STATISTICAL METHODS FOR HIGH-DIMENSIONAL DATA AND CONTINUOUS GLUCOSE MONITORING

BY TINGNI SUN

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Statistics

Written under the direction of

Professor Cun-Hui Zhang

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2012

ABSTRACT OF THE DISSERTATION

Statistical Methods for High-dimensional Data and Continuous Glucose Monitoring

by Tingni Sun

Dissertation Director: Professor Cun-Hui Zhang

This thesis contains two parts. The first part concerns three connected problems with high-dimensional data in Chapters 2-4. The second part, Chapter 5, provides dynamic Bayes models to improve the continuous glucose monitoring.

In the first part, we propose a unified scale invariant method for the estimation of parameters in linear regression, precision matrix and partial correlation. In Chapter 2, scaled Lasso is introduced to jointly estimate regression coefficients and noise level with a gradient descent algorithm. Under mild regularity conditions, we derive oracle inequalities for the prediction and estimation of the noise level and regression coefficients. These oracle inequalities provide sufficient conditions for the consistency and asymptotic normality of the noise level estimator, including certain cases where the number of variables is of greater order than the sample size. Chapter 3 considers the estimation of precision matrix, which is closely related to linear regression. The

proposed estimator is constructed via the scaled Lasso, and guarantees the fastest convergence rate under the spectrum norm. Besides the estimation of high-dimensional objects, the estimation of low-dimensional functionals of high-dimensional objects is also of great interest. A rate minimax estimator of a high-dimensional parameter does not automatically yield rate minimax estimates of its low-dimensional functionals. We consider efficient estimation of partial correlation between individual pairs of variables in Chapter 4. Numerical results demonstrate the superior performance of the proposed methods.

In the second part, we develop statistical methods to produce more accurate and precise estimates for continuous glucose monitoring. The continuous glucose monitor measures the glucose level via an electrochemical glucose biosensor, inserted into subcutaneous fat tissue, called interstitial space. We use dynamic Bayes models to incorporate the linear relationship between the blood glucose level and interstitial signal, the time series aspects of the data, and the variability depending on sensor age. The Bayes method has been tested and evaluated with an important large dataset, called "Star I", from Medtronic, Inc., composed of continuous monitoring of glucose and other measurements. The results show that the Bayesian blood glucose prediction outperforms the output of the continuous glucose monitor in the STAR 1 trial.

Acknowledgements

I am deeply grateful to my advisor, Professor Cun-Hui Zhang, for providing me with an outstanding scientific training, as well as for his unwavering support and constant encouragement. I feel very fortunate to have been working with him and learning a lot from his extensive knowledge and insights. In addition, his devotion to mathematical and statistical sciences is always a great source of inspiration for me.

I would also like to thank the members of my dissertation committee, Professors Lawrence Shepp, Lee Dicker and Yang Feng, for their comments on the manuscript and the time they dedicated to reviewing my thesis. Special thanks go to Professor Lawrence Shepp and Professor Lee Dicker for their helpful discussions on the topics in Chapter 5 of this thesis, concerning the closed-loop diabetes control. I really enjoy working with them on this wonderful project and benefit from talking with them.

Moreover, I want to thank the entire Department of Statistics and Biostatistics for supporting me in various ways and at various occasions, with special thanks to our graduate director, Professor John Kolassa, who provides me a lot of suggestions. My special thanks also go to our department chair, Professor Regina Liu, for her great advices and encouragement during my study, especially when I was struggling with making a decision about my future career. Also, I would like to thank the fellow students in our department and my friends at Rutgers for their suggestions and help in the past five years. All these people make my graduate life happy and unforgettable.

Last but not least, I would like to extend my deepest gratitude to my parents and my husband for their unconditional love and support. They have always been there for me through thick and thin, which has helped me tremendously.

Dedication

To My Parents, Yongkang Sun and Shundi Qiu

and

To My Husband, Kai Liu

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction

1.1 Statistical Inference with High-dimensional Data

The first part of this thesis concerns three connected problems in estimation and statistical inference with high dimensional data: linear regression with unknown variance, precision matrix as a high-dimensional object, and low-dimensional functionals of high-dimensional parameters.

With the development of information technologies, high-dimensional data analysis has become very important in many fields of scientific research and knowledge discovery. Linear regression, as one of simplest statistical models, has been intensively studied in certain high-dimensional settings. A focus of recent research of high-dimensional linear regression has been on the performance of penalization methods, e.g. Lasso, smoothly clipped absolute deviation (SCAD) and minimax concave (MC) penalties, etc. These studies usually require the knowledge of the noise level in linear model. However, it is non-trivial to estimate the variance of the noise when the number of covariates p is larger than the sample size n. In Chapter 2, we propose a "scaled Lasso" methodology to simultaneously estimate the regression coefficients and noise level in linear regression. It chooses an equilibrium with a sparse regression method by iteratively estimating the noise level via the mean residual square and scaling the penalty in proportion to the estimated noise level. The iterative algorithm costs little beyond the computation of a path or grid of the sparse regression estimator for penalty levels above a proper threshold. For the scaled lasso, the algorithm is a

gradient descent in a convex minimization of a penalized joint loss function for the regression coefficients and noise level. Under mild regularity conditions, we derive oracle inequalities for the prediction and estimation performances of the noise level and regression coefficients. These inequalities provide sufficient conditions for the consistency and asymptotic normality of the noise level estimator.

Estimation of inverse covariance matrix, also known as the precision matrix or concentration matrix, is one of the classic problems in multivariate statistics, since the precision matrix has an interpretation in terms of partial correlation. The partial correlation is used to measure the conditional dependency in Gaussian-Markov graphical models that are widely used in network problems of dependencies among variables. Due to the rapid advances of technologies, precision matrix estimation is also a topic of great interests in high-dimensional network problems, such as gene association, social network, etc. In Chapter 3, we propose to estimate each column of the target matrix based on the scaled Lasso, by taking advantage of the relation between linear regression and precision matrix. Under the sparsity condition on matrix degree and mild regularity conditions, we prove that the proposed estimator guarantees the fastest rate of convergence under the spectrum norm in certain high-dimensional settings where the number of variables is of greater order than the sample size. In addition, since the scaled Lasso algorithm provides a fully specified map from the space of nonnegative-definite matrices to the space of symmetric matrices, this estimator could be extended to generate an approximate inverse of a nonnegative data matrix in a general setting.

Most of the recent advances in high-dimensional data have been focused on the estimation of high-dimensional objects as in Chapters 2 and 3. However, the estimation of low-dimensional functionals of high-dimensional parameters is also of great interest. For example, instead of the covariance matrix or its inverse as linear operators, one might be more interested in the relationship between individual pairs of variables.

In Chapter 4, we consider efficient estimation and confidence interval for the partial correlation with high-dimensional Gaussian data.

1.2 Dynamic Bayes Models for Continuous Glucose Monitoring

Closed-loop diabetes control, or artificial pancreas, is a new technology that will revolutionize diabetes management. Although the current technology is mostly developed by electronic and biomedical engineers, statistics will play an ever important role in controlling the accuracy and precision of such systems. Chapter 5 concerns the estimation problem of continuous glucose monitoring, which is one of essential components in artificial pancreas system.

The continuous glucose monitor measures the glucose level via an electrochemical glucose biosensor, inserted into subcutaneous fat tissue, called interstitial space. Motivated by the mechanism of glucose sensor and continuous blood glucose monitor, we propose a statistical framework for modeling the dynamic relationship between the blood glucose level and interstitial signal. At the current stage, our Bayes model also incorporates the time series aspects of the data and the variability depending on sensor age.

The Bayes method has been developed and evaluated with an important large dataset, called "Star I", from Medtronic, Inc., composed of continuous monitoring of glucose and other measurements. The analysis shows that our blood glucose prediction outperforms the output of the continuous glucose monitor in the STAR 1 trial.

Chapter 2
Scaled Sparse Linear Regression

2.1 Introduction

This chapter concerns the simultaneous estimation of the regression coefficients and noise level in a high-dimensional linear model. High-dimensional data analysis is a topic of great current interest due to the growth of applications where the number of unknowns far exceeds the number of data points. Among statistical models arising from such applications, linear regression is one of the best understood. Penalization, convex minimization and thresholding methods have been proposed, tested with real and simulated data, and proved to control errors in prediction, estimation and variable selection under various sets of regularity conditions. These methods typically require an appropriate penalty or threshold level. A larger penalty level may lead to a simple model with large bias, while a smaller penalty level may lead to a complex noisy model due to overfitting. Scale-invariance considerations and existing theory suggest that the penalty level should be proportional to the noise level of the regression model. In the absence of knowledge of the latter level, cross-validation is commonly used to determine the former. However, cross-validation is computationally costly and theoretically poorly understood, especially for the purpose of variable selection and the estimation of regression coefficients. The penalty level selected by cross-validation is called the prediction-oracle in Meinshausen & Bühlmann (2006), which gave an example to show that the prediction-oracle solution does not lead to consistent model selection for the Lasso.

Estimation of the noise level in high-dimensional regression is interesting in its own right. Examples include quality control in manufacturing and volatility control in finance.

Our study is motivated by Städler et al. (2010). They proposed to estimate the regression coefficients and noise level by maximizing their joint log-likelihood with an ℓ_1 penalty on the regression coefficients. Their method has a unique solution due to the joint concavity of the log-likelihood under a certain transformation of the unknown parameters. However, we prove that this penalized joint maximum likelihood estimator may result in a positive bias for the estimation of the noise level. We propose an iterative algorithm that alternates between estimating the noise level via the mean residual square and scaling the penalty level in a predetermined proportion to the estimated noise level in the Lasso or minimax concave penalized selection paths. This part of results has been published in our discussion article, Sun & Zhang (2010).

In the meanwhile, Antoniadis (2010) commented on the same problem from a different perspective by raising the possibility of adding an ℓ_1 penalty to Huber's concomitant joint loss function. See, for example section 7.7 of Huber & Ronchetti (2009). Interestingly, the minimizer of this penalized joint convex loss is identical to the equilibrium of the iterative algorithm for the Lasso path. Thus, the convergence of the iterative algorithm is guaranteed by the convexity. For simplicity, we call the equilibrium of this algorithm the scaled version of the penalized regression method, for example the scaled Lasso or scaled minimax concave penalized selection, depending on the choice of penalty function. Under mild regularity conditions, we prove oracle inequalities for prediction and the joint estimation of the noise level and regression coefficients for the scaled Lasso, that imply the consistency and asymptotic normality of the scaled Lasso estimator for the noise level. We report numerical results on the performance of scaled Lasso and other scaled penalized methods. These theoretical and numerical results strongly support the use of the proposed method for high-dimensional

regression.

This chapter is organized as follows. In Section 2.2, we describe the iterative algorithm and its connection to convex minimization. In Section 2.3, we provide oracle inequalities for the scaled Lasso and prove the consistency and asymptotic normality of the estimator for the noise level. In Section 2.4, we present numerical results. Section 2.5 contains some discussion, including oracle inequalities for the Lasso with a predetermined penalty level. Section 2.6 provides all proofs.

We use the following notation throughout this chapter. For a vector $\boldsymbol{v} = (v_1, \ldots, v_p)$, $\|\boldsymbol{v}\|_q = (\sum_j |v_j|^q)^{1/q}$ denotes the ℓ_q norm with the usual extensions $\|\boldsymbol{v}\|_\infty = \max_j |v_j|$ and $\|\boldsymbol{v}\|_0 = \#\{j : v_j \neq 0\}$. For design matrices X and subsets A of $\{1, \ldots, p\}$, x_j denotes column vectors of X and $\boldsymbol{X}_A$ denotes the matrix composed of columns with indices in set A. Moreover, $x_+ = \max(x, 0)$.

2.2 An iterative algorithm

In this section, we describe the iterative algorithm for the joint estimation of regression coefficients and noise level and its connection to convex minimization.

Suppose we observe a design matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) \in \mathbb{R}^{n \times p}$ and a response vector $\boldsymbol{y} \in \mathbb{R}^n$. For penalty functions $\rho(\cdot)$, consider penalized loss functions of the form

$$L_\lambda(\boldsymbol{\beta}) = \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2}{2n} + \lambda^2 \sum_{j=1}^{p} \rho(|\beta_j|/\lambda) \tag{2.1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a vector of regression coefficients. Let the penalty $\rho(t)$ be standardized to $\dot{\rho}(0+) = 1$, where $\dot{\rho}(t) = (d/dt)\rho(t)$. A vector $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)'$ is a

critical point of the penalized loss (2.1) if and only if

$$\begin{cases} \boldsymbol{x}_j'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n = \lambda \mathrm{sgn}(\widehat{\beta}_j)\dot{\rho}(|\widehat{\beta}_j|/\lambda), & \widehat{\beta}_j \neq 0, \\ \boldsymbol{x}_j'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n \in \lambda[-1, 1], & \widehat{\beta}_j = 0. \end{cases} \tag{2.2}$$

If the penalized loss (2.1) is convex in $\boldsymbol{\beta}$, then (3.4) is the Karush–Kuhn–Tucker condition for its minimization.

Given a penalty function $\rho(\cdot)$, one still has to choose a penalty level λ to arrive at a solution of (3.4). Such a choice may depend on the purpose of estimation, since variable selection may require a larger λ than does prediction. However, scale-invariance considerations and theoretical results suggest a penalty level proportional to the noise level σ. This motivates a scaled penalized least squares estimator as a numerical equilibrium in the following iterative algorithm:

$$\begin{aligned} &\widehat{\sigma} \leftarrow \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{\mathrm{old}}\|_2/\{(1-a)n\}^{1/2}, \\ &\lambda \leftarrow \widehat{\sigma}\lambda_0, \\ &\widehat{\boldsymbol{\beta}} \leftarrow \widehat{\boldsymbol{\beta}}^{\mathrm{new}},\ L_\lambda(\widehat{\beta}^{\mathrm{new}}) \lessapprox L_\lambda(\widehat{\beta}^{\mathrm{old}}), \end{aligned} \tag{2.3}$$

where λ_0 is a prefixed penalty level, not depending on σ, $\widehat{\sigma}$ estimates the noise level, and $a \geq 0$ provides an option for a degrees-of-freedom adjustment with $a > 0$. For $p < n$ and $(a, \lambda_0) = (p/n, 0)$, (2.3) initialized with the least squares estimator $\widehat{\boldsymbol{\beta}}^{(\mathrm{lse})}$ is non-iterative and gives $\widehat{\sigma}^2 = \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(\mathrm{lse})}\|_2^2/(n-p)$. For large data sets, one may use a few passes of a gradient descent algorithm to compute $\widehat{\beta}^{\mathrm{new}}$ from $\widehat{\beta}^{\mathrm{old}}$. In the numerical experiments reported in Section 4, $\widehat{\beta}^{\mathrm{new}}$ is a solution of (3.4) for the given λ. We describe this implementation in the following two paragraphs.

The first step of our implementation is the computation of a solution path $\widehat{\beta}(\lambda)$ of (3.4) beginning from $\widehat{\beta}(\lambda) = 0$ for $\lambda = |X'y/n|_\infty$. For quadratic spline penalties

$\rho(t)$ with m knots, Zhang (2010) developed an algorithm to compute a linear spline path of solutions $\{\lambda^{(t)} \oplus \widehat{\boldsymbol{\beta}}^{(t)} : t \geq 0\}$ of (3.4) to cover the entire range of λ. This extends the least angle regression solution or Lasso path (Osborne et al., 2000a,b; Efron et al., 2004) from $m = 1$ and includes the minimax concave penalty for $m = 2$ and the smoothly clipped absolute deviation penalty (Fan & Li, 2001) for $m = 3$. An R package named plus is available for computing the solution paths for these penalties.

The second step of our implementation is the iteration (2.3) along the solution path $\beta(\lambda)$ computed in the first step. That is to use the already computed

$$\widehat{\beta}^{\text{new}} = \widehat{\boldsymbol{\beta}}(\lambda) \tag{2.4}$$

in (2.3). For the scaled Lasso, we use $a = 0$ in (2.3) and $\rho(t) = t$ in (2.1) and (3.4). For the scaled minimax concave penalized selection, we use $a = 0$ and the minimax concave penalty $\rho(t) = \int_0^t (1 - x/\gamma)_+ dx$, where $\gamma > 0$ regularizes the maximum concavity of the penalty. When $\gamma = \infty$, it becomes the scaled Lasso. The algorithm (2.3) can be easily implemented once a solution path is computed.

Consider the ℓ_1 penalty. As discussed in the introduction, (2.3) and (3.5) form an alternating minimization algorithm for the penalized joint loss function

$$L_{\lambda_0}(\boldsymbol{\beta}, \sigma) = \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2}{2n\sigma} + \frac{(1-a)\sigma}{2} + \lambda_0 \|\boldsymbol{\beta}\|_1. \tag{2.5}$$

Antoniadis (2010) suggested this jointly convex loss function as a way of extending Huber's robust regression method to high dimensions. For $a = 0$ and $\lambda = \widehat{\sigma}\lambda_0$ with fixed $\widehat{\sigma}$, $\widehat{\sigma} L_{\lambda_0}(\boldsymbol{\beta}, \widehat{\sigma}) = L_\lambda(\boldsymbol{\beta}) + \widehat{\sigma}^2/2$, so that $\widehat{\boldsymbol{\beta}} \leftarrow \widehat{\boldsymbol{\beta}}(\lambda)$ in (3.5) minimizes $L_{\lambda_0}(\boldsymbol{\beta}, \widehat{\sigma})$ over $\boldsymbol{\beta}$. For fixed $\widehat{\boldsymbol{\beta}}$, $\widehat{\sigma}^2 \leftarrow \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2/\{(1-a)n\}$ in (2.3) minimizes $L_{\lambda_0}(\widehat{\boldsymbol{\beta}}, \sigma)$ over σ. We summarize some properties of the algorithm (2.3) with (3.5) in the following proposition.

Proposition 2.1. *Let $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\lambda)$ be a solution path of (3.4) with $\rho(t) = t$. The penalized loss function (2.5) is jointly convex in $(\boldsymbol{\beta}, \sigma)$ and the algorithm (2.3) with (3.5) converges to*

$$(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}) = \arg\min_{\boldsymbol{\beta},\sigma} L_{\lambda_0}(\boldsymbol{\beta}, \sigma). \tag{2.6}$$

The resulting estimators $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{y})$ and $\widehat{\sigma} = \widehat{\sigma}(\boldsymbol{X}, \boldsymbol{y})$ are scale equivariant in $\boldsymbol{y}$ in the sense that $\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, c\boldsymbol{y}) = c\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{y})$ and $\widehat{\sigma}(\boldsymbol{X}, c\boldsymbol{y}) = |c|\widehat{\sigma}(\boldsymbol{X}, \boldsymbol{y})$. Moreover,

$$\frac{\partial}{\partial\sigma} L_{\lambda_0}\{\widehat{\boldsymbol{\beta}}(\sigma\lambda_0), \sigma\} = \frac{1-a}{2} - \frac{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\sigma\lambda_0)\|_2^2}{2n\sigma^2}. \tag{2.7}$$

Since (2.5) is not strictly convex, the joint estimator may not be unique for some give data point $(\boldsymbol{X}, \boldsymbol{y})$. However, since (2.5) is strictly convex in σ, $\widehat{\sigma}$ is always unique in (2.6) and the uniqueness of $\widehat{\beta}$ follows from that of Lasso $\widehat{\beta}(\lambda)$ at $\lambda = \widehat{\sigma}\lambda_0$. The Lasso estimator $\widehat{\boldsymbol{\beta}}(\lambda)$ is unique when the second part of (3.4) is strict in the sense of not hitting $\pm\lambda$ when $\widehat{\beta}_j = 0$, which holds almost everywhere in $(\boldsymbol{X}, \boldsymbol{y})$ for $\lambda > 0$. See, for example, Zhang (2010).

