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

Lenka Zboňáková, Ricardo Pio Monti, Wolfgang Karl Härdle

November 27, 2019

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.

1 Introduction

High-dimensional, streaming datasets pose a unique challenge to modern statisticians. To date, the challenges associated with high-dimensional and

*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.

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., 2017b].

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 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 time-varying regularization parameters as well as the associated 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 fixed regularization parameter, λ, time-varying regression coefficients can be estimated by minimizing the following convex objective:

$$L_t(\beta, \lambda) = \sum_{i=1}^{t} w_i \left(y_i - X_i^\top \beta\right)^2 + \lambda ||\beta||_1, \tag{1}$$

where $w_i > 0$ are weights indicating the importance given to past observations [Aggarwal, 2007]. For example, it is natural to allow w_i to decay monotonically in a manner which is proportional to the chronological proximity of the ith 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.

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 [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 time-varying regularization parameter in a streaming scenario. The proposed framework is motivated by adaptive filtering theory [Haykin, 2008] and seeks to iteratively update the regularization parameter via stochastic gradient descent. 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:

$$\lambda = \frac{\{\mathbf{Y} - \mathbf{X}\hat{\beta}(\lambda)\}^\top \mathbf{X}\hat{\beta}(\lambda)}{||\hat{\beta}(\lambda)||_1}, \tag{2}$$

where we have ignored the weights, w_i, and use bold notation to denote vectors and matrices respectively. Note that in equation (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 so does the associated regularization parameter, leading to an increase in sparsity of $\hat{\beta}(\lambda)$. 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 norm of the model coefficients: $||\hat{\beta}(\lambda)||_1$. As this term appears in the denominator of equation (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, 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 equation (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, equation (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.

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 $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$, 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 $(p \times p)$ 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 Bayesian information criterion (BIC) and generalized cross-validation (GCV) to select the associated regularization parameter. Finally, the gradient method proposed by Monti et al. [2018], named Real-time Adaptive Penalization (RAP), is also applied. 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 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:

$$\varepsilon_t \sim \begin{cases} \mathrm{N}(0, \sigma_1^2), & \text{for} \quad t < 200; \\ \mathrm{N}(0, \sigma_2^2), & \qquad\quad t \geq 200, \end{cases} \tag{3}$$

resulting in a significant 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 $\sigma_2 \in \{1.1, \ldots, 2\}$. 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 larger values of the ratio to yield larger changes in the choice of the regularization parameter.

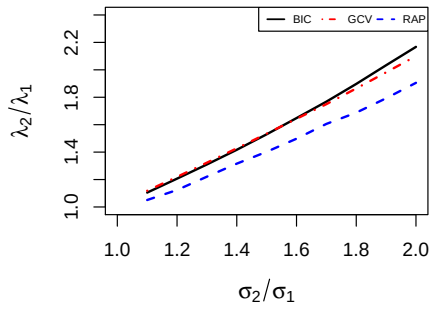

Figure 1: Relative changes of λ in dependence on relative changes of the standard deviation σ.

TVRPchangeSQR

Figure 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 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 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 normalize the BIC and RAP values of λ to fit into the interval [0, 1].

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

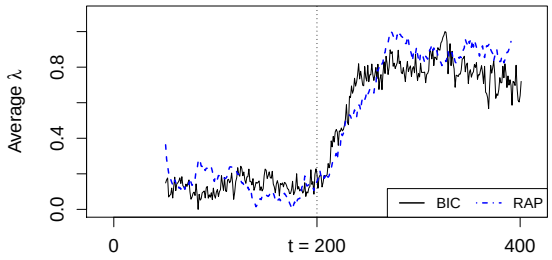

Figure 2: Standardized series of average λ over 100 scenarios with a change point at $t = 200$ and $\sigma_1 = 1$ and $\sigma_2 = 1.5$.

