$ from the numerator of λ drives its values higher. If the ρ value is too big, we speak of multicollinearity, where the denominator of λ is affected and becomes larger, which consequently causes the λ values to drop.

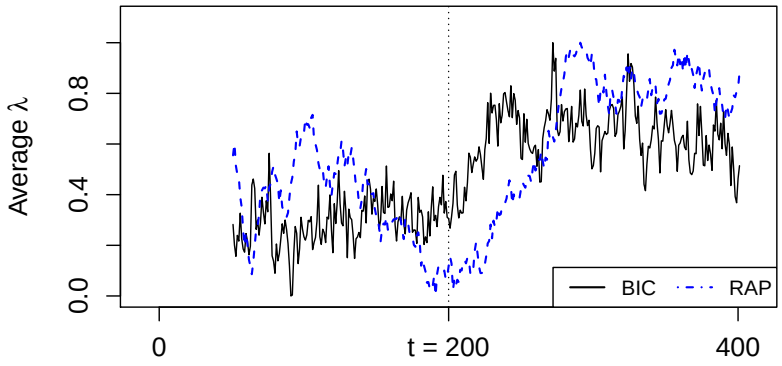

Figure 7: Standardized series of average λ over 100 scenarios with a change point at $t_0 = 200$ and $\rho_1 = 0.1$ and $\rho_2 = 0.5$, for BIC (solid) and RAP (dashed) method.
TVRPchangeSQR

In Figure 7, the change from $\rho_1 = 0.1$ to $\rho_2 = 0.5$ is depicted. We note that there is a change in λ despite the fact that the true 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 show how simultaneous changes to two properties of the data result in cancelling out the effects on the regularization parameter. The purpose of this section is, therefore, to highlight the fact that it is possible to

have a non-stationary data, where the three properties discussed previously 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 parameters β as well as changes in the variance of residuals, σ^2. Recall that the optimal choice of the regularization parameter was positively correlated with the magnitude of the residuals (see Figure 1) while being negatively correlated with q (see Figure 4). The results are presented in Figure 8. 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, whereas the changes in the L_0- or L_1-norm affect the denominator in (3).

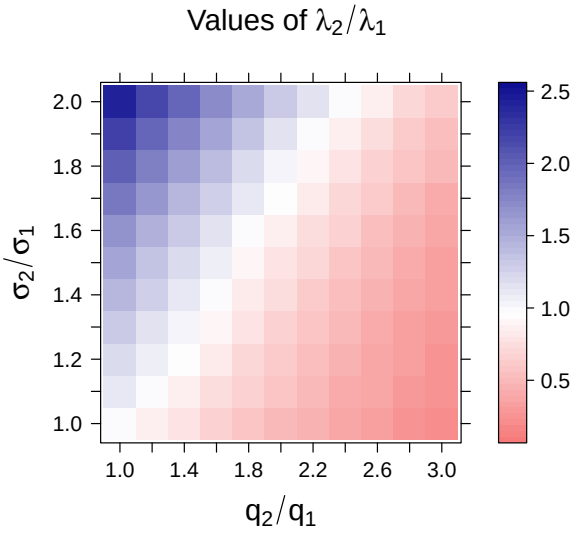

Figure 8: Relative changes of λ corresponding to the combination of relative changes of q and σ.

TVRPchangeSQR

It is important to mention, that the ratio of the values of λ before and after the change was computed by leaving out 50 observations right after the change point. In a long run, the Lasso parameter tends to stabilize itself around a specific value, but shortly after the change, one can clearly see the changes in its pattern. This is illustrated in Figure 9 for the BIC and RAP methods, where we simulated such combination of changes in q and σ which yielded the smallest change in the ratio of the values of λ, i.e. $q_1 = 5$, $q_2 = 9$ and $\sigma_1 = 1$, $\sigma_2 = 1.6$ with $\rho_1 = \rho_2 = 0.1$.