2.3 Theoretical results

In this section, we study theoretical properties of the scaled Lasso (2.6) with $a = 0$. Let $\boldsymbol{\beta}^*$ be a vector of true regression coefficients. An expert with oracular knowledge of $\boldsymbol{\beta}^*$ would estimate the noise level by the oracle maximum likelihood estimator

$$\sigma^* = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2 / n^{1/2}. \tag{2.8}$$

Under the Gaussian assumption, this is the maximum likelihood estimator for σ when β^* is known and $n(\sigma^*/\sigma)^2$ follows the χ^2_n distribution. Due to the scale equivariance of $\widehat{\sigma}$ in Proposition 1, it is natural to use σ^* as an estimation target with or without the Gaussian assumption. We derive upper and lower bounds for $\widehat{\sigma}/\sigma^* - 1$ and use them to prove the consistency and asymptotic normality of $\widehat{\sigma}$. We derive oracle inequalities for the prediction performance and the estimation of $\boldsymbol{\beta}$ under the ℓ_q loss. Throughout the sequel, $\mathrm{pr}_{\boldsymbol{\beta},\sigma}$ is the probability measure under which $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \sim N(0, \sigma^2 \boldsymbol{I}_n)$. We assume $\|\boldsymbol{x}_j\|_2^2 = n$ whenever $\mathrm{pr}_{\boldsymbol{\beta},\sigma}$ is invoked. The asymptotic theory here concerns $n \to \infty$ and allows all parameters and variables to depend on n, including $p \geq n \geq \|\boldsymbol{\beta}\|_0 \to \infty$.

2.3.1 Consistency of noise level estimator

We first provide the consistency for the estimation of σ via an oracle inequality for the prediction error of the scaled Lasso. In our first theorem, the relative error for the estimation of σ is bounded by a quantity τ_0 related to a prediction error bound $\eta(\lambda, \xi, w, T)$ in (2.9) below. For $\lambda > 0$, $\xi > 1$, $\boldsymbol{w} \in \mathbb{R}^p$, and $T \subset \{1, \ldots, p\}$, define $\delta_{w,T} = 1 - I(w = \beta^*, T = \emptyset)$ and

$$\eta(\lambda, \xi, \boldsymbol{w}, T) = \|\boldsymbol{X}\boldsymbol{\beta}^* - \boldsymbol{X}\boldsymbol{w}\|_2^2/n + (1 + \delta_{w,T})2\lambda\|\boldsymbol{w}_{T^c}\|_1 + \frac{4\xi^2\lambda^2|T|}{(\xi+1)^2\kappa^2(\xi, T)} \quad (2.9)$$

where $\kappa(\xi, T)$, the compatibility factor (van de Geer & Bühlmann, 2009), is defined as

$$\kappa(\xi, T) = \min\left\{ \frac{|T|^{1/2}\|\boldsymbol{X}\boldsymbol{u}\|_2}{n^{1/2}\|\boldsymbol{u}_T\|_1} : \boldsymbol{u} \in \mathscr{C}(\xi, T),\ \boldsymbol{u} \neq 0 \right\} \quad (2.10)$$

with the cone $\mathscr{C}(\xi, T) = \{\boldsymbol{u} : \|\boldsymbol{u}_{T^c}\|_1 \leq \xi\|\boldsymbol{u}_T\|_1\}$. Since the prediction error bound $\eta(\lambda, \xi, w, T)$ is valid for all w and T, τ_0 is related to its minimum over all w and T at

the oracle scale σ^*:

$$\tau_0 = \eta_*^{1/2}(\sigma^*\lambda_0, \xi)/\sigma^*, \quad \eta_*(\lambda, \xi) = \inf_{\boldsymbol{w}, T} \eta(\lambda, \xi, \boldsymbol{w}, T). \tag{2.11}$$

Theorem 2.1. *Let $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma})$ be as in (2.6) with $a = 0$, $\boldsymbol{\beta}^* \in \mathrm{R}^p$, σ^* in (2.8), $z^* = \|\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*)/n\|_\infty/\sigma^*$ and $\xi > 1$. When $z^* \le (1-\tau_0)\lambda_0(\xi-1)/(\xi+1)$,*

$$\max\Big(1 - \frac{\widehat{\sigma}}{\sigma^*}, 1 - \frac{\sigma^*}{\widehat{\sigma}}\Big) \le \tau_0, \quad \frac{\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2}{n^{1/2}\sigma^*} \le \frac{1}{\sigma^*}\eta_*^{1/2}\Big(\frac{\sigma^*\lambda_0}{1-\tau_0}, \xi\Big) \le \frac{\tau_0}{1-\tau_0} \tag{2.12}$$

In particular, if $\lambda_0 = A\{(2/n)\log p\}^{1/2}$ with $A > (\xi+1)/(\xi-1)$ and $\eta_*(\sigma\lambda_0, \xi)/\sigma \to 0$, then

$$\mathrm{pr}_{\boldsymbol{\beta}^*, \sigma}\big(|\widehat{\sigma}/\sigma - 1| > \epsilon\big) \to 0 \tag{2.13}$$

for all $\epsilon > 0$.

Theorem 2.1 extends to the scaled Lasso a unification of prediction oracle inequalities for a fixed penalty. With $\lambda = \sigma^*\lambda_0/(1-\tau_0)_+$, (2.12) gives $\max\{(\sigma^*\tau_0)^2, \|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2/n\} \le \eta_*(\lambda, \xi)$, or

$$\begin{aligned} & \max\{(\sigma^*\tau_0)^2, \|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2/n\} \\ \le \; & \min_w \Big\{\|\boldsymbol{X}w - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2/n + 4\widetilde{C}\lambda\sum_{j=1}^p \min(\lambda, |w_j|)\Big\} \end{aligned} \tag{2.14}$$

for a $\widetilde{C} \ge 1$, if the minimum in (2.14) is attained at a $\widetilde{w}$ with $(1+1/\xi)^2\kappa^2(\xi, \widetilde{T}) \ge 1/\widetilde{C}$, where $\widetilde{T} = \{j : |\widetilde{w}_j| > \lambda\}$. This asserts that for an arbitrary, possibly non-sparse β^*, the prediction error of the scaled Lasso is no greater than that of the best linear predictor

Xw with a sparse w for an additional capped-ℓ_1 cost of the order $\lambda \sum_j \min(\lambda, |w_j|)$. A consequence of this prediction error bound for the scaled Lasso is the consistency of the corresponding estimator of the noise level in (2.13). Due to the scale equivariance in Proposition 2.1, Theorem 2.1 and the results in the rest of the section are all scale free.

For fixed penalty λ, the upper bound $\eta(\lambda, \xi, \boldsymbol{w}, T)$ has been previously established for different w and T, with possibly different constant factors. Examples include $\eta(\lambda, \xi, \boldsymbol{\beta}^*, \emptyset) = 2\lambda\|\boldsymbol{\beta}^*\|_1$ (Greenshtein & Ritov, 2004; Greenshtein, 2006), $\eta(\lambda, \xi, \boldsymbol{\beta}^*, S_{\boldsymbol{\beta}^*}) \lesssim \lambda^2\|\beta^*\|_0$ with $S_{\boldsymbol{w}} = \{j : w_j \neq 0\}$ (van de Geer & Bühlmann, 2009), and $\min_{\boldsymbol{w}} \eta(\lambda, \xi, \boldsymbol{w}, S_{\boldsymbol{w}}) = \min_w\{\|\boldsymbol{X}\boldsymbol{\beta}^* - \boldsymbol{X}\boldsymbol{w}\|_2^2/n + O(\lambda^2\|\boldsymbol{w}\|_0)\}$ (Koltchinskii et al., 2011).

2.3.2 Asymptotic normality of noise level estimator

Now we provide sharper convergence rates and the asymptotic normality for the scaled Lasso estimation of the noise level σ. This sharper rate $\lambda\mu(\lambda, \xi)/\sigma^2$, essentially taking the square of the order τ_0 in (2.12), is based on the following ℓ_1 error bound for the estimation of $\boldsymbol{\beta}$,

$$\mu(\lambda, \xi) = (\xi + 1) \min_T \inf_{0<\nu<1} \max\Big[\frac{\|\boldsymbol{\beta}^*_{T^c}\|_1}{\nu}, \frac{\lambda|T|/\{2(1-\nu)\}}{\kappa^2\{(\xi+\nu)/(1-\nu), T\}}\Big]. \tag{2.15}$$

This ℓ_1 error bound has the interpretation

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \mu(\lambda, \xi) \leq \widetilde{C} \sum_{j=1}^{p} \min(\lambda, |\beta_j^*|), \tag{2.16}$$

if $\widetilde{C} \geq (1 + \xi) \max\{2, 1/\kappa^2(2\xi + 1, \widetilde{T})\}$ with $\widetilde{T} = \{j : |\beta_j^*| > \lambda\}$. This allows $\boldsymbol{\beta}^*$ to have many small elements, as in Zhang & Huang (2008), Zhang (2009) and Ye &

Zhang (2010). The bound $\mu(\lambda,\xi) \le (\xi+1)\lambda|S_{\beta^*}|/\{2\kappa^2(\xi,S_{\beta^*})\}$ improves upon its earlier version in van de Geer & Bühlmann (2009) by a constant factor $4\xi/(\xi+1) \in (2,4)$.

Theorem 2.2. *Let* $\{\widehat{\boldsymbol{\beta}},\widehat{\sigma},\boldsymbol{\beta}^*,\sigma^*,z^*,\xi\}$ *be as in Theorem 2.1. Set* $\tau_* = \{\lambda_0\mu(\sigma^*\lambda_0,\xi)/\sigma^*\}^{1/2}$. *(i) The following inequalities hold in the event* $z^* \le (1-\tau_*^2)\lambda_0(\xi-1)/(\xi+1)$*,*

$$\max\big(1-\widehat{\sigma}/\sigma^*, 1-\sigma^*/\widehat{\sigma}\big) \le \tau_*^2, \quad \|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1 \le \mu(\sigma^*\lambda_0,\xi)/(1-\tau_*^2). \tag{2.17}$$

(ii) Let $\lambda_0 \ge \{(2/n)\log(p/\epsilon)\}^{1/2}(\xi+1)/\{(\xi-1)(1-\tau_*^2)\}$. *For all* $\epsilon > 0$ *and* $n-2 > \log(p/\epsilon) \to \infty$*,*

$$\mathrm{pr}_{\boldsymbol{\beta}^*,\sigma}\big\{z^* \le (1-\tau_*^2)\lambda_0(\xi-1)/(\xi+1)\big\} \ge 1-\{1+o(1)\}\epsilon/\{\pi\log(p/\epsilon)\}^{1/2}.$$

If $\lambda_0 = A\{(2/n)\log p\}^{1/2}$ *with* $A > (\xi+1)/(\xi-1)$ *and* $\lambda_0\mu(\sigma\lambda_0,\xi)/\sigma \ll n^{-1/2}$*, then*

$$n^{1/2}\big(\widehat{\sigma}/\sigma-1\big) \to N(0,1/2) \tag{2.18}$$

in distribution under $\mathrm{pr}_{\boldsymbol{\beta}^*,\sigma}$.

Since $\sigma^2\tau_*^2 \approx \mu(\lambda,\xi) \le 2(\xi+1)\min_T \eta(\lambda,2\xi+1,\boldsymbol{\beta}^*,T)$ with $\lambda=\sigma\lambda_0$, the rate τ_*^2 in (2.17) is essentially the square of that in (2.12), in view of (2.11). It follows that the scaled Lasso provides a faster convergence rate than does the penalized maximum likelihood estimator for the estimation of the noise level (Städler et al., 2010; Sun &

Zhang, 2010). In particular, (2.17) implies

$$\max\left(1-\widehat{\sigma}/\sigma^*, 1-\sigma^*/\widehat{\sigma}\right) \le (\xi+1)\lambda_0^2|S_{\boldsymbol{\beta}^*}|/\{2\kappa^2(\xi, S_{\boldsymbol{\beta}^*})\} \lesssim |S_{\boldsymbol{\beta}^*}|(\log p)/n \quad (2.19)$$

with $S_{\boldsymbol{\beta}^*} = \{j : \beta_j^* \neq 0\}$, when $\kappa^2(\xi, S_{\boldsymbol{\beta}^*})$ can be treated as a constant. The bounds (2.19) and its general version (2.17) lead to the asymptotic normality (2.18) under proper assumptions. Thus, statistical inference about σ is justified with the scaled Lasso in certain large-p-smaller-n cases, for example, when $|S_{\boldsymbol{\beta}^*}|(\log p)/\sqrt{n} \to 0$ under the compatibility condition (van de Geer & Bühlmann, 2009).

2.3.3 Oracle inequalities for coefficient estimator

For a fixed penalty level, oracle inequalities for the ℓ_q error of the Lasso have been established in Bunea et al. (2007), van de Geer (2008) and van de Geer & Bühlmann (2009) for $q = 1$, Zhang & Huang (2008) and Bickel et al. (2009) for $q \in [1, 2]$, Meinshausen & Yu (2009) for $q = 2$, and Zhang (2009) and Ye & Zhang (2010) for $q \ge 1$. The bounds on $\widehat{\sigma}/\sigma^*$ in (2.17) and (2.19) allow automatic extensions of these existing ℓ_q oracle inequalities from the Lasso with fixed penalty to the scaled Lasso. We illustrate this by extending the oracle inequalities of Ye & Zhang (2010) for the Lasso and Candes & Tao (2007) for the Dantzig selector in the following corollary. Ye & Zhang (2010) used the following sign-restricted cone invertibility factor to separate conditions on the error $y - X\beta^*$ and design X in the derivation of error bounds for the Lasso:

$$F_q(\xi, S) = \inf\left\{\frac{|S|^{1/q}\|\boldsymbol{X}'\boldsymbol{X}\boldsymbol{u}\|_\infty}{n\|\boldsymbol{u}\|_q} : \boldsymbol{u} \in \mathscr{C}_-(\xi, S)\right\}, \quad (2.20)$$

where $\mathscr{C}_{-}(\xi, S) = \{\boldsymbol{u} : \|\boldsymbol{u}_{S^c}\|_1 \le \xi\|\boldsymbol{u}_S\|_1 \ne 0, u_j \boldsymbol{x}_j' \boldsymbol{X}\boldsymbol{u} \le 0, \text{for all } j \notin S\}$. The quantity (3.9) can be viewed as a generalized restricted eigenvalue comparing the ℓ_q loss and the dual norm of the ℓ_1 penalty with respect to the inner product for the least squares fit. This is more directly connected to the Karush–Kuhn–Tucker condition (3.4). Compared with the restricted eigenvalue (Bickel et al., 2009) and the compatibility factor (2.10), a main advantage of (3.9) is to allow all $q \in [1, \infty]$. In addition, (3.9) yields sharper oracle inequalities (Ye & Zhang, 2010). For $(|A|, |B|, \|\boldsymbol{u}\|_2) = (\lceil a \rceil, \lceil b \rceil, 1)$ with $A \cap B = \emptyset$, define

$$\delta_a^{\pm} = \max_{A,\boldsymbol{u}} \Big\{ \pm \Big(\|\boldsymbol{X}_A \boldsymbol{u}/n^{1/2}\|_2 - 1 \Big) \Big\}, \ \theta_{a,b} = \max_{A,B,\boldsymbol{u}} \big\| \boldsymbol{X}_A' \boldsymbol{X}_B \boldsymbol{u}/n \big\|_2. \tag{2.21}$$

The quantities in (2.21) are used in the uniform uncertainty principle (Candes & Tao, 2007) and the sparse Riesz condition (Zhang & Huang, 2008). We note that $1 - \delta_a^-$ is the minimum eigenvalue of $\boldsymbol{X}_A' \boldsymbol{X}_A/n$ among $|A| \le a$, $1 + \delta_a^+$ is the corresponding maximum eigenvalue, and $\theta_{a,b}$ is the maximum operator norm of size $a \times b$ off-diagonal sub-blocks of the Gram matrix $\boldsymbol{X}'\boldsymbol{X}/n$.

Corollary 2.1. *Suppose $\|\boldsymbol{\beta}^*_{S^c}\|_1 = 0$. Then, Theorem 2.2 holds with $\mu(\lambda, \xi)$ replaced by $\lambda|S|(2\xi)/\{(\xi+1)F_1(\xi, S)\}$, and for $z^* \le (1-\tau_*^2)\lambda_0(\xi-1)/(\xi+1)$,*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_q \le \frac{k^{1/q}(\sigma^* z^* + \widehat{\sigma}\lambda_0)}{F_q(\xi, S)} \le \frac{2\sigma^* \xi \lambda_0 k^{1/q}}{(1-\tau_*^2)(\xi+1)F_q(\xi, S)} \tag{2.22}$$

for all $1 \le q \le \infty$, where $k = |S|$. In particular, for $\xi = \sqrt{2}$ and $z^* \le (1-\tau_*^2)\lambda_0(\sqrt{2}-1)^2$,

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \le \frac{(8k)^{1/2}\lambda_0\sigma^*/(1-\tau_*^2)}{(\sqrt{2}+1)F_2(\sqrt{2}, S)} \le \frac{4k^{1/2}\lambda_0\sigma^*/(1-\tau_*^2)}{(\sqrt{2}+1)(1-\delta_{1.5k}^- - \theta_{2k,1.5k})_+}. \tag{2.23}$$

2.3.4 The key of proofs

The proofs of Theorem 2.1 and 2.2 are based on the basic inequality

$$\begin{aligned} & \|\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2/n + \|\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{X}\boldsymbol{w}\|_2^2/n \\ \leq \quad & \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2/n + 2\lambda\{\|\boldsymbol{w}\|_1 - \|\widehat{\boldsymbol{\beta}}(\lambda)\|_1\} + 2\sigma^* z^* \|\boldsymbol{w} - \widehat{\boldsymbol{\beta}}(\lambda)\|_1 \end{aligned} \quad (2.24)$$

as a consequence of the Karush–Kuhn–Tucker conditions (3.4). The version of (2.24) with $\boldsymbol{w} = \boldsymbol{\beta}^*$ is well-known (van de Geer & Bühlmann, 2009) and controls $\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2$ for sparse $\boldsymbol{\beta}^*$. When $\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{X}\boldsymbol{\beta}^*|_2^2 > \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}\boldsymbol{\beta}^*|_2^2$, (2.24) controls the excess for sparse w by the same argument. The general $\boldsymbol{w}$ is taken in Theorem 1, while $\boldsymbol{w} = \boldsymbol{\beta}^*$ is taken in Theorem 2. In both cases, (2.24) provides the cone condition in (2.10) and (3.9). This is used to derive upper and lower bounds for (2.7), the derivative of the profile loss function $L_{\lambda_0}(\widehat{\boldsymbol{\beta}}(\sigma\lambda_0), \sigma)$ with respect to σ, within a neighborhood of $\sigma/\sigma^* = 1$. The bounds for the minimizer $\widehat{\sigma}$ then follow from the joint convexity of the penalized loss (2.5).