TVRPchangeSQR

is implemented and for the RAP algorithm the adjustment is dependent on the size of the fixed forgetting factor. r. 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:

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

The time series of estimated λ values is presented in Figure 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 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:

$$||\beta_t||_0 = \begin{cases} q_1, & \text{for} \quad t < 200; \\ q_2, & \qquad t \geq 200, \end{cases} \tag{5}$$

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

Figure 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 as q_2/q_1 increases. This is to be expected, as an increase in the specified ratio implies a larger active set in the latter part of the time series. This figure provides empirical validation of the inverse correlation between the magnitude of the active set and the estimated regularization parameter.

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:

$$\rho_t = \begin{cases} \rho_1, & \text{for} \quad t < 200; \\ \rho_2, & \qquad t \geq 200, \end{cases} \tag{6}$$

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

As in the previous experiments, we visualize the relative changes of λ with respect to the relative changes of ρ in Figure 5. The time series of the estimated values of λ over the whole sample size are depicted in Figure 6.

From Figure 5 it is important to note that the changes of λ no longer demonstrate a linear dependency with the statistical property of interest. For

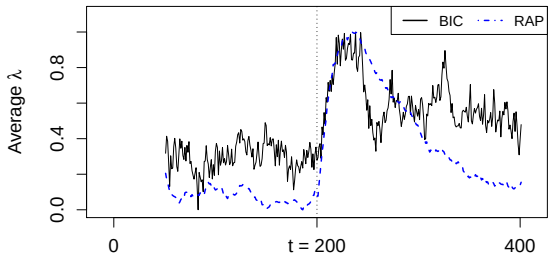

Figure 3: Standardized series of average λ over 100 scenarios with a change point at $t = 200$ and regression coefficients β_t defined by (4).

TVRPchangeB

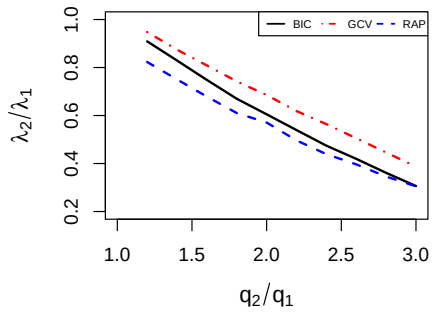

Figure 4: Relative changes of λ in dependence on relative changes of the size of the active set q.

TVRPchangeSQR

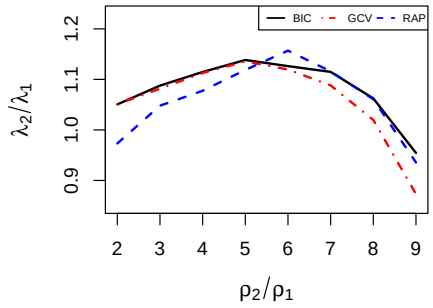

Figure 5: Relative changes of λ in dependence on relative changes of the correlation parameter ρ.

TVRPchangeSQR

$\rho_2 = 0.2, \ldots, 0.8$ 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 the case that ρ changes to the value $\rho_2 = 0.6$.

A potential explanation for the non-linear nature of the relationship demonstrated in Figure 5 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]). As such, as ρ increases, the term $\mathbf{X}^\top\mathbf{X}$ 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 Figure 6 the change from $\rho_1 = 0.1$ to $\rho_2 = 0.5$ 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 *canceling 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 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 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 Figure 1) whilst being negative correlated with q (see Figure 4). Figure 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 Figure 8. Furthermore, we note that changes in σ tend to dominate the changes in ρ with the largest changes in λ occurring for large changes in σ.

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 coef-

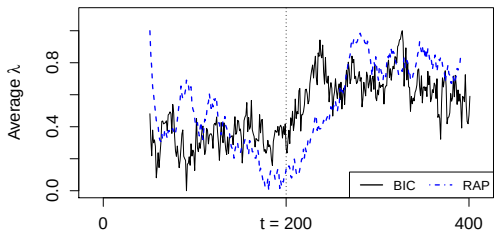

Figure 6: Standardized series of average λ over 100 scenarios with a change point at $t = 200$ and $\rho_1 = 0.1$ and $\rho_2 = 0.5$.