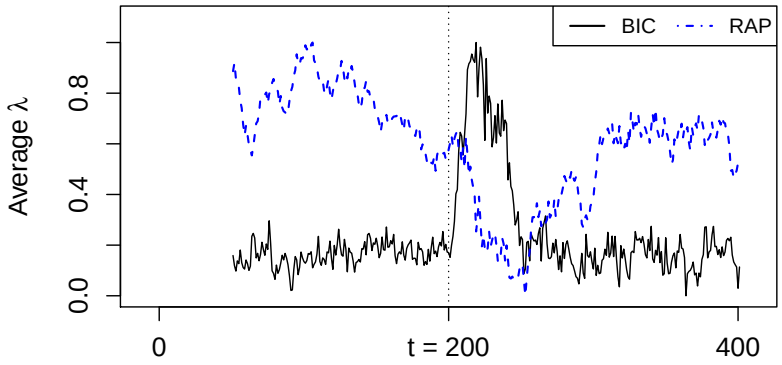

Figure 9: Standardized series of average λ over 100 scenarios with a change point at $t_0 = 200$ and $q_1 = 5$, $q_2 = 9$, $\sigma_1 = 1$, $\sigma_2 = 1.6$ and $\rho_1 = \rho_2 = 0.1$, for BIC (solid) and RAP (dashed) method.

TVRPchangeSQR

Interestingly, the pattern of the BIC method is reversed in comparison to the pattern of the RAP, but in both of the methods under consideration, there is a short-term change in the values of λ clearly visible.

Furthermore, we also 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 parameter ρ did not have a linear relationship with the regularization parameter λ. A similar non-linear relationship can be seen in Figure 10.

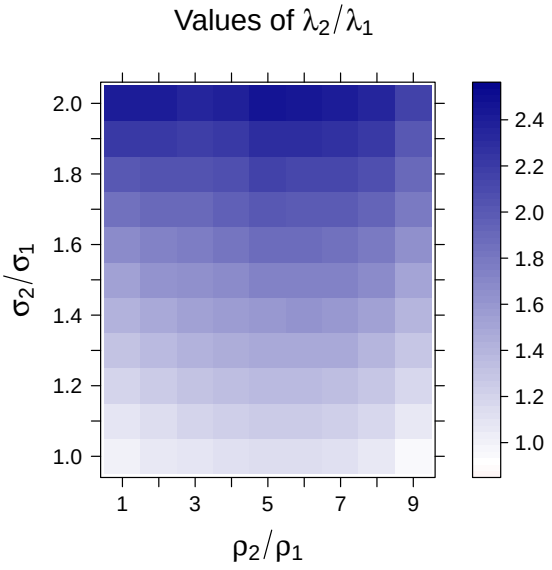

Figure 10: Relative changes of λ corresponding to the combination of relative changes of ρ and σ.

TVRPchangeSQR

The pattern of the ratio of the values of λ before and after the change stays similar for all of the fixed values of σ. With changes in σ, the Lasso parameter tends to rise linearly, as was seen before in Figure 1. Moreover, we note that the changes in σ tend to dominate the changes in the covariance parameter ρ, resulting in the occurrence of the most significant changes of λ whenever the changes in σ are large. Selecting a combination of changes which yields the smallest ratio, we created the plot of Figure 11, where, again, the pattern of the values of λ changes in the short term after the structural break.

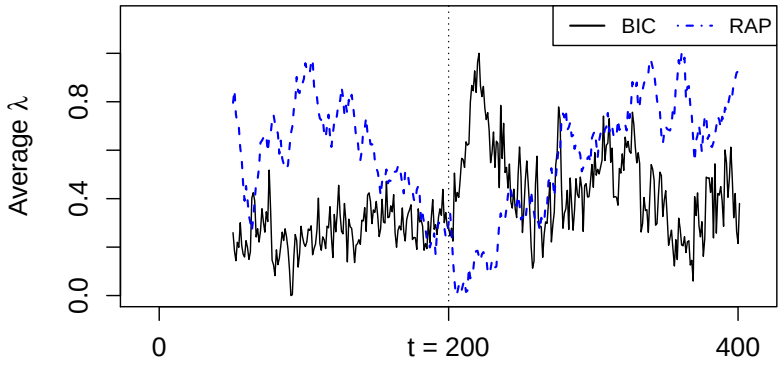

Figure 11: Standardized series of average λ over 100 scenarios with a change point at $t_0 = 200$ and $\rho_1 = 0.1$, $\rho_2 = 0.9$, $\sigma_1 = 1$, $\sigma_2 = 1.1$ and $q_1 = q_2 = 5$, for BIC (solid) and RAP (dashed) method.