2.4 Numerical studies

In this section, we present some numerical results to compare five methods: the scaled penalized methods with the ℓ_1 penalty, minimax concave penalty and smoothly clipped absolute deviation penalty, the ℓ_1 penalized maximum likelihood estimator (Städler et al., 2010), and its bias correction. The penalized maximum likelihood estimator is

$$\{\widehat{\boldsymbol{\beta}}^{(\mathrm{pmle})}, \widehat{\sigma}^{(\mathrm{pmle})}\} = \arg\max_{\boldsymbol{\beta},\sigma} \Big\{ -\frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2}{2\sigma^2 n} - \log\sigma - \lambda_0 \frac{\|\boldsymbol{\beta}\|_1}{\sigma} \Big\},$$

or equivalently the limit of the iteration $\widehat{\sigma} \leftarrow \{\boldsymbol{y}'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/n\}^{1/2}$ and $\widehat{\boldsymbol{\beta}} \leftarrow \widehat{\boldsymbol{\beta}}(\widehat{\sigma}\lambda_0)$. The bias-corrected estimator is one iteration of (2.3) with (3.5) from $\{\widehat{\boldsymbol{\beta}}^{(\mathrm{pmle})}, \widehat{\sigma}^{(\mathrm{pmle})}\}$

with $a = 0$,

$$\widehat{\sigma}^{(\mathrm{bc})} = \|\boldsymbol{y} - \widehat{\boldsymbol{\beta}}(\widehat{\sigma}^{(\mathrm{pmle})}\lambda_0)\|_2/n^{1/2},\ \widehat{\boldsymbol{\beta}}^{(\mathrm{bc})} = \widehat{\boldsymbol{\beta}}(\widehat{\sigma}^{(\mathrm{bc})}\lambda_0).$$

Two simulation examples and a real data set are considered.

2.4.1 Simulation results

Example 1

This experiment has the same setting as in Experiment 5 of Zhang (2010). We provide the description of the simulation settings in our notation as follows: $(n, p) = (600, 3000)$, the $\boldsymbol{x}_j$ are normalized columns from a Gaussian random matrix with independent and identically distributed rows and correlation $\sigma_{j,k} = \sigma_{1,2}^{|k-j|}$ between the j-th and k-th entries within each row, $\gamma = 2/(1 - \max|\boldsymbol{x}_k'\boldsymbol{x}_j|/n)$ for the minimax concave penalty and smoothly clipped absolute deviation penalty, the nonzero β_j^* are composed of five blocks of $\beta_*(1, 2, 3, 4, 3, 2, 1)'$ centered at random multiples $j_1, \ldots, j_5$ of 25, β_* sets $\|\boldsymbol{X}\boldsymbol{\beta}^*\|_2^2 = 3n$, and $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*$ is a vector of independent and identically distributed $N(0, 1)$ variables, where the true value of σ is 1. Two cases are considered: $\sigma_{1,2} = 0.1$ for low correlation and $\sigma_{1,2} = 0.9$ for high correlation.

We summarize the simulation results in Table 2.1, which provides the bias and standard error of the ratio $\widehat{\sigma}/\sigma$ and the average ℓ_2 error of the estimated β. In this simulation experiment, the scaled Lasso outperforms the penalized maximum likelihood estimator and its bias correction, which are also based on the Lasso path. The scaled concave penalized methods perform well at the universal penalty level $\lambda_2 = \{(2/n)\log p\}^{1/2}$. The simulation results also suggest that smaller λ may provide a somewhat better estimate of the noise level at the cost of a larger estimation error for the coefficients. The scaled minimax concave penalized selection demonstrates the

strongest performance for $\sigma_{1,2} = 0.1$, while the scaled Lasso is the best for $\sigma_{1,2} = 0.9$. This suggests potential improvements over the choice $\gamma = 2/(1 - \max|\boldsymbol{x}_k'\boldsymbol{x}_j|/n)$, since the minimax concave penalty becomes the ℓ_1 penalty with $\gamma = \infty$.

Table 2.1: Performance of five methods in Example 1 at three penalty levels λ_0, $\lambda_j = \{2^{j-1}(\log p)/n\}^{1/2}$, $j = 1, 2, 3$. The estimation performance across 100 replications in terms of bias ($\times 10$) and standard error ($\times 10$) of $\widehat{\sigma}/\sigma$ and the ℓ_2 estimation error ($\times 10$) of $\widehat{\boldsymbol{\beta}}$ are tabulated for each method.

		$\sigma_{1,2} = 0.1$		$\sigma_{1,2} = 0.9$	
Method		Bias $\pm$ SE ($\widehat{\sigma}/\sigma$)	$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$	Bias $\pm$ SE ($\widehat{\sigma}/\sigma$)	$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$
PMLE	λ_1	5·5 $\pm$ 0·4	8·7	2·5 $\pm$ 0·3	5·3
	λ_2	7·7 $\pm$ 0·4	12·1	3·8 $\pm$ 0·3	5·4
	λ_3	9·5 $\pm$ 0·4	15·0	5·7 $\pm$ 0·3	5·8
BC	λ_1	3·2 $\pm$ 0·4	7·8	0·3 $\pm$ 0·3	5·4
	λ_2	6·2 $\pm$ 0·5	11·4	1·3 $\pm$ 0·3	5·4
	λ_3	9·1 $\pm$ 0·5	14·9	3·2 $\pm$ 0·4	5·6
Scaled Lasso	λ_1	1·9 $\pm$ 0·4	7·3	0·1 $\pm$ 0·3	5·4
	λ_2	5·1 $\pm$ 0·5	10·9	0·7 $\pm$ 0·3	5·4
	λ_3	9·0 $\pm$ 0·5	14·9	1·9 $\pm$ 0·3	5·5
Scaled mcp	λ_1	$-$0·2 $\pm$ 0·4	4·7	0·1 $\pm$ 0·3	8·2
	λ_2	1·8 $\pm$ 0·6	7·4	0·7 $\pm$ 0·3	7·1
	λ_3	7·9 $\pm$ 0·8	14·0	1·8 $\pm$ 0·3	7·2
Scaled scad	λ_1	0·7 $\pm$ 0·4	6·0	0·1 $\pm$ 0·3	7·7
	λ_2	4·8 $\pm$ 0·6	11·0	0·7 $\pm$ 0·3	6·7
	λ_3	8·9 $\pm$ 0·5	14·9	1·9 $\pm$ 0·3	5·8

PMLE, ℓ_1 penalized maximum likelihood estimator; BC, bias-corrected estimator; mcp, minimax concave penalty; scad, smoothly clipped absolute deviation penalty; SE, standard erro

We plot the histogram of the simulated $\widehat{\sigma}$ for the scaled estimators with the penalty level λ_1 in Figure 2.1, with the approximate normal density superimposed. Since $F(\xi, S) > 1$ and $|\beta^*|_0(\log p)/n > 0.4$, the condition for the asymptotic normality in Theorem 2.2 holds marginally at the best with an approximate variance $1200^{-1/2} = 0.03$. Still, the plots demonstrate a reasonable match between the simulations and the theory, except for the scaled Lasso and scaled smoothly clipped absolute deviation penalization in the case of low correlation. For λ_2 and λ_3, the bias of the scaled estimators dominates the standard error; see Table 2.1.

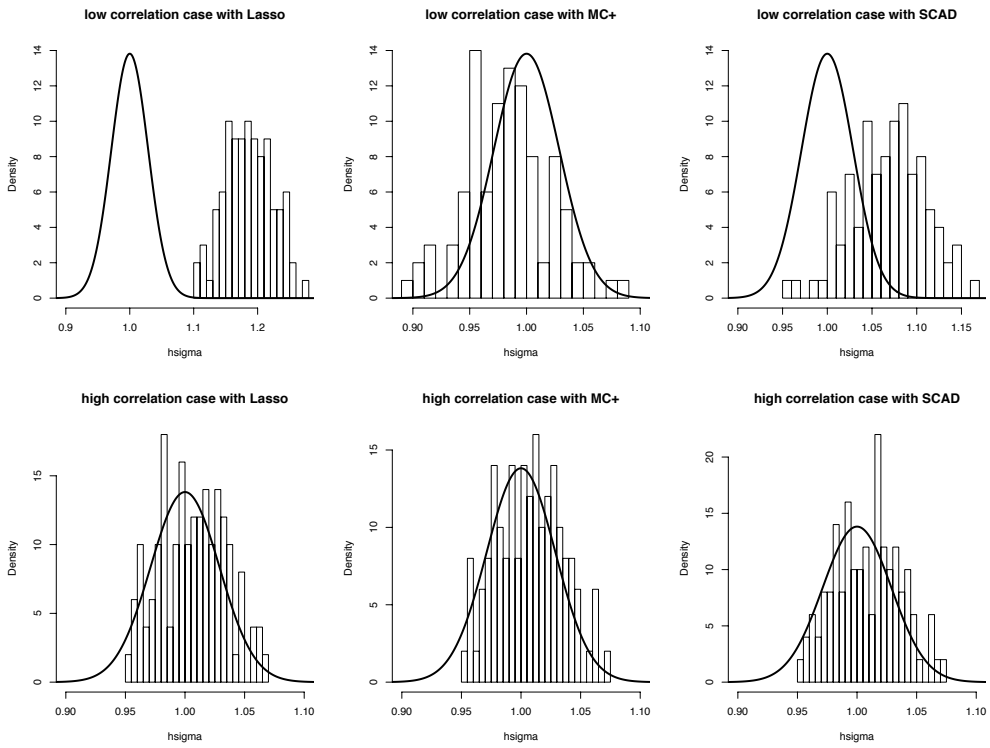

Figure 2.1: Histograms of the simulated $\widehat{\sigma}$ for the scaled estimators with penalty level λ_1. Top row: $\sigma_{1,2} = 0.1$; Bottom row: $\sigma_{1,2} = 0.9$; Left: the scaled Lasso; Middle: the scaled minimax concave penalization; Right: the scaled penalization with smoothly clipped absolute deviation.

Example 2

We compare the five estimators at the same three penalty levels $\{\lambda_1, \lambda_2, \lambda_3\}$ as in Example 1. The experiment has the setting of Example 2 in Fan et al. (2012), with the smallest signal, $b = 1/\sqrt{3}$. Fan et al. (2012) considered several joint estimators of $(\boldsymbol{\beta}, \sigma)$ using cross-validation. Their results are included without repeating their experiment. We provide their description of the simulation setting in our notation as follows: $\boldsymbol{X}$ has independent and identically distributed Gaussian rows with marginal distribution $N(0,1)$, $\mathrm{corr}(x_i, x_j) = \sigma_{1,2}$ for $1 \le i < j \le 50$ and $\mathrm{corr}(x_i, x_j) = 0$ otherwise, $(n, p) = (200, 2000)$, nonzero coefficients $\beta_j = 1/\sqrt{3}$ for $j \in S = \{1, 2, 3\}$, and $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \sim N(0, \sigma^2 \boldsymbol{I})$ with $\sigma = 1$. Two configurations are considered: independent columns $\boldsymbol{x}_j$ with $\sigma_{1,2} = 0$ and correlated first 50 columns $\boldsymbol{x}_j$ with $\sigma_{1,2} = 0.5$. Again, we set $\gamma = 2/(1 - \max |\boldsymbol{x}_k' \boldsymbol{x}_j|/n)$ for the concave penalties.

Our simulation results are presented in the top section of Table 2.2, while the results in Fan et al. (2012) are included in the bottom section. In addition to the bias and the standard error for the estimation of the noise level σ, we report the average model size $|\widehat{S}|$ and the relative frequency of sure screening $\widehat{S} \supset S$ as in Fan et al. (2012), where $\widehat{S} = \{j : \widehat{\beta}_j \neq 0\}$ is the selected model.

The scaled Lasso at λ_1 and the scaled minimax concave penalized selection at λ_2 clearly outperform the cross-validation methods for the estimation of σ, especially for the standard error. The results with the average model size and sure screening probability show that cross-validation methods select about 30 variables when the true model size is 3. The scaled estimators choose correct models at the universal penalty level $\{(2/n) \log p\}^{1/2}$ with large probabilities.

Table 2.2: Performance of five methods in Example 2 at three penalty levels λ_0, $\lambda_j = \{2^{j-1}(\log p)/n\}^{1/2}$, $j = 1, 2, 3$, and the results in Fan et al. (2012). The estimation performance across 100 replications in terms of average bias ($\times 10$) and standard error ($\times 10$) of $\widehat{\sigma}/\sigma$, average model size and relative frequency of sure screening are tabulated for each method.

		$\sigma_{1,2} = 0$			$\sigma_{1,2} = 0.5$		
Method		Bias $\pm$ SE ($\widehat{\sigma}/\sigma$)	AMS	SSP	Bias $\pm$ SE ($\widehat{\sigma}/\sigma$)	AMS	SSP
PMLE	λ_1	1·7 $\pm$ 0·6	7·8	1·0	1·5 $\pm$ 0·5	9·8	1·0
	λ_2	2·6 $\pm$ 0·6	3·1	1·0	2·5 $\pm$ 0·5	5·2	1·0
	λ_3	3·7 $\pm$ 0·7	2·0	0·3	3·8 $\pm$ 0·5	3·8	1·0
BC	λ_1	0·5 $\pm$ 0·6	12·3	1·0	0 $\pm$ 0·5	15·7	1·0
	λ_2	1·6 $\pm$ 0·6	3·2	1·0	0·7 $\pm$ 0·5	5·8	1·0
	λ_3	3·4 $\pm$ 0·7	2·1	0·4	2·0 $\pm$ 0·6	4·3	1·0
Scaled Lasso	λ_1	$-$0·1 $\pm$ 0·6	15·0	1·0	$-$0·5$\pm$ 0·6	18·8	1·0
	λ_2	1·3 $\pm$ 0·7	3·2	1·0	0·4 $\pm$ 0·6	6·1	1·0
	λ_3	3·2 $\pm$ 0·7	2·1	0·4	1·3 $\pm$ 0·6	4·5	1·0
Scaled mcp	λ_1	$-$1·3 $\pm$ 0·8	14·3	1·0	$-$0·8$\pm$ 0·6	14·0	1·0
	λ_2	$-$0·1 $\pm$ 0·6	3·2	1·0	0·1 $\pm$ 0·6	3·3	1·0
	λ_3	1·5 $\pm$ 1·4	2·5	0·6	0·7 $\pm$ 0·6	3·0	1·0
Scaled scad	λ_1	$-$0·6 $\pm$ 0·6	14·1	1·0	$-$0·5$\pm$ 0·6	14·4	1·0
	λ_2	0·8 $\pm$ 0·9	3·1	1·0	0·3 $\pm$ 0·6	3·9	1·0
	λ_3	3·2 $\pm$ 0·7	2·2	0·4	1·2 $\pm$ 0·6	3·9	1·0
N-LASSO		$-$5·3 $\pm$ 2·0	36·6	1·0	$-$4·6 $\pm$ 2·0	29·6	1·0
RCV-SIS		0·2 $\pm$ 1·4	50·0	0·9	$-$0·1 $\pm$ 1·4	50·0	1·0
RCV-ISIS		0·5 $\pm$ 1·7	30·9	0·7	0·2 $\pm$ 1·2	29·0	0·8
RCV-LASSO		0 $\pm$ 1·3	31·1	0·9	$-$0·3 $\pm$ 1·1	26·5	1·0
P-SCAD		$-$1·4 $\pm$ 1·1	30·0	1·0	$-$1·2 $\pm$ 1·7	29·9	1·0
CV-SCAD		0·7 $\pm$ 1·2	30·0	1·0	0·9 $\pm$ 1·3	29·9	1·0
P-LASSO		$-$0·8 $\pm$ 2·1	36·5	1·0	$-$0·9 $\pm$ 1·5	29·6	1·0
CV-LASSO		1·4 $\pm$ 1·1	36·5	1·0	0·8 $\pm$ 1·0	29·6	1·0

PMLE, ℓ_1 penalized maximum likelihood estimator; BC, bias-corrected estimator; mcp, minimax concave penalty; scad, smoothly clipped absolute deviation penalty; N, naive; RCV, refitted cross-validation; SIS, sure independent screening; ISIS, iterative SIS; P, plug-in method with degrees-of-freedom correction; CV, cross-validation; SE, standard error; AMS, average model size; SSP, relative frequency of sure screening.

Table 2.3: Selected probe sets by four methods in the real data example: the Lasso with cross-validation, the Lasso with adjusted cross-validation, the scaled Lasso and minimax concave penalized selection at $\lambda_2 = \{2(\log p)/n\}^{1/2}$. The estimated coefficients ($\times 10^3$) are tabulated for each method.

Probe ID	C-V lasso		C-V lasso/LSE		Scaled lasso		Scaled MCP	
#cov	200	3000	200	3000	200	3000	200	3000
1369353_at	$-9 \cdot 12$	$-7{\cdot}13$*	$-7{\cdot}09$	$-2{\cdot}79$*	$-7{\cdot}3$	$-4{\cdot}03$*		
1370052_at$^{\triangle}$		3·65						
1370429_at	$-3{\cdot}22$		$-8{\cdot}94$*	$-11{\cdot}06$	$-8{\cdot}78$*	$-9{\cdot}36$	$-16{\cdot}37$*	
1371242_at	$-6{\cdot}66$							
1374106_at	8·88*	10·58*	7·33*	6·14*	7·45*	7·01*	8·47*	10·02*
1374131_at	4·07	0·80						
1375585_at$^{\triangle}$						0·58		
1384204_at			0·70		0·70			
1387060_at$^{\triangle}$		3·50*						
1388538_at$^{\triangle}$		1·42						
1389584_at	17·16*	25·39*	20·07*	19·61*	19·97*	21·18*	45·75*	50·49*
1393979_at	$-1{\cdot}81$		$-0{\cdot}22$		$-0{\cdot}4$			
1379079_at$^{\triangle}$		$-1{\cdot}43$*						
1379495_at		4·84	1·73		1·71	$1 \cdot 00$		
1379971_at	13·56*	13·1	11·19*	8·81	11·25*	9·52		
1380033_at	8·69		2·76		2·97		6·75*	
1380070_at$^{\triangle}$		0·19						
1381787_at	$-2{\cdot}05$		$-2{\cdot}01$		$-2{\cdot}11$			
1382452_at$^{\triangle}$		12·93				1·63		12·91*
1382835_at	12·64	5·79	3·73		4·15			
1383110_at	9·03*	19·99	15·10*	16·43	14·97*	16·69	15·80*	23·01*
1383522_at	3·03*		*		*			
1383673_at	5·54	6·12*	6·07	6·15*	6·08	6·47*		
1383749_at	$-13{\cdot}86$	$-10{\cdot}85$*	$-10{\cdot}84$	$-6{\cdot}7$*	$-11{\cdot}02$	$-8{\cdot}07$*	$-2{\cdot}74$ *	$-1{\cdot}11$*
1383996_at	25·01*	17·82*	18·61*	14·30*	18·88*	15·52*	25·07*	19·19*
1385687_at$^{\triangle}$		$-0{\cdot}99$						
1386683_at				4·60*		2·90*		
1390788_a_at	0·92							
1392692_at$^{\triangle}$		1·74						
1393382_at	2·43							
1393684_at	1·59							
1395076_at$^{\triangle}$						0·23		
1397489_at$^{\triangle}$		3·33						
Model size	19	20	15	10	15	14	7	6
$\widehat{\lambda} = \widehat{\sigma}\lambda_0$	0.0103	0.0163	0.025	0.035	0.0243	0.0315	0.0244	0.0304

C-V Lasso/LSE, the Lasso estimator with the adjusted cross-validation; mcp, minimax concave penalty; #cov, the number of covariates considered; $\triangle$, probes not in the smaller set of 200 probes; * , covariates selected by stability selection.