TVRPchangeSQR

ficients together with their correlations. The results, provided in Figure 9, 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 change in the active set only.

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. 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 examples presented in this section provide a clear illustration that non-stationary data are present in a wide range of applications.

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. By analyzing 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 a part of the Human Connectome Project (HCP). This dataset consists of measurements of 15 distinct brain regions taken during an Emotion task, as described

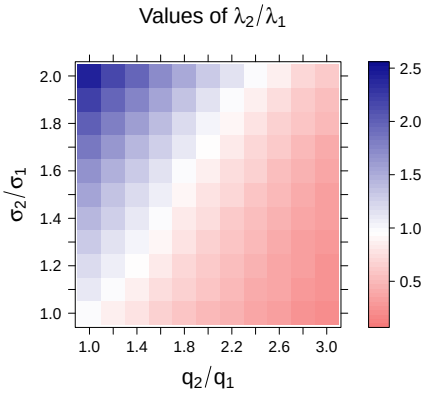

Figure 7: Relative changes of λ corresponding to the combination of relative changes of q and σ.

TVRPchangeSQR

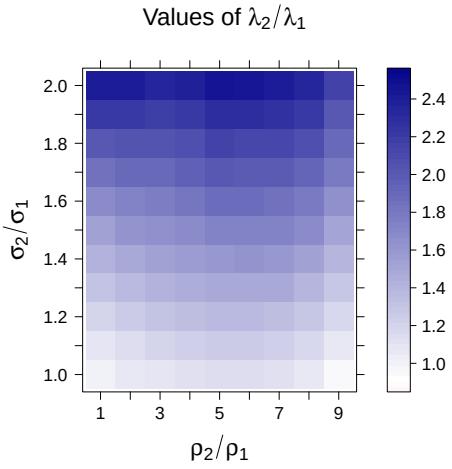

Figure 8: Relative changes of λ corresponding to the combination of relative changes of ρ and σ.

TVRPchangeSQR

in Barch et al. [2013]. 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 [Monti et al., 2017b] or changes

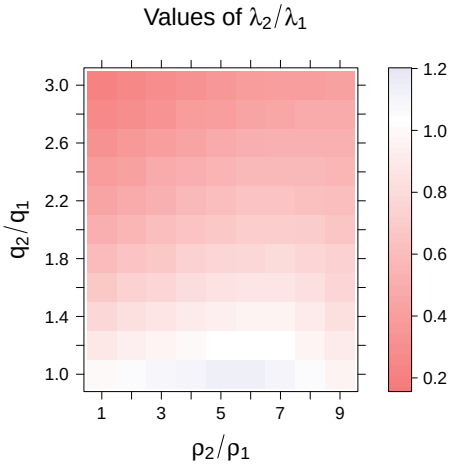

Figure 9: Relative changes of λ corresponding to the combination of relative changes of q and ρ.

TVRPchangeSQR

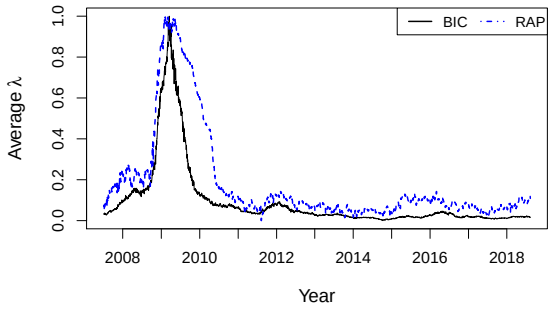

Figure 10: Standardized series of average λ in the US stock returns data, daily observations from January 3, 2007 to August 10, 2018.

TVRPfrm

across subjects [Monti et al., 2017a].