TVRPchangeSQR

Finally, we also consider the combination of changes in the L_0-norm together with changes in the covariance parameter ρ. Note, that the changes in these parameters are strongly coupled due to the effects of multicollinearity induced by simultaneously increasing the number of non-zero regression coefficients together with their correlations. The results, provided in Figure 12, highlight these dependencies. For the values of ρ near $\rho = 0.5$, there are some combinations which cancel each other. For the extreme parts of the heatmap, e.g. $\rho_2 = 0.2$ or $\rho_2 = 0.9$, the pattern is clearly driven by the changes in the active set only.

For a better illustration of the changes, we include Figure 13, where the pattern of λ values for the combination of changes with the smallest effect on the Lasso parameter in the long run is considered.

Similarly to what was observed in the previous figures, there is a clear change in the

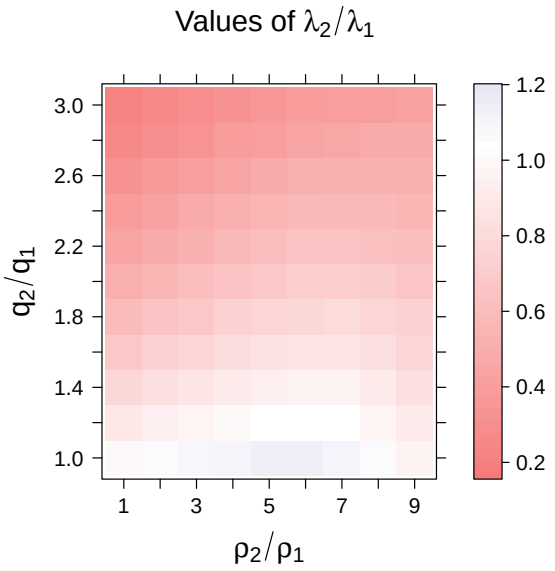

Figure 12: Relative changes of λ corresponding to the combination of relative changes of q and ρ.

TVRPchangeSQR

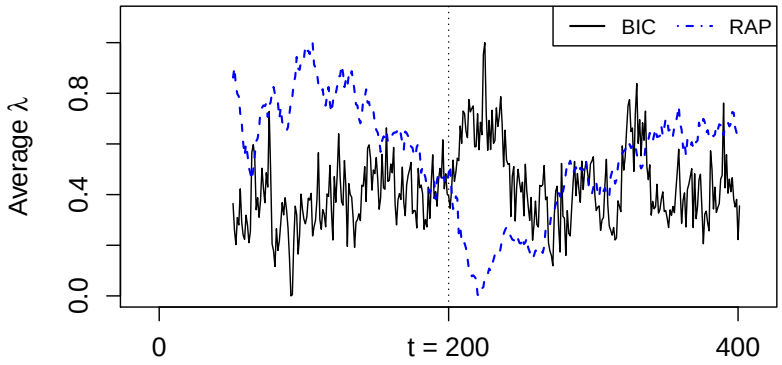

Figure 13: Standardized series of average λ over 100 scenarios with a change point at $t_0 = 200$ and $q_1 = 5$, $q_2 = 6$, $\rho_1 = 0.1$, $\rho_2 = 0.4$ and $\sigma_1 = \sigma_2 = 1$, for BIC (solid) and RAP (dashed) method.

TVRPchangeSQR

pattern in Figure 13, too.

3.2 Application to financial and neuroimaging data

Until now we have provided extensive empirical evidence based on a variety of simulations, each varying one or more of the statistical properties of the data. However, it is interesting to investigate, whether the patterns of λ values are connected to some specific occasions in a real data analysis.

For this purpose, we consider 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 January 3, 2007, to August 10, 2018, see Table 1. The companies listed on NASDAQ are ordered by the market capitalization and downloaded from Yahoo Finance. This sample is particularly interesting as it covers the financial crisis of 2008 and 2009. By analysing this data, it is hoped that we may be able to understand the statistical properties which directly precede similar financial crises, thereby potentially providing some form of advanced warning. The second dataset we consider corresponds to fMRI data collected as part of the Human Connectome Project (HCP). This dataset consists of measurements of 15 distinct brain regions taken during an emotion task, as described in Barch et al. (2013). Data was analysed 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 different tasks (Monti et al., 2017b) or changes across subjects (Monti et al., 2017a).