2.4.2 Real data example

We study a data set containing 18976 probes for 120 rats, which is reported in Scheetz et al. (2006). Our goal is to find probes that are related to that of gene TRIM32, which has been found to cause Bardet–Biedl syndrome, a genetically heterogeneous disease of multiple organ systems including the retina. We consider linear regression with the probe from TRIM32, 1389163_at, as the response variable. As in Huang et al. (2008), we focus on 3000 probes with the largest variances among the 18975 covariates and consider two approaches. The first approach is to regress on these $p = 3000$ probes. The second approach is to regress on the 200 probes among the 3000 with the largest marginal correlation coefficients with TRIM32. For the cross-validation Lasso, we randomly partition the data 1000 times, each with a training set of size 80 and a validation set of size 40. For each partition, the penalty level λ is selected by minimizing the prediction mean squared error in the validation set. Then we compute the Lasso estimator with all 120 observations at the penalty level equal to the median of the selected penalty levels with the 1000 random partitions. Since cross-validation tends to choose a larger model, we also consider an adjusted version using the cross-validated error of the least squares estimator with covariates selected by the Lasso. For the minimax concave penalty, we set $\gamma = 2/(1 - \sigma_{0.95}) = 6.37$, where $\sigma_{0.95}$ is the 95% quantile of $|\boldsymbol{x}_k'\boldsymbol{x}_j|/n$.

Table 2.3 shows the probe sets identified by four methods: the cross-validation lasso, its adjusted version, the scaled lasso at at universal penalty level $\lambda_2 = \{2(\log p)/n\}^{1/2}$, and the minimax concave penalized selection at the same penalty level. We apply stability selection (Meinshausen & Bühlmann, 2010) to check the reliability of selection. Let $W_1, \ldots, W_p$ be independent variables with $P(W = 0.2) =$

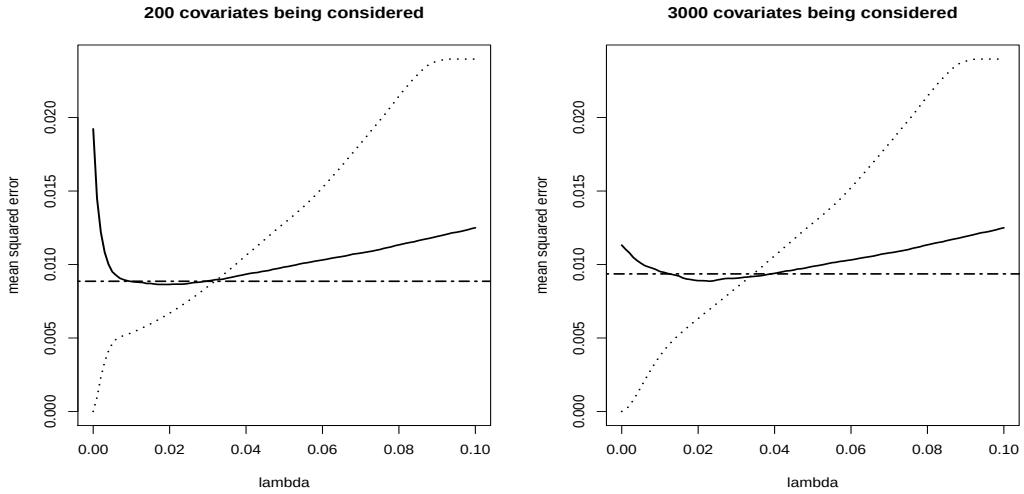

Figure 2.2: Mean squared error of the Lasso estimator against penalty level λ. Solid line: testing error for fixed λ; dotted line: training error for fixed λ; dashed line: testing error for the scaled Lasso.

$P(W = 1) = 1/2$ and

$$\widehat{\boldsymbol{\beta}}^W = \arg\min_{\boldsymbol{b}} \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2}{2n} + \widehat{\lambda} \sum_{j=1}^{p} |b_j|/W_j,$$

where $\widehat{\lambda}$ is the penalty level chosen by individual methods. The stability selection selects variables with nonzero estimated $\widehat{\beta}_j^W$ over 50 times in 100 replications. We observe that the scaled minimax concave penalized selector produces most sparse and most stable selection, followed by the adjusted cross-validation, the scaled lasso and then the plain cross-validation. The selection results are consistent among the four methods in the sense that the selected models are almost nested. Since the model size is between 6 and 8 by stability selection in all 8 cases and by the scaled minimax concave penalized selection for both $p = 200$ and $p = 3000$, these two methods provide most consistent results. The scaled lasso and the adjusted cross-validation yield identical lasso and stability selections for $p = 200$ and identical stability selection for $p = 3000$.

We also compare the prediction performance of the scaled Lasso with that of the Lasso with the best fixed penalty level. We compute the scaled estimators in 1000 replications. In each replication, the dataset is split at random into a training set with 80 observations and a test set with 40 observations. The prediction mean squared error is computed within the test set, while the scaled estimators and the Lasso estimator with fixed penalty level λ are computed based on the training set. Figure 2.2 demonstrates that in prediction, the scaled Lasso with λ_0 chosen as $\lambda_2 = \{2(\log p)/n\}^{1/2}$ performs almost as well as the Lasso with the optimal fixed λ.

In addition, we compare the prediction performance of all the estimators mentioned in this section. In each replication, we compute the penalized maximum likelihood estimator, its bias-correction, and scaled penalization methods based on the training set of 80 observations. For cross-validation, the training set of 80 observations is further partitioned at random 100 times into two groups of sizes 60 and 20, and a penalty level is selected by minimizing the estimated loss in the smaller group for the Lasso estimator based on the larger group. This selected penalty is then used for the Lasso with the entire training set. Thus, the cross-validation Lasso is also based on the training set with 80 observations. For the adjusted cross-validation with the least squares cross-validated error, two estimators are considered: the Lasso estimator with the λ selected by the adjusted cross-validation, and the least squares estimator with the covariates selected by the Lasso. In Table 2.4, we present the medians of the prediction mean squared error and the selected model size in the 200 replications. The scaled Lasso has comparable prediction performance as cross-validation. Again, Table 2.4 suggests that original cross-validation tends to choose larger models, while adjusted cross-validation leads to results comparable with the scaled Lasso.

Table 2.4: Prediction performance of eight methods in the real data example at three penalty levels λ_0, $\lambda_j = \{2^{j-1}(\log p)/n\}^{1/2}(j = 1, 2, 3)$. The prediction mean squared error ($\times 10^2$), the estimated model size and the correlation coefficient ($\times 10^2$) between fitted and observed responses are tabulated for each method.

		#cov = 200			#cov = 3000		
Method		P-MSE	$\|\widehat{\boldsymbol{\beta}}\|_0$	corr	P-MSE	$\|\widehat{\boldsymbol{\beta}}\|_0$	corr
PMLE	λ_1	0·94	12	67·1	0·97	12	63·5
	λ_2	0·97	9	63·5	1·04	7	59·8
	λ_3	1·09	6	57·6	1·23	3	52·2
BC	λ_1	0·93	13	68·2	0·96	15	64·6
	λ_2	0·95	10	64·7	1·01	9	60·9
	λ_3	1·04	7	59·4	1·17	4	53·1
Scaled Lasso	λ_1	0·93	13	68·4	0·96	17	64·3
	λ_2	0·94	10	65·2	0·98	10	61·7
	λ_3	1·02	7	60·8	1·13	5	53·9
Scaled mcp	λ_1	1·03	6	66·4	1·08	8	62·3
	λ_2	1·03	5	63·4	1·06	5	60·0
	λ_3	1·12	3	59·1	1·18	2	54·9
Scaled scad	λ_1	1·00	11	68·9	1·01	14	65·1
	λ_2	0·95	10	68·8	0·98	10	65·9
	λ_3	1·01	8	65·0	1·09	5	59·7
C-V Lasso		0·94	15	69·0	0·99	25	63·8
C-V Lasso/LSE1		0·97	11	64·8	0·98	12	62·5
C-V Lasso/LSE2		0·97	11	66·8	1·09	12	62·6

PMLE, ℓ_1 penalized maximum likelihood estimator; BC, bias-corrected PMLE; mcp, minimax concave penalty; scad, smoothly clipped absolute deviation penalty; C-V Lasso/LSE1, the Lasso with adjusted cross-validation; C-V Lasso/LSE2, the least squares estimator with the Lasso selection and adjusted cross-validation, #cov, the number of covariates considered; corr, the correlation coefficient between fitted and observed responses; P-MSE, prediction mean squared error.

2.5 Discussion

Although theory for the scaled Lasso is more complete, several theoretical and simulation studies (Fan & Peng, 2004; Zhang, 2010) have supported the use of concave penalized least squares estimators with the minimax concave penalty or the smoothly clipped absolute deviation penalty. For variable selection, the Lasso requires a restrictive irrepresentability condition on the design matrix (Meinshausen & Bühlmann, 2006; Zhao & Yu, 2006), while concave penalized least squares estimators require weaker conditions. Model selection consistency implies oracle properties in estimation and prediction in the sense of matching the performance of an oracle expert with the knowledge of the set of relevant variables $S = \{j : \beta_j^* \neq 0\}$. An important issue with the concave penalized least squares estimator is the multiplicity of local minimizers of the penalized loss. In this regard, Zhang (2010) proved selection consistency and oracle properties of the minimax concave penalization estimator for the local minimum computed by the penalized linear unbiased selection algorithm.

Throughout this chapter, we have considered $\lambda_0 = A\{(2/n)\log p\}^{1/2}$ with $A > 1$. This choice is somewhat conservative from a number of points of view. Simulation results suggest that the requirement $A > 1$ is a mathematical technicality. If $|X'\varepsilon/n|_\infty \leq \lambda_*$ with large probability for a standard normal vector ε, the theoretical results in this paper are all valid under $\mathrm{pr}_{\boldsymbol{\beta},\sigma}$ when λ_0 is replaced by the smaller $\min(\lambda_0, A\lambda_*)$. The value of λ_* can be estimated by simulation with the given X and a separately generated ε. A somewhat sharper choice of λ_0 is $A\{(2/n)\log(p/s)\}^{1/2}$ with the unknown $s = |\beta^*|_0$ (Zhang, 2010), or its simulated version with $\lambda_* = \max_{|T|=s} |X_T'\varepsilon|_2/|T|^{1/2}$. The difference between the two λ_0 is limited unless $\log p = \{1 + o(1)\}\log n$.

In the proof of our theoretical results for the scaled Lasso, we use oracle inequalities for fixed penalty which unify and somewhat sharpen existing results. We now discuss

this. Define

$$\eta^*(\lambda,\xi) = \min_T \ 2^{-1}\Big[\eta(\lambda,\xi,\boldsymbol{\beta}^*,T) + \big\{\eta^2(\lambda,\xi,\boldsymbol{\beta}^*,T) - 16\lambda^2\|\boldsymbol{\beta}^*_{T^c}\|_1^2\big\}^{1/2}\Big] \quad (2.25)$$

as a sharper version of $\eta(\lambda,\xi,\boldsymbol{\beta}^*,T)$ in (2.9).

Theorem 2.3. *Let $\widehat{\boldsymbol{\beta}}(\lambda)$ be the minimizer of (2.1) with $\rho(t) = t$. Let $\boldsymbol{\beta}^* \in \mathbb{R}^p$ be a target vector and $\xi > 1$. Then, in the event $\|\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*)\|_\infty/n \le \lambda(\xi-1)/(\xi+1)$, we have*

$$\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2/n \le \min\big\{\eta_*(\lambda,\xi), \eta^*(\lambda,\xi)\big\} \quad (2.26)$$

with $\eta_*(\lambda,\xi)$ in (2.11). Moreover, in the same event and with $\mu(\lambda,\xi)$ in (2.15),

$$\|\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^*\|_1 \le \mu(\lambda,\xi). \quad (2.27)$$

The interpretations of (2.26) and (2.27) are given in (2.14) and (2.16), along with their relationship to several existing results. We note here that the condition $\kappa(\xi,S) \asymp 1$ for (2.14) and (2.16), weaker than the parallel condition on the restricted eigenvalue (Bickel et al., 2009), can be slightly weakened by using $F_1(\xi,S)$ in (3.9) (Ye & Zhang, 2010).

A parallel study Belloni et al. (2011) expressed the same estimator of β alone in a different format and considered a different algorithm, called square-root Lasso. We note that (2.3) can be easily implemented with existing algorithms, while the loss function in β in Belloni et al. (2011) is identical to the minimum of (2.5) over σ. Also, we note that (2.3) and (3.5) allow concave penalties and degrees of freedom

adjustments as in Zhang (2010).

2.6 Proofs

Here we prove Proposition 2.1, Theorem 2.3, Theorem 2.1 and then Theorem 2.2.

Proof of Proposition 2.1. (i) Since $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\sigma\lambda_0)$ is a solution of (3.4) at $\lambda = \sigma\lambda_0$,

$$\left\{(\partial/\partial \boldsymbol{w}) L_{\lambda_0}(\boldsymbol{w}, \sigma)\Big|_{\boldsymbol{w}=\widehat{\boldsymbol{\beta}}(\sigma\lambda_0)}\right\}_j = 0,$$

for all $\widehat{\beta}_j(\sigma\lambda_0) \neq 0$. Since $\{j : \widehat{\beta}_j(\lambda)\}$ is unchanged in a neighborhood of $\sigma\lambda_0$, $[(\partial/\partial\sigma)\{\widehat{\boldsymbol{\beta}}(\sigma\lambda_0)/\sigma\}]_j = 0$ for $\widehat{\beta}_j(\sigma\lambda_0) = 0$. Thus,

$$\frac{\partial}{\partial\sigma} L_{\lambda_0}\{\widehat{\boldsymbol{\beta}}(\sigma\lambda_0), \sigma\} = \frac{\partial}{\partial t} L_{\lambda_0}\{\widehat{\boldsymbol{\beta}}(\sigma\lambda_0), t\}\Big|_{t=\sigma} = \frac{1-a}{2} - \frac{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\sigma\lambda_0)\|_2^2}{2n\sigma^2}.$$

(ii) The convergence of (2.3) and (3.5) follows from the joint convexity of $L_{\lambda_0}(\beta, \sigma)$. The scale invariance follows from $L_0(c\boldsymbol{\beta}, c\sigma; \boldsymbol{X}, c\boldsymbol{y}) = cL_0(\boldsymbol{\beta}, \sigma; \boldsymbol{X}, \boldsymbol{y})$, where $L_0(\boldsymbol{\beta}, \sigma; \boldsymbol{X}, \boldsymbol{y})$ expresses the dependence of (2.5) on the data $(\boldsymbol{X}, \boldsymbol{y})$. □

Proof of Theorem 2.3. (i) Let $\widehat{\beta} = \widehat{\beta}(\lambda)$. Since $\sigma^* z^* = \|\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*)\|_\infty/n$ and $\dot{\rho}(|\widehat{\beta}_j|/\lambda) = 1$ for $\widehat{\beta}_j \neq 0$, the inner product of $w - \widehat{\beta}$ and the Karush–Kuhn–Tucker condition (3.4) yields

$$(\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{w})'(\boldsymbol{X}\widehat{\beta} - X\beta^*)/n \leq \lambda(\|\boldsymbol{w}\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1) + \sigma^* z^* \|\boldsymbol{w} - \widehat{\boldsymbol{\beta}}\|_1.$$

Since $2(\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{w})'(\boldsymbol{X}\widehat{\beta} - X\beta^*) = \|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \|\boldsymbol{X}\boldsymbol{h}\|_2^2 - \|\boldsymbol{X}\boldsymbol{\beta}^* - \boldsymbol{X}\boldsymbol{w}\|_2^2$, this gives the basic inequality (2.24). Let $\boldsymbol{h} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. Since $\sigma^* z^* \leq \lambda(\xi-1)/(\xi+1)$, $\lambda\{\|\boldsymbol{w}\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1\} + \sigma^* z^* \|\boldsymbol{w} - \widehat{\boldsymbol{\beta}}\|_1$ is no greater than $b\|(\boldsymbol{w} - \widehat{\boldsymbol{\beta}})_T\|_1 + 2\lambda\|\boldsymbol{w}_{T^c}\|_1 - (b/\xi)\|(\boldsymbol{w} - \widehat{\boldsymbol{\beta}})_{T^c}\|_1$ with $b = 2\xi\lambda/(\xi+1)$. Thus, (2.24) implies

$$\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{w}\|_2^2/n + \|\boldsymbol{X}\boldsymbol{h}\|_2^2/n + (2b/\xi)\|(\boldsymbol{w} - \widehat{\boldsymbol{\beta}})_{T^c}\|_1 \leq 2c + 2b\|(\boldsymbol{w} - \widehat{\boldsymbol{\beta}})_T\|_1 \quad (2.28)$$

with $c = \|\boldsymbol{X}\boldsymbol{\beta}^* - \boldsymbol{X}\boldsymbol{w}\|_2^2/(2n) + 2\lambda\|\boldsymbol{w}_{T^c}\|_1$. For $T = \emptyset$ and $w = \beta^*$, (2.28) directly yields $|Xh|_2^2/n \leq c = 2\lambda|\beta^*|_1$. For general $\{w, T\}$, we want to prove

$$\|\boldsymbol{X}\boldsymbol{h}\|_2^2/n \leq \eta(\lambda, \xi, \boldsymbol{w}, T) = 2c + b^2/a, \quad a = \kappa^2(\xi, T)/|T|.$$

It suffices to consider $\|\boldsymbol{X}\boldsymbol{h}\|_2^2/n \geq 2c$. In this case, $\widehat{\boldsymbol{\beta}} - \boldsymbol{w} \in \mathscr{C}(\xi, T)$ by (2.28), so that by (2.10)

$$a\|(\boldsymbol{w} - \widehat{\boldsymbol{\beta}})_T\|_1^2 \leq \|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{w}\|_2^2/n. \quad (2.29)$$

Let $x = \|(\boldsymbol{w} - \widehat{\boldsymbol{\beta}})_T\|_1$ and $y = \|\boldsymbol{X}\boldsymbol{h}\|_2^2/n$. It follows from (2.28) and (2.29) that $ax^2 + y \leq 2c + 2bx$. For such (x, y), $y - 2c \leq \max_x\{2bx - ax^2\} = b^2/a$. This gives $y \leq 2c + b^2/a = \eta(\lambda, \xi, \boldsymbol{w}, T)$.