The modelling procedure employed for both of the datasets consisted in regressing each of the components of the multivariate time series on the rest. 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 Figure 10 and for the fMRI data the graphical output can be seen in Figure 11.

From Figure 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 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 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 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 the emotion task (and denoted in blue in Figure 11) and the latter the neutral task (denoted in red in Figure 11). From Figure 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 a data and how these relate to changes in cognitive state [Calhoun et al., 2014].

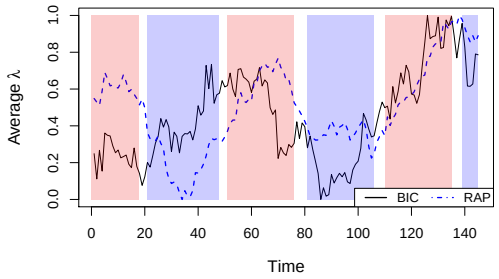

Figure 11: Standardized series of average λ in the fMRI dataset. Distinct tasks are indicated by the background color (red indicates a neutral task, blue indicates an emotion task and white denotes the resting period).

TVRPfmri

4 Discussion

In this work, we have highlighted and provided extensive 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 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 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 Figure 7.

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.

References

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

Deanna M Barch, Gregory C Burgess, Michael P Harms, Steven E Petersen, Bradley L Schlaggar, Maurizio Corbetta, Matthew F Glasser, Sandra Curtiss, Sachin Dixit, Cindy Feldt, et al. Function in the Human Connectome: Task-fMRI and Individual Differences in Behavior. *NeuroImage*, 80:169–189, 2013.

León Bottou. Large-Scale Machine Learning with Stochastic Gradient Descent. *Proc. COMPSTAT'2010*, pages 177–186, 2010. ISSN 0269-2155.

Vince D Calhoun, Robyn Miller, Godfrey Pearlson, and Tulay Adalı. The Chronnectome: Time-Varying Connectivity Networks as the Next Frontier in fMRI Data Discovery. *Neuron*, 84(2):262–274, 2014.

Victor Chernozhukov, Wolfgang K Härdle, Chen Huang, and Weining Wang. Lasso-Driven Inference in Time and Space. *arXiv preprint arXiv:1806.05081*, 2018.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright Hastie. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman&Hall/CRC, 2015.

Simon Haykin. *Adaptive Filter Theory*. Pearson Education, 2008.

Ricardo Pio Monti, Christoforos Anagnostopoulos, and Giovanni Montana. Learning Population and Subject-Specific Brain Connectivity Networks via Mixed Neighborhood Selection. *The Annals of Applied Statistics*, 11(4):2142–2164, 2017a.

Ricardo Pio Monti, Romy Lorenz, Rodrigo M Braga, Christoforos Anagnostopoulos, Robert Leech, and Giovanni Montana. Real-Time Estimation of Dynamic Functional Connectivity Networks. *Human Brain Mapping*, 38(1): 202–220, 2017b.

Ricardo Pio Monti, Christoforos Anagnostopoulos, and Giovanni Montana. Adaptive Regularization for Lasso Models in the Context of Nonstationary Data Streams. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 2018. doi: 10.1002/sam.11390.

Michael R Osborne, Brett Presnell, and Berwin A Turlach. On the Lasso and its Dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337, 2000.

R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. Available at http://www.R-project.org/, 2014.

Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B*, pages 267–288, 1996.

Lining Yu, Wolfgang K Härdle, Lukas Borke, and Thijs Benschop. FRM : a Financial Risk Meter Based on Penalizing Tail Events Occurrence. *SFB 649 Discussion Paper 2017-003*, 2017.

Lenka Zboňáková, Wolfgang K Härdle, and Weining Wang. Time Varying Quantile Lasso. In *Applied Quantitative Finance, 3rd ed.*, pages 331–353, 2017. doi: 10.1007/978-3-662-54486-0.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320, 2005.