The modelling procedure for both of the datasets consists of regressing each of the components of the multivariate time series on the rest. This way we get either 100 or 15 sequences of the Lasso parameter values, for the financial and neuroimaging data respectively, which are then averaged and normalized to the [0, 1] interval as before. The resulting time series for the US stock market data are depicted in Figure 14 and for the fMRI data the graphical output can be seen in Figure 15.

From Figure 14 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 as a selecting rule 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.

Figure 15 shows the time series of the average regularization parameter over eight

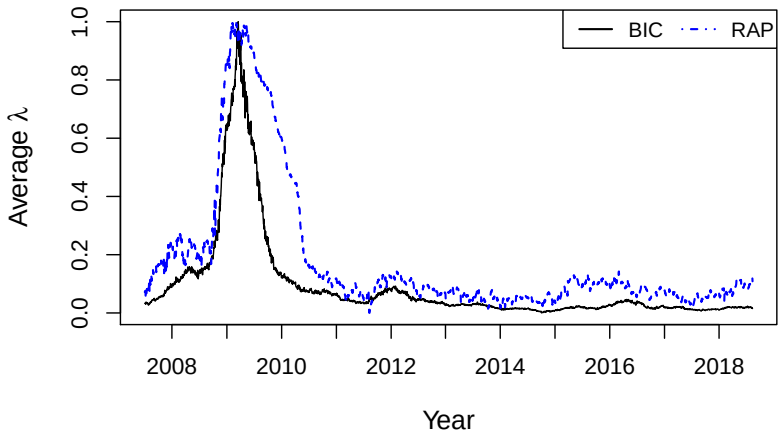

Figure 14: Standardized series of average λ in the US stock returns data, daily observations from January 3, 2007, to August 10, 2018, for BIC (solid) and RAP (dashed) method.

TVRPfrm

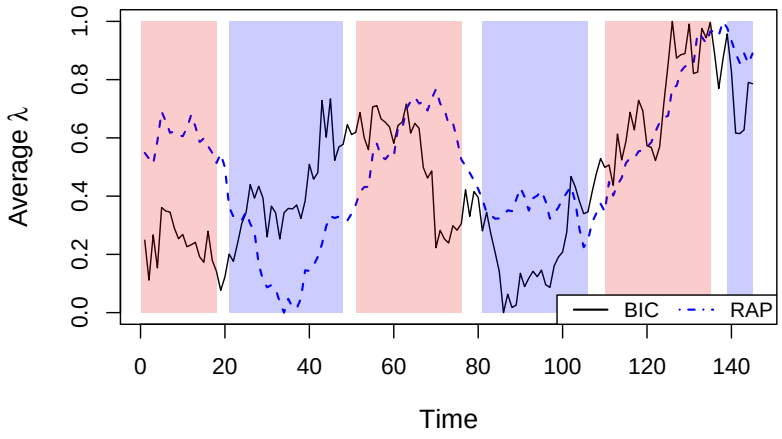

Figure 15: Standardized series of average λ in the fMRI dataset, for BIC (solid) and RAP (dashed) method. Distinct tasks are indicated by the background colour (red indicates a neutral task, blue indicates an emotion task and white denotes the resting period).

TVRPfmri

subjects completing an emotion related task. The task required participants to perform a series of trails presented in blocks. The trails either required them to decide 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 to be the emotion task (denoted in blue in Figure 15) and the latter the neutral task (denoted in red in Figure 15). From Figure 15 we see clear changes in the estimated regularization

parameter induced by the 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 a data and how these relate to the changes in a cognitive state (Calhoun et al., 2014).

4 Discussion

In this work, we have highlighted and provided extensive empirical evidence for 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 Osborne et al. (2000). 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 Figure 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 Figure 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 Figure 6.

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 Figure 8.

Finally, we conclude by 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. Further, the specific pattern observable throughout the time series of the penalty parameter might be of interest if one considers change point detection related problems.

References

Aggarwal, C. C. (2007). *Data Streams: Models and Algorithms*. Springer.

Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., Glasser, M. F., Curtiss, S., Dixit, S., Feldt, C., Nolan, D., Bryant, E., Hartley, T., Footer, O., Bjork, J. M., Poldrack, R., Smith, S., Johansen-Berg, H., Snyder, A. Z., and VanEssen, D. C. (2013). Function in the Human Connectome: Task-fMRI and Individual Differences in Behavior. *Neuroimage*, 80:169–189.

Bottou, L. (2010). Large-Scale Machine Learning with Stochastic Gradient Descent. In *Proc. COMPSTAT'2010*, pages 177–186.

Calhoun, V. D., Miller, R., Pearlson, G., and Adali, T. (2014). The Chronnectome: Time-Varying Connectivity Networks as the Next Frontier in fMRI Data Discovery. *Neuron*, 84:262–274.

Chernozhukov, V., Härdle, W. K., Huang, C., and Wang, W. (2018). Lasso-Driven Inference in Time and Space. arXiv preprint arXiv:1806.05081.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12:2121–2159.

Hastie, T., Tibshirani, R., and Hastie, M. W. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC.

Haykin, S. S. (2008). *Adaptive Filter Theory*. Pearson Education.

Monti, R. P., Anagnostopoulos, C., and Montana, G. (2017a). Learning Population and Subject-Specific Brain Connectivity Networks via Mixed Neighborhood Selection. *The Annals of Applied Statistics*, 11:2142–2164.

Monti, R. P., Anagnostopoulos, C., and Montana, G. (2018). Adaptive Regularization for Lasso Models in the Context of Nonstationary Data Streams. *Statistical Analysis and Data Mining: The ASA Data Science Journal*.

Monti, R. P., Lorenz, R., Braga, R. M., Anagnostopoulos, C., Leech, R., and Montana, G. (2017b). Real-Time Estimation of Dynamic Functional Connectivity Networks. *Human Brain Mapping*, 38:202–220.

Osborne, M. R., Presnell, B., and Turlach, B. A. (2000). On the Lasso and its Dual. *Journal of Computational and Graphical Statistics*, 9:319–337.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. Available at http://www.R-project.org/.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, 58:267–288.

Yu, L., Härdle, W. K., Borke, L., and Benshop, T. (2017). FRM: A Financial Risk Meter Based on Penalizing Tail Events Occurrence. SFB 649 Discussion Paper 2017-003.

Zboňáková, L., Härdle, W. K., and Wang, W. (2017). Time Varying Quantile Lasso. In *Appplied Quantitative Finance, 3rd ed.*, pages 331–353. Springer.

Zou, H. and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society: Series B*, 67:301–320.