For $\boldsymbol{w} = \boldsymbol{\beta}^*$, it suffices to consider the case $y > c = 2\lambda\|\boldsymbol{\beta}^*_{T^c}\|_1$, where the cone condition holds for $\widehat{\beta} - \beta^*$. Now, (x, y) satisfies $ax^2 \leq y \leq c + bx$. The maximum of

y, attained at $ax^2 = c + bx$, is

$$c + b\{b + (b^2 + 4ac)^{1/2}\}/(2a) = \left[\eta(\lambda, \xi, \boldsymbol{\beta}^*, T) + \{\eta^2(\lambda, \xi, \boldsymbol{\beta}^*, T) - 4c^2\}^{1/2}\right]/2 \quad (2.30)$$

(ii) Let $0 < \nu < 1$ and $T \subset \{1, \ldots, p\}$. It follows from (2.28) with $\boldsymbol{w} = \boldsymbol{\beta}^*$ that

$$(1 + \xi)\|\boldsymbol{X}\boldsymbol{h}\|_2^2/n + 2\lambda\|\boldsymbol{h}_{T^c}\|_1 \leq 2\lambda(\xi + 1)\|\boldsymbol{\beta}^*_{T^c}\|_1 + 2\xi\lambda\|\boldsymbol{h}_T\|_1.$$

It suffices to consider $\nu|\boldsymbol{h}|_1 \geq (\xi + 1)|\boldsymbol{\beta}^*_{T^c}|_1$. In this case

$$(1 + \xi)\|\boldsymbol{X}\boldsymbol{h}\|_2^2/n + 2\lambda(1 - \nu)\|\boldsymbol{h}_{T^c}\|_1 \leq 2\lambda(\xi + \nu)\|\boldsymbol{h}_T\|_1.$$

Thus, $(1 - \nu)\|\boldsymbol{h}_{T^c}\|_1 \leq (\xi + \nu)\|\boldsymbol{h}_T\|_1$, or equivalently $\boldsymbol{h} \in \mathscr{C}\{(\xi + \nu)/(1 - \nu), T\}$. It follows from (2.10) that $\|\boldsymbol{X}\boldsymbol{h}\|_2^2/n \geq \|\boldsymbol{h}_T\|_1^2\kappa^2\{(\xi + \nu)/(1 - \nu), T\}/|T|$, so that

$$(1 + \xi)\|\boldsymbol{h}_T\|_1^2\kappa^2\{(\xi + \nu)/(1 - \nu), T\}/|T| + 2(1 - \nu)\lambda\|\boldsymbol{h}_{T^c}\|_1 \leq 2(\xi + \nu)\lambda\|\boldsymbol{h}_T\|_1. \quad (2.31)$$

Let $x = \|\boldsymbol{h}_T\|_1$ and $y = \|\boldsymbol{h}_{T^c}\|_1$. Write (2.31) as $ax^2 + by \leq cx$. Subject to this inequality, the maximum of $x + y$ is $\max_{x \geq 0}\{x + (cx - ax^2)/b\}$. This maximum, attained at $2ax = b + c$, is $x(b + c)/(2b) = (b + c)^2/(4ab)$. Thus,

$$\|\boldsymbol{h}\|_1 \leq \frac{\{2(\xi + 1)\lambda\}^2|T|}{4(1 + \xi)\kappa^2\{(\xi + \nu)/(1 - \nu), T\}\{2(1 - \nu)\lambda\}} = \frac{(\xi + 1)\lambda|T|/(1 - \nu)}{2\kappa^2\{(\xi + \nu)/(1 - \nu), T\}}.$$

This gives $\|\boldsymbol{h}\|_1 \leq \mu(\lambda, \xi)$ for $\nu|\boldsymbol{h}|_1 \geq (\xi + 1)|\boldsymbol{\beta}^*_{T^c}|_1$. □

Proof of Theorem 2.1. Assume $\tau_0 < 1$ without loss of generality. Consider $t \geq \sigma^*(1-\tau_0)$ and the penalty level $\lambda = t\lambda_0$ for the Lasso. Since $z^*\sigma^* \leq \sigma^*(1-\tau_0)\lambda_0(\xi-1)/(\xi+1) \leq \lambda(\xi-1)/(\xi+1)$ and $\sigma^* = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2/n^{1/2}$, the Cauchy–Schwarz inequality and (2.26) imply

$$\big|\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(t\lambda_0)\|_2/n^{1/2} - \sigma^*\big| \leq \|\boldsymbol{X}\widehat{\boldsymbol{\beta}}(t\lambda_0) - \boldsymbol{X}\boldsymbol{\beta}^*\|_2/n^{1/2} \leq \eta_*^{1/2}(t\lambda_0, \xi).$$

Since $\eta_*^{1/2}(t\lambda_0, \xi) \leq \sigma^*\tau_0$ for $t < \sigma^*$, the derivative (2.7) of the loss with $a = 0$ satisfies

$$2t^2\frac{\partial}{\partial t}L_{\lambda_0}\{\widehat{\boldsymbol{\beta}}(t\lambda_0), t\} = t^2 - \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(t\lambda_0)\|_2^2/n \leq t^2 - (\sigma^*)^2(1-\tau_0)^2 = 0$$

at $t = \sigma^*(1-\tau_0)$. This implies $\widehat{\sigma} \geq \sigma^*(1-\tau_0)$ by the strict convexity of the profile loss (2.5) in σ. For $t > \sigma^*$, $\eta_*^{1/2}(t\lambda_0, \xi) \leq t\tau_0$ by (2.9) and (2.11), so that at $t = \sigma^*/(1-\tau_0)$,

$$t^2 - \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(t\lambda_0)\|_2^2/n \geq t^2 - \big(\sigma^* + t\tau_0\big)^2 \geq 0.$$

This implies $\sigma^* \geq \widehat{\sigma}(1-\tau_0)$ by the strict convexity of (2.5) in σ. Thus, the first part of (2.12) holds. Moreover,

$$\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2/n^{1/2} \leq \eta_*^{1/2}(\widehat{\sigma}\lambda_0, \xi) \leq \eta_*^{1/2}\{\sigma^*\lambda_0/(1-\tau_0), \xi\} \leq \sigma^*\tau_0/(1-\tau_0).$$

Finally, since $\mathrm{pr}_{\boldsymbol{\beta},\sigma}[\|\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/n\|_\infty \leq \sigma\{(2/n)\log p\}^{1/2}] \to 1$, (2.13) follows from (2.12). □

The proof of Theorem 2 requires the following lemma.

Lemma 2.1. *Let T_m have the t-distribution with m degrees of freedom. Then, there exists $\epsilon_m \to 0$ such that for all $t > 0$*

$$\mathrm{pr}\big[T_m^2 > m\{e^{2t^2/(m-1)} - 1\}\big] \leq (1+\epsilon_m)e^{-t^2}/(\pi^{1/2}t). \tag{2.32}$$

Proof of Lemma 2.1. Let $x = [m\{e^{2t^2/(m-1)} - 1\}]^{1/2}$. Since T_m has the t-distribution,

$$\begin{aligned}
\mathrm{pr}\Big(T_m^2 > x^2\Big) &= \frac{2\Gamma\{(m+1)/2\}}{\Gamma(m/2)(m\pi)^{1/2}} \int_x^\infty \Big(1 + \frac{u^2}{m}\Big)^{-(m+1)/2} du \\
&\leq \frac{2\Gamma\{(m+1)/2\}}{x\Gamma(m/2)(m\pi)^{1/2}} \int_x^\infty \Big(1 + \frac{u^2}{m}\Big)^{-(m+1)/2} u du \\
&= \frac{2\Gamma\{(m+1)/2\}m}{x\Gamma(m/2)(m\pi)^{1/2}(m-1)} \Big(1 + \frac{x^2}{m}\Big)^{-(m-1)/2}.
\end{aligned}$$

Since $x \geq t\{2m/(m-1)\}^{1/2}$,

$$\mathrm{pr}\Big(T_m^2 > x^2\Big) \leq \frac{\sqrt{2}\Gamma\{(m+1)/2\}}{\Gamma(m/2)(m-1)^{1/2}} \frac{e^{-t^2}}{t\pi^{1/2}} = (1+\epsilon_m)\frac{e^{-t^2}}{t\pi^{1/2}},$$

where $\epsilon_m = \{2/(m-1)\}^{1/2}\Gamma\{(m+1)/2\}/\Gamma(m/2) - 1 \to 0$ as $m \to \infty$. □

Proof of Theorem 2.2. We need to express τ_*^2 as a function of σ at $\sigma = \sigma^*$ in the proof. Define

$$\phi(\sigma) = \lambda_0\mu(\sigma\lambda_0, \xi)/\sigma,\ \phi_+ = \frac{\phi(\sigma^*)\xi}{(\xi+1)\{1-\phi(\sigma^*)\}_+},\ \phi_- = \frac{\phi(\sigma^*)(\xi-1)}{\xi+1}.$$

We have $\tau_*^2 = \phi(\sigma^*) < 1$, $\phi_- \leq \phi(\sigma^*)$ and $\phi_+ \leq \phi(\sigma^*)/(1-\phi(\sigma^*))$.

(i) Consider $z^* \leq (1-\phi_-)\lambda_0(\xi-1)/(\xi+1)$. Let $\lambda = t\lambda_0$ and $\boldsymbol{h} = \widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^*$.

Since $\|\boldsymbol{X}'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^*)/n\|_\infty = z^*\sigma^*$, the Karush–Kuhn–Tucker condition (3.4) gives

$$\begin{aligned} -(z^*\sigma^*+\lambda)\|\boldsymbol{h}\|_1 &\le (\boldsymbol{X}\boldsymbol{h})'\{\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^*+\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda)\}/n \\ &= (\sigma^*)^2-\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda)\|_2^2/n \\ &= (\boldsymbol{X}\boldsymbol{h})'\{2(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^*)-\boldsymbol{X}\boldsymbol{h}\}/n \le 2z^*\sigma^*\|\boldsymbol{h}\|_1 \quad (2.33) \end{aligned}$$

as lower and upper bounds for $(\sigma^*)^2-\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda)\|_2^2/n$. This is a key point in the proof.

For $t \ge \sigma^*(1-\phi_-)$, $z^*\sigma^* \le t\lambda_0(\xi-1)/(\xi+1) = \lambda(\xi-1)/(\xi+1)$, so that (2.27) in Theorem 2.3 implies $\|\boldsymbol{h}\|_1 \le \mu(t\lambda_0,\xi)$. It follows (2.33) that for $t=\sigma^*(1-\phi_-)$,

$$\begin{aligned} t^2-\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}(t\lambda_0)\|_2^2/n &\le t^2-(\sigma^*)^2+2z^*\sigma^*\mu(t\lambda_0,\xi) \\ &\le 2t(t-\sigma^*)+2t\lambda_0(\xi-1)(\xi+1)^{-1}\mu(\sigma^*\lambda_0,\xi)=0, \end{aligned}$$

due to $\phi_- = (\xi-1)(\xi+1)^{-1}\phi(\sigma^*) = (\xi-1)(\xi+1)^{-1}\lambda_0\mu(\sigma^*\lambda_0,\xi)/\sigma^*$. As in the proof of Theorem 2.1, we find $\widehat{\sigma}/\sigma^* \ge 1-\phi_-$ by (2.7) and the strict convexity of (2.5) in σ.

Now we prove $\widehat{\sigma}/\sigma^* \le 1+\phi_+$. For $t>\sigma^*$, $\mu(t\lambda_0,\xi) \le (t/\sigma^*)\mu(\sigma^*\lambda_0,\xi)$ by (2.15). Thus, since $(\xi-1)/(\xi+1)+1 = 2\phi_+\{1-\phi(\sigma^*)\}/\phi(\sigma^*)$ and $\phi_+ \le (1+\phi_+)\phi(\sigma^*)$, for $t/\sigma^* = 1+\phi_+$ (2.33) and (2.27) imply that

$$\begin{aligned} & t^2-\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}(t\lambda_0)\|_2^2/n \\ \ge\ & t^2-(\sigma^*)^2-(z^*\sigma^*+t\lambda_0)\mu(t\lambda_0,\xi) \end{aligned}$$

$$
\begin{aligned}
&\geq (t+\sigma^*)\sigma^*\phi_+ - \{(\xi-1)/(\xi+1)+1+\phi_+\}t\lambda_0\mu(\sigma^*\lambda_0,\xi) \\
&= (\sigma^*)^2\big((2+\phi_+)\phi_+ - [2\phi_+\{1-\phi(\sigma^*)\}/\phi(\sigma^*)+\phi_+](1+\phi_+)\phi(\sigma^*)\big) \\
&= (\sigma^*)^2\phi_+\{\phi(\sigma^*)(1+\phi_+)-\phi_+\} > 0.
\end{aligned}
$$

It follows that $\widehat{\sigma}/\sigma^* \leq 1+\phi_+$ by convexity.

Since $1-\phi_- \leq \widehat{\sigma}/\sigma^* \leq 1+\phi_+$, $\|\widehat{\boldsymbol{\beta}}(\widehat{\sigma}\lambda_0)-\boldsymbol{\beta}^*\|_1 \leq \mu(\widehat{\sigma}\lambda_0,\xi) \leq \mu(\sigma^*\lambda_0,\xi)(1+\phi_+)$. This completes the proof of (2.17).

(ii) Let $z_j = \boldsymbol{x}_j'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^*)/(n\sigma^*)$ with $z^* = \max_{j\leq p}|z_j|$. Under $\mathrm{pr}_{\boldsymbol{\beta}^*,\sigma}$, $\varepsilon^* = \boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^*$ is a vector of independent and identically distributed normal variables with zero mean. Since $\sigma^* = \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^*\|/n^{1/2}$, $z_j/\{(1-z_j^2)/(n-1)\}^{1/2}$ follows a t-distribution with $n-1$ degrees of freedom. Lemma 2.1 with $m = n-1$ and $t^2 = \log(p/\epsilon) > 2$ implies

$$
\begin{aligned}
&\mathrm{pr}_{\boldsymbol{\beta}^*,\sigma}\Big[\frac{(n-1)z_j^2}{1-z_j^2} > (n-1)\{e^{2t^2/(n-2)}-1\}\Big] \\
\leq\ & \frac{1+\epsilon_{n-1}}{\pi^{1/2}t}e^{-t^2} = \frac{(1+\epsilon_{n-1})\epsilon/p}{\{\pi\log(p/\epsilon)\}^{1/2}}.
\end{aligned} \tag{2.34}
$$

Since $e^a - 1 \leq \sum_{k=1}^{\infty} a^k/2^{k-1} = a/(1-a/2)$ for any $0 < a < 2$,

$$
(n-1)\{e^{2t^2/(n-2)}-1\} \leq \frac{2(n-1)t^2/(n-2)}{1-t^2/(n-2)} \leq \frac{2(n-1)t^2/n}{1-2t^2/n}. \tag{2.35}
$$

The combination of (2.34) and (2.35) yields

$$
\mathrm{pr}_{\boldsymbol{\beta}^*,\sigma}\big[|z_j| > \{2\log(p/\epsilon)/n\}^{1/2}\big] \ = \ \mathrm{pr}_{\boldsymbol{\beta}^*,\sigma}\Big\{\frac{(n-1)z_j^2}{1-z_j^2} > \frac{2(n-1)t^2/n}{1-2t^2/n}\Big\}
$$

$$\begin{aligned} &\leq \ \mathrm{pr}_{\boldsymbol{\beta}^*,\sigma}\Big\{\frac{(n-1)z_j^2}{1-z_j^2} > (n-1)(e^{\frac{2t^2}{n-2}}-1)\Big\} \\ &\leq \ (1+\epsilon_{n-1})(\epsilon/p)/\{\pi\log(p/\epsilon)\}^{1/2}. \end{aligned}$$

Since $\lambda_0 \geq \{(2/n)\log(p/\epsilon)\}^{1/2}(\xi+1)/\{(\xi-1)(1-\phi_-)\}$, this bounds the tail probability of $z^* = \max_{j\leq p}|z_j|$ by the union bound. Since $n(\sigma^*/\sigma)^2$ follows the χ^2_n distribution, $n^{1/2}(\sigma^*/\sigma - 1)$ converges to $N(0, 1/2)$ in distribution, which then implies (2.18) by (2.17) under $\phi(\sigma) = o(n^{-1/2})$. □

Chapter 3
Estimation of Matrix Inversion

3.1 Introduction

This chapter concerns the estimation of the matrix inversion $\mathbf{\Theta}^*$ satisfying $\overline{\mathbf{\Sigma}}\mathbf{\Theta}^* \approx \boldsymbol{I}$, given a data matrix $\overline{\mathbf{\Sigma}}$. When $\overline{\mathbf{\Sigma}}$ is a sample covariance matrix, our problem is the estimation of the inverse of the corresponding population covariance matrix. The inverse covariance matrix is also called precision matrix or concentration matrix. With the dramatic advances in technology, the number of covariates is of greater order than the sample size n in many statistical and engineering applications. In this case, the sample covariance matrix is always singular and thus it is difficult to compute the precision matrix. In such cases, a certain type of sparsity condition is required for proper estimation the precision matrix and for theoretical investigation of the estimation problem. In this paper, we will impose for simplicity an ℓ_0 (maximum degree) sparsity condition on the target inverse matrix $\mathbf{\Theta}^*$.

Many approaches have been proposed to estimate the sparse inverse matrix in the high dimensional setting. The ℓ_1 penalization is one of the most popular methods. Lasso-type methods, or convex minimization algorithms with the ℓ_1 penalty on all entries of $\mathbf{\Theta}^*$, have been discussed by Banerjee et al. (2008), Friedman et al. (2008), and more, and by Yuan & Lin (2007) with ℓ_1 penalization on the off-diagonal matrix only. This is refereed to as the graphical Lasso (GLasso) due to the connection of the precision matrix to Gaussian Markov graphical models. In this GLasso framework, Rothman et al. (2008) proved the convergence rate $\{((p+s)/n)\log p\}^{1/2}$ in the

Frobenius norm and $\{(s/n)\log p\}^{1/2}$ in the spectrum norm, where s is the number of nonzero entries in the off-diagonal matrix. Ravikumar et al. (2008) provided sufficient conditions for model selection consistency of this ℓ_1-regularized MLE. Lam & Fan (2009) studied on a general penalty function and achieved a sharper bound of order $\{(s/n)\log p\}^{1/2}$ under the Frobenius norm for the ℓ_1 penalty. Similar convergence rates have been also studied under the Frobenius norm in a unified framework for penalized estimation in Negahban et al. (2009). Since the spectrum norm can be controlled via the Frobenius norm, this provides a sufficient condition $(s/n)\log p \to 0$ for the convergence under the spectrum norm to the unknown precision matrix. This is a very strong condition since s is of the order dp for banded precision matrices, where d is the matrix degree, i.e. the largest number of nonzero entries in the columns.

Some recent work suggests a weaker sufficient condition with the matrix degree. Yuan (2010) estimated each column of the inverse matrix by Dantzig selector and then seek a symmetric matrix close to the column estimation. When ℓ_1 norm of the precision matrix is bounded, this method can achieve a convergence rate of order $d\{(\log p)/n\}^{1/2}$ based on several matrix norms. The CLIME estimator, introduced by Cai et al. (2011), has the same order of convergence rate, which uses the plug-in method with Dantzig selector to estimate each column, but followed by a simpler symmetrization step. They also require the boundedness of the ℓ_1 norm of the unknown. In Yang & Kolaczyk (2010), the Lasso is applied to estimate the columns of the target matrix under the assumption of equal diagonal, and the estimation error is studied in the Frobenius norm for $p = n^{\nu}$. This column-by-column idea reduces a graphical model to a regression model. It was first introduced in Meinshausen & Bühlmann (2006) for identifying nonzero variables in a graphical model, called neighborhood selection.

In this chapter, we propose to apply the scaled Lasso, introduced in Chapter 2, column by column to estimate a precision matrix in the high dimensional setting. Based on the connection of precision matrix to linear regression by the block inversion

formula, we construct a column estimator with the scaled Lasso, a joint estimator for the regression coefficients and noise level. Since we only need the sample covariance matrix in our procedure, this estimator could be extended to generate an approximate inverse of a nonnegative data matrix in a general setting. This scaled Lasso algorithm provides a fully specified map from the space of nonnegative-definite matrices to the space of symmetric matrices. For each column, the penalty level of the scaled Lasso is determined by data via convex minimization, without using cross-validation.

We study theoretical properties of the proposed estimator for a precision matrix under a normality assumption. More precisely, we assume that the data matrix is the sample covariance matrix $\overline{\mathbf{\Sigma}} = \boldsymbol{X}'\boldsymbol{X}/n$, where the rows of $\boldsymbol{X}$ are iid $N(0, \mathbf{\Sigma}^*)$. Under conditions on the spectrum norm and degree of the inverse of $\mathbf{\Sigma}^*$, we prove that the proposed estimator guarantees the rate of convergence of order $d\{(\log p)/n\}^{1/2}$ in the spectrum norm. The conditions are weaker than those in the existing analyses of other ℓ_1 algorithms, which typically require the boundedness of the ℓ_1 norm. When the ℓ_1 norm of the target matrix diverges to infinity, the analysis of the proposed estimator guarantees a faster convergence rate than that of the existing literature. We state this main result of this chapter in the following theorem.

Theorem 3.1. *Let $\widehat{\mathbf{\Theta}}$ be the scaled Lasso estimator, defined in (3.5), (3.6) and (3.7) below, based on n iid observations from $N(0, \mathbf{\Sigma}^*)$. Let ρ_* and ρ^* be the smallest and largest eigenvalues of correlation matrix of $\mathbf{\Sigma}^*$, Θ^* be the inverse of $\mathbf{\Sigma}^*$ and $d = \max_i \#\{j : \Theta^*_{ij} \neq 0\}$ be the maximum degree of $\mathbf{\Theta}^*$. Suppose that $d\sqrt{(\log p)/n} \to 0$, the diagonal entries of the target matrix $\mathbf{\Theta}^*$ are uniformly bounded, ρ_* is bounded from 0 and $(\rho^*/\rho_*)\{(d/n)\log p\}^{1/2} < a$ for a small fixed a. Then, the spectrum norm of the*

estimation error $\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*$ *is bounded by*

$$\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_2 = O_P(d\sqrt{(\log p)/n}).$$

The convergence of the proposed scaled Lasso estimator under the sharper spectrum norm condition on $\boldsymbol{\Theta}^*$, instead of the stronger bounded ℓ_1 condition, is not entirely technical. It is a direct consequence of the faster convergence rate of the scaled Lasso estimator of the noise level in linear regression. To the best of our knowledge, it is unclear if other ℓ_1 algorithms also achieve this fast convergence rate, either for the estimation of the noise level in linear regression or for the estimation of a precision matrix under the spectrum norm. However, it is still possible that this difference between the scaled Lasso and other methods is due to potentially coarser specification of the penalty level in other algorithms (e.g. cross validation) or a less accurate error bound in other analyses.

The chapter is organized as follows. In Section 3.2, we present the estimation for the inversion of a nonnegative definite matrix via the scaled Lasso. In Section 3.3, we study error bounds of the proposed estimator for precision matrix. Simulation studies are presented in Section 3.4. In Section 3.5, we discuss oracle inequalities for the scaled Lasso with unnormalized predictors and the estimation of inverse correlation matrix. Section 3.6 includes all the proofs.

We use the following notation throughout this chapter. For a vector $\boldsymbol{v} = (v_1, \ldots, v_p)$, $\|\boldsymbol{v}\|_q = (\sum_j |v_j|^q)^{1/q}$ is the ℓ_q norm with the special $\|\boldsymbol{v}\| = \|\boldsymbol{v}\|_2$ and the usual extensions $\|\boldsymbol{v}\|_\infty = \max_j |v_j|$ and $\|\boldsymbol{v}\|_0 = \#\{j : v_j \neq 0\}$. For matrices $\boldsymbol{M}$, $\boldsymbol{M}_{j,*}$ is the j-th column of $\boldsymbol{M}$, $\boldsymbol{M}_{A,B}$ represents the submatrix of $\boldsymbol{M}$ with rows in A and columns in B, $\|\boldsymbol{M}\|_q = \sup_{\|\boldsymbol{v}\|_q=1} \|\boldsymbol{M}\boldsymbol{v}\|_q$ is the ℓ_q matrix norm. In particular, $\|\cdot\|_2$ is the spectrum norm for symmetric matrices. Moreover, we denote the set $\{j\}$

by j and denote the set $\{1, \ldots, p\} \setminus \{j\}$ by $-j$ in the subscripts.

3.2 Matrix inversion via scaled Lasso

Let $\overline{\boldsymbol{\Sigma}}$ be a nonnegative-definite data matrix and $\boldsymbol{\Theta}^*$ be a positive-definite target matrix with $\overline{\boldsymbol{\Sigma}}\boldsymbol{\Theta}^* \approx \boldsymbol{I}$. In this section, we describe the relationship between positive-definite matrix inversion and linear regression and propose an estimator for $\boldsymbol{\Theta}^*$ via scaled Lasso, a joint convex minimization for the estimation of regression coefficients and noise level.

We use scaled Lasso to estimate $\boldsymbol{\Theta}^*$ column by column. Define $\sigma_j > 0$ and $\boldsymbol{\beta} \in \mathbb{R}^{p \times p}$ by

$$\sigma_j^2 = (\boldsymbol{\Theta}^*_{jj})^{-1}, \quad \boldsymbol{\beta}_{*,j} = -\boldsymbol{\Theta}^*_{*,j}\sigma_j^2 = -\boldsymbol{\Theta}^*_{*,j}(\boldsymbol{\Theta}^*_{jj})^{-1}. \tag{3.1}$$

In the matrix form, we have the following relationship

$$\text{diag}\boldsymbol{\Theta}^* = \text{diag}(\sigma_j^{-2}, j = 1, \ldots, p), \quad \boldsymbol{\Theta}^* = -\boldsymbol{\beta}(\text{diag}\boldsymbol{\Theta}^*). \tag{3.2}$$

Let $\boldsymbol{\Sigma}^* = (\boldsymbol{\Theta}^*)^{-1}$. Since $(\partial/\partial \boldsymbol{b}_{-j})\boldsymbol{b}'\boldsymbol{\Sigma}^*\boldsymbol{b} = 2\boldsymbol{\Sigma}^*_{-j,*}\boldsymbol{b} = 0$ at $\boldsymbol{b} = \boldsymbol{\beta}_{*,j}$, one may estimate the j-th column of $\boldsymbol{\beta}$ by minimizing the ℓ_1 penalized quadratic loss. In order to shrink the estimation coefficients on the same scale, we adjust the penalty function with a normalizing factor, which leads to the ℓ_1 penalized quadratic loss as follows,

$$\boldsymbol{b}'\overline{\boldsymbol{\Sigma}}\boldsymbol{b}/2 + \lambda \sum_{k=1}^{p} \overline{\Sigma}_{kk}^{1/2} |b_k|$$

subject to $b_j = -1$. This is actually the Lasso for a linear regression model with normalized preditors. In practice, we first normalize the predictors by the weights

$\overline{\Sigma}_{kk}^{-1/2}(k \neq j)$ and then the minimization problem can be solved by algorithms for the Lasso estimation. This is similar to Yuan (2010) and Cai et al. (2011) who used the Dantzig selector to estimate each column. However, one still needs to choose a penalty level λ and to estimate σ_j to recover $\mathbf{\Theta}^*$ via (3.2). A solution to resolve these two issues is the scaled Lasso:

$$\{\widehat{\boldsymbol{\beta}}_{*,j}, \widehat{\sigma}_j\} = \arg\min_{\boldsymbol{b},\sigma} \Big\{ \frac{\boldsymbol{b}'\overline{\mathbf{\Sigma}}\boldsymbol{b}}{2\sigma} + \frac{\sigma}{2} + \lambda_0 \sum_{k=1}^{p} \overline{\Sigma}_{kk}^{1/2} |b_k| : b_j = -1 \Big\} \tag{3.3}$$

where $\lambda_0 = A\sqrt{2(\log p^2/\epsilon)/n}$ with a fixed $A > 1$. This is actually (2.6) with normalized parameters $\overline{\Sigma}_{kk}^{1/2} b_k$. Since $\boldsymbol{\beta}'\mathbf{\Sigma}^*\boldsymbol{\beta} = (\text{diag}\mathbf{\Theta}^*)^{-1}\mathbf{\Theta}^*(\text{diag}\mathbf{\Theta}^*)^{-1}$,

$$\text{diag}\big(\boldsymbol{\beta}'\mathbf{\Sigma}^*\boldsymbol{\beta}\big) = (\text{diag}\mathbf{\Theta}^*)^{-1} = \text{diag}(\sigma_j^2, j = 1, \ldots, p).$$

Thus, (3.3) is expected to yield consistent estimates of σ_j.

In Chapter 2, an iterative algorithm (2.3) is provided to compute the scaled Lasso estimator (2.6). We rewrite the algorithm in the form of matrices. For each $j \in \{1, \ldots, p\}$, the Lasso path is given by the estimates $\widehat{\boldsymbol{\beta}}_{-j,j}(\lambda)$ satisfying the following KKT conditions, for all $k \neq j$,

$$\begin{cases} \overline{\Sigma}_{kk}^{-1/2}\overline{\mathbf{\Sigma}}_{k,*}\widehat{\boldsymbol{\beta}}_{*,j}(\lambda) = -\lambda \text{sgn}(\widehat{\beta}_{k,j}(\lambda)), & \widehat{\beta}_{k,j} \neq 0, \\ \overline{\Sigma}_{kk}^{-1/2}\overline{\mathbf{\Sigma}}_{k,*}\widehat{\boldsymbol{\beta}}_{*,j}(\lambda) \in \lambda[-1, 1], & \widehat{\beta}_{k,j} = 0, \end{cases} \tag{3.4}$$

where $\widehat{\boldsymbol{\beta}}_{jj}(\lambda) = -1$. Based on the Lasso path $\widehat{\boldsymbol{\beta}}_{*,j}(\lambda)$, the scaled Lasso estimator $\{\widehat{\boldsymbol{\beta}}_{*,j}, \widehat{\sigma}_j\}$ is computed iteratively by

$$\widehat{\sigma}_j^2 \leftarrow \widehat{\boldsymbol{\beta}}_{*,j}'\overline{\mathbf{\Sigma}}\widehat{\boldsymbol{\beta}}_{*,j}, \quad \lambda \leftarrow \widehat{\sigma}_j\lambda_0, \quad \widehat{\boldsymbol{\beta}}_{*,j} \leftarrow \widehat{\boldsymbol{\beta}}_{*,j}(\lambda). \tag{3.5}$$

Here the penalty level of the Lasso is determined by the data without using cross-validation. We then simply take advantage of the relationship (3.2) and compute the coefficients and noise levels by the scaled Lasso for each column

$$\text{diag}\widetilde{\boldsymbol{\Theta}} = \text{diag}(\widehat{\sigma}_j^{-2}, j = 1, \ldots, p), \quad \widetilde{\boldsymbol{\Theta}} = -\widehat{\boldsymbol{\beta}}(\text{diag}\widetilde{\boldsymbol{\Theta}}). \tag{3.6}$$

It is noticed that a good estimator for $\boldsymbol{\Theta}^*$ should be a symmetric matrix. However, the estimator $\widetilde{\boldsymbol{\Theta}}$ does not have to be symmetric. We improve this estimator by using a symmetrization step as in Yuan (2010),

$$\widehat{\boldsymbol{\Theta}} = \underset{\boldsymbol{M}:\boldsymbol{M}^T=\boldsymbol{M}}{\arg\min} \|\boldsymbol{M} - \widetilde{\boldsymbol{\Theta}}\|_1, \tag{3.7}$$

which can be solved by linear programming. Alternatively, semidefinite programming, which is somewhat more expensive computationally, can be used to produce a nonnegative definite $\widehat{\boldsymbol{\Theta}}$ in (3.7). According to the definition, the new estimator $\widehat{\boldsymbol{\Theta}}$ has the same ℓ_1 error rate as $\widetilde{\boldsymbol{\Theta}}$. A nice property for symmetric matrix is that the spectrum norm is bounded by the ℓ_1 matrix norm. The ℓ_1 matrix norm can be given more explicitly as the maximum ℓ_1 norm of the columns, while the ℓ_∞ matrix norm is the maximum ℓ_1 norm of the rows. Hence, for any symmetric matrix, the ℓ_1 matrix norm is equivalent to the ℓ_∞ matrix norm, so the spectrum norm can be bounded by either of them. Since both our estimator and the target matrix are symmetric, the error bound based on the spectrum norm could be studied by bounding the ℓ_1 error, as typically done in the existing literature. We will discuss these error bounds in Section 3.3.

To sum up, we propose to estimate the matrix inversion by (3.5), (3.6) and (3.7). The iterative algorithm (3.5) computes the regression coefficients and noise level based on a Lasso path determined by (3.4). Then (3.6) translates the resulting estimators

of (3.5) to column estimators and thus a preliminary matrix estimator is constructed. Finally, the symmetrization step (3.7) produces a symmetric estimate for our target matrix.

3.3 Error bounds for precision matrix

In this section, we study the error $\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*$ for the inverse of a covariance matrix, which is our primary example of the target matrix. From now on, we suppose that the data matrix is the sample covariance matrix $\overline{\mathbf{\Sigma}} = \boldsymbol{X}'\boldsymbol{X}/n$, where the rows of $\boldsymbol{X}$ are iid $N(0, \mathbf{\Sigma}^*)$, and the target matrix is $\mathbf{\Theta}^* = (\mathbf{\Sigma}^*)^{-1}$.

Let ρ_* and ρ^* be the smallest and the largest eigenvalues of the correlation matrix $(\text{diag}\mathbf{\Sigma}^*)^{-1/2}\mathbf{\Sigma}^*(\text{diag}\mathbf{\Sigma}^*)^{-1/2}$. Define $S_j = \{i \neq j : \Theta^*_{i,j} \neq 0\}$ and the degree of the matrix

$$d = \deg(\mathbf{\Theta}^*) = \max_j |S_j| + 1.$$

The following theorem gives the convergence rate based on the ℓ_1 matrix norm (ℓ_∞ matrix norm) and spectrum norm.

Theorem 3.2. *Let $\epsilon \in (0, 1/4)$ and $\lambda_0 = A\{(2/n)\log(p^2/\epsilon)\}^{1/2}$ with $A > 1$. Suppose that $\{d(\log p)/n\}^{1/2}\rho^*/\rho_* < a$ for a small fixed a. Then with probability greater than $1 - 4\epsilon$,*

$$\begin{aligned} \|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_2 &\leq \|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_1 = \|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_\infty \\ &\leq C_1\lambda_0^2 d\|\mathbf{\Theta}^*\|_1\rho_*^{-1} + C_2\lambda_0 d\max\Theta^*_{kk}\rho_*^{-1} \end{aligned} \tag{3.8}$$

where C_1 and C_2 are constants depending on $\{A, a\}$ only.

Since the entries of $\mathbf{\Theta}^*$ are bounded by the maximum of the diagonal, the ℓ_1 matrix norm $\|\mathbf{\Theta}^*\|_1$ is of the same order as the matrix degree d. Thus, the inequality (3.8) provides a convergence rate of the order $d\lambda_0$ for either the ℓ_1 matrix norm or the spectrum norm under the conditions $d\{(\log p)/n\}^{1/2} \to 0$, $\rho_*^{-1} = O(1)$ and $\max(\Theta^*)_{kk} = O(1)$. The first condition is the main sparsity condition, and the other two are actually conditions on the ℓ_2 norm of the target matrix. To achieve the same convergence rate, Yuan (2010) and Cai et al. (2011) both imposed the condition $d\{(\log p)/n\}^{1/2} \to 0$ and the boundedness of the ℓ_1 norm of the unknown. We replace the ℓ_1 condition by the weaker boundedness of the spectrum norm of the unknown. The spectrum norm condition on the unknown is not only weaker, but also natural for the convergence in spectrum norm. The extra condition $\{d(\log p)/n\}^{1/2}\rho^*/\rho_* < a$ here is not strong. Under the conditions $d\{(\log p)/n\}^{1/2} \to 0$ and $\rho_*^{-1} = O(1)$, the extra condition only requires $\rho^*/d^{1/2}$ to be small and it allows ρ^* to diverge to infinity.

This sharper error bound in the spectrum norm is a consequence of using the scaled Lasso estimator (3.3). We prove a convergence rate of order $\lambda_0^2 d$ for the scaled Lasso estimation of the noise levels σ_j in Chapter 2. With this faster rate of convergence, the estimation error in the diagonal is no longer the main term and thus the condition of the bounded ℓ_1 norm of $\mathbf{\Theta}^*$ can be weakened.

The consistency of the scaled Lasso estimation for the noise level is based on the ℓ_1 error bound for the regression coefficients. Oracle inequalities for the ℓ_1 error of the Lasso have been studied with various conditions, including the restricted isometry condition (Candes & Tao, 2007), the compatibility condition (van de Geer, 2007) and the sign-restricted cone invertibility factor (Ye & Zhang, 2010) among others. Chapter 2 extends these oracle inequalities for the scaled Lasso. Here we use the version under

the condition of ℓ_1 sign-restricted cone invertibility factor (SCIF)

$$SCIF_1(\xi, S; \mathbf{\Sigma}) = \inf \left\{ \frac{|S| \cdot \|\mathbf{\Sigma} \boldsymbol{u}\|_\infty}{\|\boldsymbol{u}\|_1} : \boldsymbol{u} \in \mathscr{C}_-(\xi, S) \right\} > 0, \tag{3.9}$$

with the cone $\mathscr{C}(\xi, S) = \{\boldsymbol{u} \in R^{p-1} : \|\boldsymbol{u}_{S^c}\|_1 \leq \xi \|\boldsymbol{u}_S\|_1\}$ and the sign-restricted cone $\mathscr{C}_-(\xi, S) = \{\boldsymbol{u} \in \mathscr{C}(\xi, S) : u_j \mathbf{\Sigma}_{j,*} \boldsymbol{u} \leq 0, \forall j \notin S\}$. It is proved that, conditional on $\boldsymbol{X}_{*,-j}$,

$$\left| \frac{\widehat{\sigma}_j}{\sigma_j} - 1 \right| = O_p(1)|S_j|\lambda_0^2, \quad \|\widehat{\boldsymbol{\beta}}_{-j,j} - \boldsymbol{\beta}_{-j,j}\|_1 / \sigma_j = O_p(1)|S_j|\lambda_0, \tag{3.10}$$

under the condition that $SCIF_1(\xi, S_j; \overline{\mathbf{\Sigma}}_{-j})$ is bounded away from 0. This is guaranteed by the conditions of Theorem 3.2. The error bound of ℓ_1 matrix norm then follows from (3.10).

3.4 Simulation results

In this section, we compare the proposed matrix estimator based on scaled Lasso with graphical Lasso and CLIME (Cai et al., 2011). Three models are considered. The first two models are the same as model 1 and model 2 in Cai et al. (2011). Model 2 was also studied in Rothman et al. (2008).

- Model 1: $\Theta_{ij} = 0.6^{|i-j|}$.
- Model 2: Let $\mathbf{\Theta} = \boldsymbol{B} + \delta \boldsymbol{I}$, where each off-diagonal entry in $\boldsymbol{B}$ is generated independently and equals to 0.5 with probability 0.1 or 0 with probability 0.9. δ is chosen such that the condition number of $\mathbf{\Theta}^*$ is p. Finally, we rescale the matrix $\mathbf{\Theta}^*$ to the unit in diagonal.
- Model 3: The diagonal of the target matrix has unequal values. $\mathbf{\Theta} =$

$D^{1/2}\mathbf{\Omega}D^{1/2}$, where $\Omega_{ij} = 0.6^{|i-j|}$ and D is a diagonal matrix with diagonal elements $d_{ii} = (4i + p - 5)/\{5(p-1)\}$.