JPM	JP Morgan Chase & Co.	RF	Regions Financial Corporation
BAC	Bank of America Corporation	IX	Orix Corp Ads
WFC	Wells Fargo & Company	FITB	Fifth Third Bancorp
HSBC	HSBC Holdings PLC	WLTW	Willis Towers Watson Public Limited Company
C	Citigroup Inc.	HIG	Hartford Financial Services Group, Inc. (The)
RY	Royal Bank of Canada	BAP	Credicorp Ltd.
TD	Toronto Dominion Bank (The)	SHG	Shinhan Financial Group Co Ltd
HDB	HDFC Bank Limited	KB	KB Financial Group, Inc.
AXP	American Express Company	HBAN	Huntington Bancshares Incorporated
USB	U.S. Bancorp	BEN	Franklin Resources, Inc.
GS	Goldman Sachs Group, Inc. (The)	SIVB	SVB Financial Group
BLK	BlackRock, Inc.	CMA	Comerica Incorporated
ITUB	Itau Unibanco Holding S.A.	MKL	Markel Corporation
WBK	Westpac Banking Corporation	L	Loews Corporation
BNS	Bank of Nova Scotia (The)	ETFC	E*TRADE Financial Corporation
LFC	China Life Insurance Company Limited	BCH	Banco De Chile
SCHW	The Charles Schwab Corporation	NMR	Nomura Holdingd, Inc. ADR
PUK	Prudential PLC	EFX	Equifax, Inc.
LYG	Lloyds Banking Group PLC	PFG	Principal Financial Group Inc.
BBD	Banco Bradesco S.A.	BSAC	Banco Santander Chile
SMFG	Sumitomo Mitsui Financial Group, Inc.	XL	XL Group Ltd.
ING	ING Group, N.V.	LNC	Lincoln National Corporation
BK	Bank Of New York Mellon Corporation (The)	RJF	Raymond James Financial, Inc.
SPGI	S&P Global Inc.	AJG	Arhtur J. Gallagher & Co.
BMO	Bank of Montreal (BMO)	AEG	Aegon NV
COF	Capital One Financial Corporation	ACGL	Arch Capital Group Ltd.
MFG	Mizuho Financial Group, Inc.	ROL	Rollins, Inc.
BBVA	Banco Bilbao Viscaya Argentaria S.A.	CINF	Cincinnati Financial Corporation
MMC	Marsh & McLennan Companies, Inc.	FNF	Fidelity National Financial, Inc.
BCS	Barclays PLC	CIB	BanColombia S.A.
PRU	Prudential Financial, Inc.	ZION	Zions Bancorporation
CM	Canadian Imperial Bank of Commerce	AFG	American Financial Group, In.c
CS	Credit Suisse Group	TMK	Torchmark Corporation
PGR	Progressive Corporation (The)	Y	Alleghany Corporation
MFC	Manulife Financial Corp	SEIC	SEI Investments Company
AFL	Aflac Incorporated	EWBC	East West Bancorp, Inc.
ALL	Allstate Corporation (The)	WRB	W.R. Berkley Corporation
AON	Aon PLC	RE	Everest Re Group, Ltd.
TRV	The Travelers Companies, Inc.	CACC	Credit Acceptance Corporation
STI	SunTrust Banks, Inc.	BRO	Brown & Brown, Inc.
AMTD	TD Ameritrade Holding Corporation	AMG	Affiliated Managers Group, Inc.
MCO	Moodys Corporation	UNM	Unum Group
STT	State Street Corporation	CBSH	Commerce Bancshares, Inc.
IBN	ICICI Bank Limited	CFR	Cullen/Frost Bankers, Inc.
TROW	T. Rowe Price Group, Inc.	MKTX	MarketAxess Holdings, Inc.
MTB	M&T Bank Corporation	AIZ	Assurant, Inc.
DB	Deutsche Bank AG	BOKF	BOK Financial Corporation
NTRS	Northern Trust Corporation	ORI	Old Republic International Corporation
SLF	Sun Life Financial, Inc.	PACW	PacWest Bancorp
KEY	KeyCorp	PBCT	People's United Financial, Inc.

Table 1: List of 100 largest financial companies listed on NASDAQ (accessed in August 2018).

IRTG 1792 Discussion Paper Series 2018

For a complete list of Discussion Papers published, please visit irtg1792.hu-berlin.de.

001 "Data Driven Value-at-Risk Forecasting using a SVR-GARCH-KDE Hybrid" by Marius Lux, Wolfgang Karl Härdle and Stefan Lessmann, January 2018.

002 "Nonparametric Variable Selection and Its Application to Additive Models" by Zheng-Hui Feng, Lu Lin, Ruo-Qing Zhu asnd Li-Xing Zhu, January 2018.

003 "Systemic Risk in Global Volatility Spillover Networks: Evidence from Option-implied Volatility Indices " by Zihui Yang and Yinggang Zhou, January 2018.

004 "Pricing Cryptocurrency options: the case of CRIX and Bitcoin" by Cathy YH Chen, Wolfgang Karl Härdle, Ai Jun Hou and Weining Wang, January 2018.

005 "Testing for bubbles in cryptocurrencies with time-varying volatility" by Christian M. Hafner, January 2018.

006 "A Note on Cryptocurrencies and Currency Competition" by Anna Almosova, January 2018.

007 "Knowing me, knowing you: inventor mobility and the formation of technology-oriented alliances" by Stefan Wagner and Martin C. Goossen, February 2018.

008 "A Monetary Model of Blockchain" by Anna Almosova, February 2018.

009 "Deregulated day-ahead electricity markets in Southeast Europe: Price forecasting and comparative structural analysis" by Antanina Hryshchuk, Stefan Lessmann, February 2018.

010 "How Sensitive are Tail-related Risk Measures in a Contamination Neighbourhood?" by Wolfgang Karl Härdle, Chengxiu Ling, February 2018.