For each model, we generate a training sample of size 100 from a multivariate normal distribution with mean zero and covariance matrix $\mathbf{\Sigma} = \mathbf{\Theta}^{-1}$ and an independent sample of size 100 from the same distribution for validating the tuning parameter λ for the graphical Lasso and CLIME. The GLasso and CLIME estimators are computed based on training data with various λ's and we choose λ by minimizing likelihood loss $\{\text{trace}(\overline{\mathbf{\Sigma}}\widehat{\mathbf{\Theta}}) - \log\det(\widehat{\mathbf{\Theta}})\}$ on the validation sample. The proposed scaled Lasso estimator is computed based on the training sample alone with the penalty level $\lambda_0 = \{(\log p)/n\}^{1/2}$. Consider 6 different dimensions $p = 30, 60, 90, 150, 300, 1000$ and replicate 100 times for each case. The CLIME estimators for $p = 300$ and $p = 1000$ are not computed due to the computational costs.

Table 3.1 presents the mean and standard deviation of estimation errors based on 100 replications. The estimation error is measured by several matrix norms: spectrum norm, matrix ℓ_1 norm and Frobenius norm. We can see that scaled Lasso estimator, labelled as SLasso, outperforms the graphical Lasso (GLasso) in all cases, while it has a comparable peformance with the CLIME.

Table 3.1: Estimation errors under various matrix norms of scaled Lasso, GLasso and CLIME for three models.

Model 1									
	Spectrum norm			Matrix ℓ_1 norm			Frobenius norm		
p	SLasso	GLasso	CLIME	SLasso	GLasso	CLIME	SLasso	GLasso	CLIME
30	2.41(0.08)	2.49(0.14)	2.29(0.21)	2.93(0.11)	3.09(0.11)	2.92(0.17)	4.09(0.12)	4.24(0.26)	3.80(0.36)
60	2.61(0.05)	2.94(0.05)	2.68(0.10)	3.10(0.09)	3.55(0.07)	3.27(0.09)	6.16(0.10)	7.15(0.15)	6.32(0.28)
90	2.67(0.05)	3.07(0.03)	2.87(0.09)	3.19(0.08)	3.72(0.06)	3.42(0.07)	7.73(0.11)	9.25(0.12)	8.42(0.31)
150	2.74(0.04)	3.19(0.02)	3.05(0.04)	3.28(0.08)	3.88(0.06)	3.55(0.06)	10.22(0.13)	12.55(0.09)	11.68(0.20)
300	2.80(0.03)	3.29(0.01)	NA	3.38(0.07)	4.06(0.05)	NA	14.77(0.11)	18.44(0.09)	NA
1000	2.87(0.03)	3.39(0.00)	NA	3.52(0.06)	4.44(0.07)	NA	27.59(0.12)	35.11(0.06)	NA
Model 2									
	Spectrum norm			Matrix ℓ_1 norm			Frobenius norm		
p	SLasso	GLasso	CLIME	SLasso	GLasso	CLIME	SLasso	GLasso	CLIME
30	0.75(0.08)	0.82(0.07)	0.81(0.09)	1.32(0.16)	1.49(0.15)	1.45(0.18)	1.90(0.10)	1.84(0.09)	1.87(0.11)
60	1.07(0.05)	1.15(0.06)	1.19(0.08)	1.97(0.16)	2.21(0.12)	2.20(0.23)	3.31(0.08)	3.18(0.13)	3.42(0.09)
90	1.49(0.04)	1.54(0.05)	1.61(0.04)	2.63(0.16)	2.89(0.16)	2.90(0.17)	4.50(0.06)	4.40(0.11)	4.65(0.08)
150	1.98(0.03)	2.02(0.05)	2.06(0.03)	3.31(0.17)	3.60(0.15)	3.65(0.19)	6.02(0.05)	6.19(0.16)	6.33(0.08)
300	2.85(0.02)	2.89(0.02)	NA	4.50(0.14)	4.92(0.17)	NA	9.35(0.05)	9.79(0.05)	NA
1000	5.35(0.01)	5.52(0.01)	NA	7.30(0.21)	7.98(0.15)	NA	18.34(0.05)	20.81(0.02)	NA
Model 3									
	Spectrum norm			Matrix ℓ_1 norm			Frobenius norm		
p	SLasso	GLasso	CLIME	SLasso	GLasso	CLIME	SLasso	GLasso	CLIME
30	1.75(0.10)	2.08(0.10)	1.63(0.19)	2.24(0.14)	2.59(0.10)	2.17(0.20)	2.52(0.10)	2.91(0.16)	2.37(0.25)
60	2.09(0.08)	2.63(0.04)	2.10(0.10)	2.58(0.13)	3.10(0.05)	2.65(0.14)	3.81(0.09)	4.84(0.08)	3.98(0.13)
90	2.24(0.07)	2.84(0.03)	2.38(0.18)	2.72(0.12)	3.30(0.06)	2.91(0.12)	4.79(0.08)	6.25(0.08)	5.37(0.37)
150	2.40(0.06)	3.06(0.02)	2.76(0.05)	2.89(0.11)	3.45(0.04)	3.18(0.09)	6.35(0.09)	8.43(0.07)	7.75(0.08)
300	2.54(0.05)	3.26(0.01)	NA	3.05(0.10)	3.58(0.03)	NA	9.20(0.09)	12.41(0.04)	NA
1000	2.68(0.05)	3.47(0.01)	NA	3.26(0.09)	3.73(0.03)	NA	17.2(0.09)	23.55(0.02)	NA

3.5 Discussion

In the proof of the theoretical results for the proposed estimator, we use oracle inequalities for the estimation error associated with a linear model without normalizing the predictors. In the discussion section, we describe this aspect of our results. Consider a linear model as follows,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 I_n).$$

Let $\boldsymbol{\Sigma} = \boldsymbol{X}'\boldsymbol{X}/n$, $\boldsymbol{D} = \text{diag}\boldsymbol{\Sigma}$, $\widetilde{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{D}^{1/2}$ and $\widetilde{\boldsymbol{\Sigma}} = \boldsymbol{D}^{-1/2}\overline{\boldsymbol{\Sigma}}\boldsymbol{D}^{-1/2}$. In order to penalize the coefficients on the same scale, we use a weighted ℓ_1 norm of the coefficients as the penalty function. Consider the estimator

$$\{\widehat{\boldsymbol{\beta}}, \widehat{\sigma}\} = \arg\min_{\boldsymbol{b},\sigma} \Big\{ \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda_0 \|\boldsymbol{D}^{1/2}\boldsymbol{b}\|_1 \Big\}. \tag{3.11}$$

This is actually the scaled Lasso as we use in matrix estimation in Section 2. It is equivalent to the estimation based on normalized predictors:

$$\{\widehat{\boldsymbol{\alpha}}, \widehat{\sigma}\} = \arg\min_{\boldsymbol{a},\sigma} \Big\{ \frac{\|\boldsymbol{y} - \widetilde{\boldsymbol{X}}\boldsymbol{a}\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda_0 \|\boldsymbol{a}\|_1 \Big\} \tag{3.12}$$

with $\widehat{\boldsymbol{\beta}} = \boldsymbol{D}^{-1/2}\widehat{\boldsymbol{\alpha}}$.

The following theorem gives the oracle inequalities for the estimation of regression coefficients and noise level.

Theorem 3.3. *Let* $\{\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\sigma}\}$ *be as in (3.11) and (3.12),* $\sigma^* = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2/n^{1/2}$, $S = \{k : \beta_k \neq 0\}$, $z^* = \|\widetilde{\boldsymbol{X}}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/n\|_\infty/\sigma^*$ *and* $\xi > 1$.

(i)In the event $z^* \le (1+\tau^+)^{-1/2}\lambda_0(\xi-1)/(\xi+1)$,

$$\frac{1}{1+\tau^+} \le \Big(\frac{\widehat{\sigma}}{\sigma^*}\Big)^2 \le \frac{1}{1-\tau^-}, \quad \|\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}\|_1 \le \frac{(1+\xi)\tau^-\sigma^*}{2\xi\lambda_0(1-\tau^-)^{1/2}}, \tag{3.13}$$

$$\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1 \le \frac{(1+\xi)\tau^-\sigma^*}{2\xi\lambda_0(1-\tau^-)^{1/2}\min_k D_{kk}^{1/2}}, \tag{3.14}$$

where $\tau^- = \phi_1(\xi)\lambda_0^2|S|/SCIF_1(\xi,S;\widetilde{\boldsymbol{\Sigma}})$ *and* $\tau^+ = \phi_2(\xi)\lambda_0^2|S|/SCIF_1(\xi,S;\widetilde{\boldsymbol{\Sigma}})$ *with constants* $\xi>1$, $\phi_1(\xi)=4\xi^2/(1+\xi)^2$ *and* $\phi_2(\xi)=4\xi(\xi-1)/(1+\xi)^2$.

(ii)Let $\lambda_0 \ge \{(2/n)\log(p/\epsilon)\}^{1/2}(\xi+1)/\{(\xi-1)(1-\tau_-)\}$. *For* $n-2>\log(p/\epsilon)\to\infty$,

$$P\big\{z^* \le (1-\tau_-)\lambda_0(\xi-1)/(\xi+1)\big\} \ge 1-(1+o(1))\epsilon/\sqrt{\pi\log(p/\epsilon)}.$$

Theorem 3.3 is an immediate extension from the oracle inequalities for the scaled Lasso in Chapter 2. With an extra condition that $\boldsymbol{x}_k'\boldsymbol{x}_k/n (k=1,\ldots,p)$ are uniformly bounded from zero, the estimators have the same convergence rate as that for a regression model with normalized predictors. The error rates (3.10) follows from Theorem 3.3 and are used to prove the convergence rate of matrix estimation.

3.6 Proofs

In this section, we provide the proofs of Theorem 3.3 and Theorem 3.2. Theorem 3.1 is a brief version of Theorem 3.2, so we omit the proof.

Proof of Theorem 3.3. The inequalities (3.13) are parallel to (2.17). The only difference is that here we use the ℓ_1 bound under the condition of the sign-restricted cone invertibility factor (SCIF). Since $\widehat{\boldsymbol{\beta}} = \boldsymbol{D}^{-1/2}\widehat{\boldsymbol{\alpha}}$, (3.14) follows from the second

inequality in (3.13). □

Proof of Theorem 3.2. Let $\xi > (A+1)/(A-1)$, $(\sigma_j^*)^2 = \boldsymbol{\beta}_{*,j}'\overline{\boldsymbol{\Sigma}}\boldsymbol{\beta}_{*,j}$, $z_{(j),k} = \overline{\boldsymbol{\Sigma}}_{kk}^{-1/2}|\overline{\boldsymbol{\Sigma}}_{k,*}\boldsymbol{\beta}_{*,j}|/\sigma_j^*$ and $z_{(j)}^* = \max_{k\neq j} z_{(j),k}$. By Theorem 4, in the event $z_{(j)}^* \leq (1+\tau_{(j)}^+)^{-1/2}\lambda_0(\xi-1)/(\xi+1)$,

$$\frac{1}{1+\tau_{(j)}^+} \leq \Big(\frac{\widehat{\sigma}_j}{\sigma_j^*}\Big)^2 \leq \frac{1}{1-\tau_{(j)}^-}, \quad \sum_{k\neq j}\overline{\Sigma}_{kk}^{1/2}|\widehat{\beta}_{k,j}-\beta_{k,j}| \leq \frac{(1+\xi)\tau_{(j)}^-\sigma_j^*}{2\xi\lambda_0(1-\tau_{(j)}^-)^{1/2}}, \quad (3.15)$$

where

$$\tau_{(j)}^- = \phi_1(\xi)\lambda_0^2|S_j|/SCIF_1(\xi,S_j;\widetilde{\boldsymbol{\Sigma}}_{-j}), \quad \tau_{(j)}^+ = \phi_2(\xi)\lambda_0^2|S_j|/SCIF_1(\xi,S_j;\widetilde{\boldsymbol{\Sigma}}_{-j}).$$

We first derive some probabilistic bounds for some useful quantities. Since $\overline{\boldsymbol{\Sigma}} = \boldsymbol{X}'\boldsymbol{X}/n$ and the rows of $\boldsymbol{X}$ follow a multivariate normal distribution with covariance matrix $\boldsymbol{\Sigma}^*$, we have $\sigma_j^* = \|\boldsymbol{x}_j - \boldsymbol{X}_{-j}\boldsymbol{\beta}_{-j,j}\|/\sqrt{n}$ and $z_{(j),k} = \widetilde{\boldsymbol{x}}_k(\boldsymbol{x}_j - \boldsymbol{X}_{-j}\boldsymbol{\beta}_{-j,j})/\sigma_j^*$. Thus, $n(\sigma_j^*/\sigma_j)^2$ follows a χ^2 distribution with n degrees of freedom and thus

$$P\big\{\big|(\sigma_j^*/\sigma_j)^2 - 1\big| > \sqrt{(8/n)\log(2p/\epsilon)}\big\} \leq \epsilon/p. \quad (3.16)$$

Also, we have that $z_{(j),k}/\{(1-z_{(j),k}^2)/(n-1)\}^{1/2}$ follows a t-distribution with $n-1$ degrees of freedom. By Lemma 2.1 with $m = n-1$ and $t^2 = \log(p^2/\epsilon) > 2$,

$$P\big\{|z_{(j),k}| > \sqrt{2\log(p^2/\epsilon)/n}\big\} \;\leq\; (1+\epsilon_{n-1})(\epsilon/p^2)/\sqrt{\pi\log(p^2/\epsilon)}.$$

Thus,

$$P\Big\{ \max_j |z^*_{(j)}| > \sqrt{2\log(p^2/\epsilon)/n}\Big\} \;\; \leq \;\; \epsilon, \tag{3.17}$$

i.e. the events $z^*_{(j)} \leq (1+\tau^+_{(j)})^{-1/2}\lambda_0(\xi-1)/(\xi+1)(j=1,\ldots,p)$ occur with probability greater than $1-\epsilon$. Since $\overline{\Sigma}_{kk} \sim \Sigma^*_{kk}\chi^2_n/n$, we have

$$P\Big\{\big|\overline{\Sigma}_{kk}/\Sigma^*_{kk}-1\big| > \sqrt{(8/n)\log(2p/\epsilon)}\Big\} \leq \epsilon/p. \tag{3.18}$$

So there exists a small ζ, such that $\max|\overline{\Sigma}_{kk}/\Sigma^*_{kk}-1| < \zeta$ holds for all k with probability greater than $1-\epsilon$.

Now we need to bound $SCIF_1(\xi, S_j; \widetilde{\boldsymbol{\Sigma}}_{-j})$, for all j, with probability greater than $1-\epsilon$ under the given conditions, where $S_j = \{i \neq j : \beta_{i,j} \neq 0\}$. Let $\boldsymbol{Z} = \boldsymbol{X}(\text{diag}\boldsymbol{\Sigma}^*)^{-1/2}$. We discuss the bounds for $SCIF_1$ within the event $\max|\overline{\Sigma}_{kk}/\Sigma^*_{kk}-1| < \zeta$.

For $(|A|,|B|,\|\boldsymbol{u}\|,\|\boldsymbol{v}\|_r) = (\lceil a\rceil,\lceil b\rceil,1,1)$ with $A\cap B=\emptyset$, we define

$$\delta^{\pm}_a = \delta^{\pm}_a(\boldsymbol{X}) = \max_{A,\boldsymbol{u}}\Big\{ \pm\Big(\|\boldsymbol{X}'_A\boldsymbol{X}_A\boldsymbol{u}/n\| - 1\Big)\Big\},\; \theta^{(2)}_{a,b} = \theta^{(2)}_{a,b}(\boldsymbol{X}) = \max_{A,B,\boldsymbol{u},\boldsymbol{v}} \boldsymbol{v}'\boldsymbol{X}'_A\boldsymbol{X}_B\boldsymbol{u}/n.$$

For any subset $T\subset\{1,\ldots,p\}$, we have

$$\theta^{(2)}_{a,b} \geq \theta^{(2)}_{a,b}(\boldsymbol{X}_T), \quad \delta^{\pm}_a \geq \delta^{\pm}_a(\boldsymbol{X}_T), \quad \theta^{(2)}_{a,b} \leq (1+\delta^+_a)^{1/2}(1+\delta^+_b)^{1/2} \leq 1+\delta^+_{a\vee b}. \tag{3.19}$$

By Proposition 2(i) in Zhang & Huang (2008), we have

$$P\Big\{(1-c)^2\rho_* \le 1-\delta_m^-(\boldsymbol{Z}) \le 1+\delta_m^+(\boldsymbol{Z}) \le (1+c)^2\rho^*\Big\} \ge 1-\epsilon, \tag{3.20}$$

where $c = \sqrt{m/n} + \sqrt{(2m/n)\log(2p/\epsilon)}(1+o(1))$. We also have

$$\begin{aligned} 1+\delta_a^+(\widetilde{\boldsymbol{X}}) &\le \max(\Sigma^*_{kk}/\overline{\Sigma}_{kk})(1+\delta_a^+(\boldsymbol{Z})) = (1+\delta_a^+(\boldsymbol{Z}))/(1-\zeta), \\ 1-\delta_a^-(\widetilde{\boldsymbol{X}}) &\ge \min(\Sigma^*_{kk}/\overline{\Sigma}_{kk})(1-\delta_a^-(\boldsymbol{Z})) = (1-\delta_a^-(\boldsymbol{Z}))/(1+\zeta). \end{aligned} \tag{3.21}$$

Let $k_j = |S_j|$. It follows from the shifting inequality in Ye and Zhang (2010) with $\ell \ge d$ that

$$\begin{aligned} SCIF_1(\xi, S_j; \widetilde{\boldsymbol{\Sigma}}_{-j}) &\ge \frac{1}{1+\xi}\big(1-\delta^-_{k_j+\ell}(\widetilde{\boldsymbol{X}}_{-j}) - \xi\sqrt{\frac{k_j}{4\ell}}\theta^{(2)}_{4\ell,k_j+\ell}(\widetilde{\boldsymbol{X}}_{-j})\big) \\ &\ge \frac{1}{1+\xi}\big\{1-\delta^-_{4\ell}(\widetilde{\boldsymbol{X}}) - \xi\sqrt{\frac{d}{4\ell}}(1+\delta^+_{4\ell}(\widetilde{\boldsymbol{X}}))\big\} \\ &\ge \frac{1}{1+\xi}\big\{\frac{1-\delta^-_{4\ell}(\boldsymbol{Z})}{1+\zeta} - \xi\sqrt{\frac{d}{4\ell}}\frac{1+\delta^+_{4\ell}(\boldsymbol{Z})}{1-\zeta}\big\} \end{aligned}$$

The second and the third inequalities follow from (3.19) and (3.21), respectively. Let $m = 4\ell$ in (3.20) with $\ell = d(\xi\rho^*/\rho_*)^2 > d$. Then

$$SCIF_1(\xi, S_j; \widetilde{\boldsymbol{\Sigma}}_{-j}) \ge \frac{\rho_*}{1+\xi}\big\{\frac{(1-c)^2}{1+\zeta} - \frac{(1+c)^2}{2(1-\zeta)}\big\}.$$

Under the condition $(\rho^*/\rho_*)\{(d/n)\log p\}^{1/2} < a$ for a small fixed a, c is also very small. Thus, with probability greater than $1-\epsilon$, $SCIF_1(\xi, S_j; \widetilde{\boldsymbol{\Sigma}}_{-j})$ are bounded by

$C\rho_*$ for all j, where C is a constant only depending on $\{\xi, \zeta, a\}$.