011 "How to Measure a Performance of a Collaborative Research Centre" by Alona Zharova, Janine Tellinger-Rice, Wolfgang Karl Härdle, February 2018.

012 "Targeting customers for profit: An ensemble learning framework to support marketing decision making" by Stefan Lessmann, Kristof Coussement, Koen W. De Bock, Johannes Haupt, February 2018.

013 "Improving Crime Count Forecasts Using Twitter and Taxi Data" by Lara Vomfell, Wolfgang Karl Härdle, Stefan Lessmann, February 2018.

014 "Price Discovery on Bitcoin Markets" by Paolo Pagnottoni, Dirk G. Baur, Thomas Dimpfl, March 2018.

015 "Bitcoin is not the New Gold - A Comparison of Volatility, Correlation, and Portfolio Performance" by Tony Klein, Hien Pham Thu, Thomas Walther, March 2018.

016 "Time-varying Limit Order Book Networks" by Wolfgang Karl Härdle, Shi Chen, Chong Liang, Melanie Schienle, April 2018.

017 "Regularization Approach for NetworkModeling of German EnergyMarket" by Shi Chen, Wolfgang Karl Härdle, Brenda López Cabrera, May 2018.

018 "Adaptive Nonparametric Clustering" by Kirill Efimov, Larisa Adamyan, Vladimir Spokoiny, May 2018.

019 "Lasso, knockoff and Gaussian covariates: a comparison" by Laurie Davies, May 2018.

IRTG 1792, Spandauer Straße 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.

IRTG 1792 Discussion Paper Series 2018

For a complete list of Discussion Papers published, please visit irtg1792.hu-berlin.de.

020 "A Regime Shift Model with Nonparametric Switching Mechanism" by Haiqiang Chen, Yingxing Li, Ming Lin and Yanli Zhu, May 2018.

021 "LASSO-Driven Inference in Time and Space" by Victor Chernozhukov, Wolfgang K. Härdle, Chen Huang, Weining Wang, June 2018.

022 " Learning from Errors: The case of monetary and fiscal policy regimes" by Andreas Tryphonides, June 2018.

023 "Textual Sentiment, Option Characteristics, and Stock Return Predictability" by Cathy Yi-Hsuan Chen, Matthias R. Fengler, Wolfgang Karl Härdle, Yanchu Liu, June 2018.

024 "Bootstrap Confidence Sets For Spectral Projectors Of Sample Covariance" by A. Naumov, V. Spokoiny, V. Ulyanov, June 2018.

025 "Construction of Non-asymptotic Confidence Sets in 2 -Wasserstein Space" by Johannes Ebert, Vladimir Spokoiny, Alexandra Suvorikova, June 2018.

026 "Large ball probabilities, Gaussian comparison and anti-concentration" by Friedrich Götze, Alexey Naumov, Vladimir Spokoiny, Vladimir Ulyanov, June 2018.

027 "Bayesian inference for spectral projectors of covariance matrix" by Igor Silin, Vladimir Spokoiny, June 2018.

028 "Toolbox: Gaussian comparison on Eucledian balls" by Andzhey Koziuk, Vladimir Spokoiny, June 2018.

029 "Pointwise adaptation via stagewise aggregation of local estimates for multiclass classification" by Nikita Puchkin, Vladimir Spokoiny, June 2018.

030 "Gaussian Process Forecast with multidimensional distributional entries" by Francois Bachoc, Alexandra Suvorikova, Jean-Michel Loubes, Vladimir Spokoiny, June 2018.

031 "Instrumental variables regression" by Andzhey Koziuk, Vladimir Spokoiny, June 2018.

032 "Understanding Latent Group Structure of Cryptocurrencies Market: A Dynamic Network Perspective" by Li Guo, Yubo Tao and Wolfgang Karl Härdle, July 2018.

033 "Optimal contracts under competition when uncertainty from adverse selection and moral hazard are present" by Natalie Packham, August 2018.

034 "A factor-model approach for correlation scenarios and correlation stress-testing" by Natalie Packham and Fabian Woebbeking, August 2018.

035 "Correlation Under Stress In Normal Variance Mixture Models" by Michael Kalkbrener and Natalie Packham, August 2018.

036 "Model risk of contingent claims" by Nils Detering and Natalie Packham, August 2018.

037 "Default probabilities and default correlations under stress" by Natalie Packham, Michael Kalkbrener and Ludger Overbeck, August 2018.

038 "Tail-Risk Protection Trading Strategies" by Natalie Packham, Jochen Papenbrock, Peter Schwendner and Fabian Woebbeking, August 2018.

IRTG 1792, Spandauer Straße 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.

IRTG 1792 Discussion Paper Series 2018

For a complete list of Discussion Papers published, please visit irtg1792.hu-berlin.de.

039 "Penalized Adaptive Forecasting with Large Information Sets and Structural Changes" by Lenka Zbonakova, Xinjue Li and Wolfgang Karl Härdle, August 2018.

040 "Complete Convergence and Complete Moment Convergence for Maximal Weighted Sums of Extended Negatively Dependent Random Variables" by Ji Gao YAN, August 2018.

041 "On complete convergence in Marcinkiewicz-Zygmund type SLLN for random variables" by Anna Kuczmaszewska and Ji Gao YAN, August 2018.

042 "On Complete Convergence in Marcinkiewicz-Zygmund Type SLLN for END Random Variables and its Applications" by Ji Gao YAN, August 2018.

043 "Textual Sentiment and Sector specific reaction" by Elisabeth Bommes, Cathy Yi-Hsuan Chen and Wolfgang Karl Härdle, September 2018.

044 "Understanding Cryptocurrencies" by Wolfgang Karl Härdle, Campbell R. Harvey, Raphael C. G. Reule, September 2018.

045 "Predicative Ability of Similarity-based Futures Trading Strategies" by Hsin-Yu Chiu, Mi-Hsiu Chiang, Wei-Yu Kuo, September 2018.

046 "Forecasting the Term Structure of Option Implied Volatility: The Power of an Adaptive Method" by Ying Chen, Qian Han, Linlin Niu, September 2018.

047 "Inferences for a Partially Varying Coefficient Model With Endogenous Regressors" by Zongwu Cai, Ying Fang, Ming Lin, Jia Su, October 2018.

048 "A Regime Shift Model with Nonparametric Switching Mechanism" by Haiqiang Chen, Yingxing Li, Ming Lin, Yanli Zhu, October 2018.

049 "Strict Stationarity Testing and GLAD Estimation of Double Autoregressive Models" by Shaojun Guo, Dong Li, Muyi Li, October 2018.

050 "Variable selection and direction estimation for single-index models via DC-TGDR method" by Wei Zhong, Xi Liu, Shuangge Ma, October 2018.

051 "Property Investment and Rental Rate under Housing Price Uncertainty: A Real Options Approach" by Honglin Wang, Fan Yu, Yinggang Zhou, October 2018.

052 "Nonparametric Additive Instrumental Variable Estimator: A Group Shrinkage Estimation Perspective" by Qingliang Fan, Wei Zhong, October 2018.

053 "The impact of temperature on gaming productivity: evidence from online games" by Xiaojia Bao, Qingliang Fan, October 2018.

054 "Topic Modeling for Analyzing Open-Ended Survey Responses" by Andra-Selina Pietsch, Stefan Lessmann, October 2018.

055 "Estimation of the discontinuous leverage effect: Evidence from the NASDAQ order book" by Markus Bibinger, Christopher Neely, Lars Winkelmann, October 2018.

056 "Cryptocurrencies, Metcalfe's law and LPPL models" by Daniel Traian Pele, Miruna Mazurencu-Marinescu-Pele, October 2018.

057 "Trending Mixture Copula Models with Copula Selection" by Bingduo Yang, Zongwu Cai, Christian M. Hafner, Guannan Liu, October 2018.

IRTG 1792, Spandauer Straße 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.

IRTG 1792 Discussion Paper Series 2018

For a complete list of Discussion Papers published, please visit irtg1792.hu-berlin.de.

058 "Investing with cryptocurrencies – evaluating the potential of portfolio allocation strategies" by Alla Petukhina, Simon Trimborn, Wolfgang Karl Härdle, Hermann Elendner, October 2018.

059 "Towards the interpretation of time-varying regularization parameters in streaming penalized regression models" by Lenka Zbonakova, Ricardo Pio Monti, Wolfgang Karl Härdle, October 2018.

IRTG 1792, Spandauer Straße 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.