Now we are ready to bound ℓ_1 of the column of $\widetilde{\mathbf{\Theta}} - \mathbf{\Theta}$ by (3.15), (3.16), (3.17), (3.18) and the uniform bound for $SCIF_1$. The following inequalities hold with probability greater than $1 - 4\epsilon$:

$$\begin{aligned}
\|\widetilde{\mathbf{\Theta}}_{\cdot j} - \mathbf{\Theta}^*_{\cdot j}\|_1 &\leq |\widetilde{\mathbf{\Theta}}_{jj} - \mathbf{\Theta}^*_{jj}| + \|\widetilde{\mathbf{\Theta}}_{-j,j} - \mathbf{\Theta}^*_{-j,j}\|_1 \\
&\leq \|\mathbf{\Theta}_{\cdot j}\|_1 \cdot \Big|\Big(\frac{\widehat{\sigma}_j}{\sigma_j}\Big)^{-2} - 1\Big| + \Big(\frac{\widehat{\sigma}_j}{\sigma^*_j}\Big)^{-2}\Big(\frac{\sigma^*_j}{\sigma_j}\Big)^{-1}\sigma_j^{-1}\frac{\|\widehat{\boldsymbol{\beta}}_{-j,j} - \boldsymbol{\beta}_{-j,j}\|_1}{\sigma^*_j} \\
&\leq \|\mathbf{\Theta}_{\cdot j}\|_1 \frac{C_1' \lambda_0^2 |S_j|}{\rho_*} + \frac{C_2' \lambda_0 |S_j|}{\sigma_j \big(\min_k \mathbf{\Sigma}^*_{kk}\big)^{1/2} \rho_*}.
\end{aligned}$$

The first two inequalities just use some simple algebra, while the last one put (3.15), (3.16), (3.17), (3.18) and the uniform bound for $SCIF_1$ together. The constants C_1' and C_2' only depend on $\{A, a\}$. Therefore, the ℓ_1 error of the matrix estimator $\widetilde{\mathbf{\Theta}}$ is bounded by

$$\|\widetilde{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_1 \leq C_3' \lambda_0^2 d \|\mathbf{\Theta}^*\|_1 \rho_*^{-1} + C_4' \lambda_0 d \max \Theta^*_{kk} \rho_*^{-1}.$$

Then the upper bound for $\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_1$ follows from the triangle inequality and the definition of $\widehat{\mathbf{\Theta}}$, since $\|\widehat{\mathbf{\Theta}} - \widetilde{\mathbf{\Theta}}\|_1 \leq \|\mathbf{\Theta}^* - \widetilde{\mathbf{\Theta}}\|_1$.

For any matrix $\boldsymbol{M}$ and vector $\boldsymbol{u}$, $\boldsymbol{v}$, we have

$$\boldsymbol{u}'\boldsymbol{M}\boldsymbol{v} = \sum_{i,j} M_{ij} u_i v_j \leq \big(\sum_{i,j} M_{ij} u_i^2 \sum_{i,j} M_{ij} v_j^2\big)^{1/2} \leq \big(\|\boldsymbol{M}\|_\infty \cdot \|\boldsymbol{M}\|_1\big)^{1/2} \|\boldsymbol{u}\| \cdot \|\boldsymbol{v}\|.$$

So $\|\boldsymbol{M}\|_2^2 \leq \|\boldsymbol{M}\|_\infty \cdot \|\boldsymbol{M}\|_1$. For the symmetric matrix $\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}$, we have

$$\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_2 \leq \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_\infty \leq \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_1.$$

The desired error bounds based the spectrum norm then follows. □

Chapter 4

Estimation and Statistical Inference for Partial Correlation

4.1 Introduction

Most of the recent advances in high-dimensional data have been focused on the estimation of high-dimensional objects as in Chapters 2 and 3. However, the estimation of low-dimensional functionals of high-dimensional parameters is also of great interest. For example, instead of the covariance matrix or its inverse as linear operators, one might be more interested in the relationship between individual pairs of variables. Chapter 3 provides a good estimator for precision matrix in terms of matrix norms, but it still remains unclear if this leads to a good estimator for partial correlations. In fact, a rate minimax estimator of a high-dimensional parameter does not automatically yield rate minimax estimates of its low-dimensional functionals.

In this chapter, we propose a method for estimating partial correlations of individual pairs of covariates, a feature of importance in a graphical model, with high-dimensional Gaussian data. In Chapter 2, the scaled Lasso is proposed to jointly estimate the coefficient vector and noise level (variance) in univariate linear regression. The variance of the noise is estimated by the sample variance of the estimated residuals, where we apply the scaled Lasso to estimate the linear effects from all covariates. We extend this method for partial correlation estimation. Our basic idea is to solve two sparse linear regression problems to remove linear effects of other covariates, and then compute the sample correlation as our estimator. More generally, this method can be

further extended to estimate any low-dimensional functional of conditional covariance matrix, by considering multivariate linear regression.

We study the asymptotic performance of the partial correlation estimator by taking advantage of the scale invariance and fast convergence rate of scaled Lasso. The asymptotic theory automatically justifies statistical inference for the partial correlation. These results only require a capped ℓ_1 sparsity and allow the target partial correlation matrix to have many elements of small and moderate magnitude. This nature make our scaled Lasso methodology very practical due to the tolerance of many small signals.

This chapter is organized as follows. In Section 4.2, we define an estimation problem of low-dimensional functionals of high-dimensional objects, and introduce our scale invariant method. In Section 4.3, we study asymptotic properties of the proposed estimator. Numerical results are presented in Section 4.4. Section 4.5 includes some final remarks.

We use the following notation throughout this chapter. For a vector $\boldsymbol{v} = (v_1, \ldots, v_p)$, $\|\boldsymbol{v}\|_q = (\sum_j |v_j|^q)^{1/q}$ is the ℓ_q norm with the special $\|\boldsymbol{v}\| = \|\boldsymbol{v}\|_2$ and the usual extensions $\|\boldsymbol{v}\|_\infty = \max_j |v_j|$ and $\|\boldsymbol{v}\|_0 = \#\{j : v_j \neq 0\}$. For matrices $\boldsymbol{M}$, $\boldsymbol{M}_{j,*}$ is the j-th column of $\boldsymbol{M}$, $\boldsymbol{M}_{A,B}$ represents the submatrix of $\boldsymbol{M}$ with rows in A and columns in B, Moreover, we denote the set $\{j\}$ by j and denote the set $\{1, \ldots, p\} \setminus \{j\}$ by $-j$ in the subscripts.

4.2 Estimation of low-dimensional functionals

In this section, we state a more general problem of low-dimensional functionals of high-dimensional objects, and propose a scale invariant estimation method via scaled Lasso. Our primary example is the partial correlation.

The partial correlation is of great interest in Gaussian-Markov graphical models. Let $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ be a $N(0, \boldsymbol{\Sigma})$ random vector. The partial correlation between

X_j and X_k, say r_{jk}, is their conditional correlation given all other variables. It can be also written as the error correlation in the linear regression of $\boldsymbol{X}_{\{j,k\}}$ against $\boldsymbol{X}_{\{j,k\}^c}$. In general, for any proper subset $A \subset \{1, \ldots, p\}$, a multivariate linear regression model can be written as

$$\boldsymbol{X}_A = \boldsymbol{X}_{A^c}\boldsymbol{\beta}_{A^c,A} + \boldsymbol{\varepsilon}_A. \tag{4.1}$$

For $A = \{j, k\}$, the partial correlation r_{jk} is the correlation between the two entries of $\boldsymbol{\varepsilon}_A$. It is well known that the conditional distribution of $\boldsymbol{X}_A$ given $\boldsymbol{X}_{A^c}$ follows the normal distribution

$$\boldsymbol{X}_A|\boldsymbol{X}_{A^c} \sim N(\boldsymbol{X}_{A^c}\boldsymbol{\Sigma}_{A^c,A^c}^{-1}\boldsymbol{\Sigma}_{A^c,A}, \boldsymbol{\Sigma}_A - \boldsymbol{\Sigma}_{A,A^c}\boldsymbol{\Sigma}_{A^c,A^c}^{-1}\boldsymbol{\Sigma}_{A^c,A}).$$

Thus, the coefficient matrix in (4.1) is $\boldsymbol{\beta}_{A^c,A} = \boldsymbol{\Sigma}_{A^c,A^c}^{-1}\boldsymbol{\Sigma}_{A^c,A}$ and the residual $\boldsymbol{\varepsilon}_A$ follows the multivariate normal distribution $N(0, \boldsymbol{\Sigma}_A - \boldsymbol{\Sigma}_{A,A^c}\boldsymbol{\Sigma}_{A^c,A^c}^{-1}\boldsymbol{\Sigma}_{A^c,A})$. Let $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ be the precision matrix. It follows easily from the block inversion formula that the covariance matrix for the residual $\boldsymbol{\varepsilon}_A$ is $\boldsymbol{\Theta}_A^{-1} = \boldsymbol{\Sigma}_A - \boldsymbol{\Sigma}_{A,A^c}\boldsymbol{\Sigma}_{A^c,A^c}^{-1}\boldsymbol{\Sigma}_{A^c,A}$. Thus,

$$r_{jk} = -\boldsymbol{\Theta}_{jk}/(\boldsymbol{\Theta}_{jj}\boldsymbol{\Theta}_{kk})^{1/2}. \tag{4.2}$$

Throughout the sequel, we consider the slightly more general problem of estimating a smooth function of $\boldsymbol{\Theta}_A^{-1}$, say $\tau = \tau(\boldsymbol{\Theta}_A^{-1})$, where set A is of bounded size.

For $|A| = 1$ and $\tau(s) = s$, our estimation target is $\tau(\boldsymbol{\Theta}_j^{-1}) = \boldsymbol{\Theta}_j^{-1} = \sigma_j^2$. This was done in Chapter 2, where the scaled Lasso is used to jointly estimate the coefficient vector and noise level in univariate linear regression. In the same spirit, we extend this method to $|A| > 1$. Let $\boldsymbol{X}_j \in \mathbb{R}^n$ be the j-th column of $\boldsymbol{X}$. For each $j \in A$, we apply

the scaled Lasso to the univariate linear regression of $\boldsymbol{X}_j$ against $\boldsymbol{X}_{A^c}$ as follows:

$$\{\widehat{\boldsymbol{\beta}}_{A^c,j},\ \widehat{\sigma}_j\} = \underset{\boldsymbol{b}_{A^c},\sigma}{\arg\min} \Big\{ \frac{\|\boldsymbol{X}_j - \boldsymbol{X}_{A^c}\boldsymbol{b}_{A^c}\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \sum_{\ell \in A^c} \|\boldsymbol{X}_\ell\|_2 |b_\ell| / \sqrt{n} \Big\}. \quad (4.3)$$

This is actually the scaled Lasso estimator (2.6) with normalized parameters $\|\boldsymbol{X}_\ell\|_2|b_\ell|/\sqrt{n}$. It is the same estimator as in (3.3), where it is determined by the sample covariance matrix. Here we express it in the form of linear regression as what we have in Chapter 2. Let $\widehat{\boldsymbol{\beta}}_{A^c,j}, j \in A$, be the columns of $\widehat{\boldsymbol{\beta}}_{A^c,A}$ and $\boldsymbol{z}_A = \boldsymbol{X}_A - \boldsymbol{X}_{A^c}\widehat{\boldsymbol{\beta}}_{A^c,A}$. In Chapter 2, the noise level estimator is the sample standard deviation of the approximate residuals as follows

$$\widehat{\Theta}_{jj}^{-1} = \widehat{\sigma}_j^2 = \boldsymbol{z}_j'\boldsymbol{z}_j/n.$$

For $|A| > 1$, since the covariance matrix for the residual vector $\boldsymbol{\varepsilon}_A$ is $\boldsymbol{\Theta}_A^{-1}$, this above estimator could be extended. We estimate $\boldsymbol{\Theta}_A^{-1}$ by the sample covariance matrix of the estimated residuals

$$\widehat{\boldsymbol{\Theta}}_A^{-1} = \boldsymbol{z}_A'\boldsymbol{z}_A/n,$$

and the smooth function $\tau = \tau(\boldsymbol{\Theta}_A^{-1})$ by a plug-in step

$$\widehat{\tau} = \tau(\boldsymbol{z}_A^\top \boldsymbol{z}_A/n). \quad (4.4)$$

Consider our primary example: the estimation of partial correlation of X_j and X_k. Let $A = \{j,k\}$ and $\{\boldsymbol{z}_j, \boldsymbol{z}_k\}$ be the estimated residuals of the bivariate linear regression. Since the partial correlation r_{jk} is the correlation between residuals, we

estimate it by the sample correlation

$$\widehat{r}_{jk} = \frac{\boldsymbol{z}_j^\top \boldsymbol{z}_k}{\|\boldsymbol{z}_j\|_2 \|\boldsymbol{z}_k\|_2}. \tag{4.5}$$

This is a special case of (4.4) with the function

$$\tau(\widehat{\boldsymbol{\Theta}}^{-1}_{\{j,k\}}) = S_{12}/(S_{11}S_{22})^{1/2}, \quad \text{where} \quad \widehat{\boldsymbol{\Theta}}^{-1}_{\{j,k\}} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}.$$

4.3 Theoretical properties

Suppose we have a data matrix $\boldsymbol{X} \in \mathrm{R}^{n\times p}$ with iid rows from $N(0, \boldsymbol{\Sigma})$. An oracle expert observing both $\boldsymbol{X}$ and $\boldsymbol{\varepsilon}_A = \boldsymbol{X}_A - \boldsymbol{X}_{A^c}\boldsymbol{\beta}_{A^c,A}$ may estimate τ by the oracle MLE

$$\tau^* = \tau(\boldsymbol{\varepsilon}_A^\top \boldsymbol{\varepsilon}_A/n) \tag{4.6}$$

due to the sufficiency of $\boldsymbol{\varepsilon}_A$ for $\boldsymbol{\Theta}_A$. In Chapter 2, the scaled Lasso estimator for the noise level has been proven to be within $o(n^{-1/2})$ of the oracle τ^* under certain "large-p-smaller-n" settings. The following theorem gives an error bound for the estimator $\widehat{\tau}$ in (4.4) by comparing it with the oracle MLE (4.6).

Theorem 4.1. *Suppose $\tau : \mathbb{R}^{A\times A} \to \mathbb{R}$ is a unit Lipschitz function in a neighborhood $\{M : \|M - \Theta_A^{-1}\|_2 \le \eta_0\}$. Let $\widehat{\tau}$ be given by (4.4) with $\lambda = \{3(\log p)/n\}^{1/2}$ in (4.3). Let $s_A = \max_{j\in A}\sum_{k\in A^c}\min(1, |\Theta_{jk}|/\lambda)$. Suppose that for a fixed M_0, $\|\Theta\|_2 + \|\Sigma\|_2 \le M_0$. Then, there exist constants $a_0 > 0$ and $C_0 < \infty$, both depending on $\{\eta_0, M_0\}$ only,*

such that for $s_A \leq a_0 n/\log p$*,*

$$\mathbb{P}\Big\{\big|\widehat{\tau} - \tau^*\big| > C_0 s_A(\log p)/n\Big\} \leq p^{-1/3}, \tag{4.7}$$

where τ^* *is the oracle MLE (4.6).*

Since the oracle MLE τ^* in (4.6) is based on an $|A|$-dimensional regular multivariate normal model $\varepsilon_A \sim N(0, \Theta_A^{-1})$ and $|A|$ is bounded, τ^* is efficient. This gives the efficiency of $\widehat{\tau}$.

Now consider the estimation of the partial correlation (4.2). With $A = \{j, k\}$ in (4.3), the scaled Lasso estimator is given by (4.5) and the oracle estimator is given by

$$r_{jk}^* = \boldsymbol{\varepsilon}_j^\top \boldsymbol{\varepsilon}_k / (\|\boldsymbol{\varepsilon}_j\|_2 \|\boldsymbol{\varepsilon}_k\|_2). \tag{4.8}$$

The following corollary gives the sufficient conditions for the asymptotic normality of the partial correlation estimator.

Corollary 4.1. *Let* r_{jk}^* *and* $\widehat{r}_{jk}$ *be given by (4.8) and (4.5). Suppose the conditions of Theorem 4.1 hold with* $A = \{j, k\}$ *and* $s = s_A$*. Then,*

$$\widehat{r}_{jk} - r_{jk}^* = O_P(s(\log p)/n). \tag{4.9}$$

Consequently, if $s(\log p)/\sqrt{n} \to 0$*, then*

$$\sqrt{n}(\widehat{r}_{jk} - r_{jk})/(1 - \widehat{r}_{jk}^2) \xrightarrow{D} N(0, 1).$$

Since r_{jk}^* is the (oracle) MLE of the correlation based on iid bivariate normal

observations, $\sqrt{n}(r^*_{jk} - r_{jk})$ converges to $N(0, (1 - r^2_{jk})^2)$ in distribution. Thus, Corollary 4.1 directly follows from Theorem 4.1.

A major difference between our theory and existing work based on variable selection is that Θ is allowed to have many elements of small and moderate magnitude in Theorem 4.1 and Corollary 4.1. This is similar to Zhang & Zhang (2011) where statistical inference of regression coefficients is considered.

4.4 Simulation results

We present some simulation results to demonstrate the performance of the scaled Lasso for partial correlation. Two examples are considered. The first example is a five-diagonal precision matrix with $\Theta_{jj} = 1$, $\Theta_{j-1,j} = \Theta_{j,j-1} = 0.6$, and $\Theta_{j-2,j} = \Theta_{j,j-2} = 0.1$. In the second example, we set $\Theta_{jk} = 0.6^{|j-k|}$ (no entry of the precision matrix is exactly zero). The partial correlations are computed by $r_{jk} = -\Theta_{jk}/(\Theta_{jj}\Theta_{kk})^{1/2}$. We generate a random sample of size $n = 100$ from $N(0, \Sigma)$ with $\Sigma = \Theta^{-1}$. The scaled Lasso estimator is computed with $\lambda = \{(\log p)/n\}^{1/2}$. For each example, we consider $p = 200$ and $p = 1000$.

Table 4.1: Mean and standard error of the scaled Lasso estimator for the partial correlation and the ratio of the simulated and theoretical MSEs, $\kappa = \text{MSE}/\{(1 - r^2_{jk})^2/n\}$.

Example 1: five-diagonal precision matrix						
	$r_{12} = -0.6$		$r_{13} = -0.1$		$r_{14} = 0$	
p	Mean$\pm$SE($\widehat{r}$)	κ	Mean$\pm$SE($\widehat{r}$)	κ	Mean$\pm$SE($\widehat{r}$)	κ
200	-0.626 $\pm$ 0.055	0.894	-0.042 $\pm$ 0.083	1.037	-0.010 $\pm$ 0.097	0.937
1000	-0.643 $\pm$ 0.056	1.214	-0.043 $\pm$ 0.088	1.104	-0.007 $\pm$ 0.089	0.797
Example 2: exponential decay precision matrix						
	$r_{12} = -0.6$		$r_{13} = -0.36$		$r_{14} = -0.216$	
p	Mean$\pm$SE($\widehat{r}$)	κ	Mean$\pm$SE($\widehat{r}$)	κ	Mean$\pm$SE($\widehat{r}$)	κ
200	-0.551 $\pm$ 0.064	1.602	-0.236 $\pm$ 0.079	2.846	-0.042 $\pm$ 0.100	4.412
1000	-0.539 $\pm$ 0.079	2.425	-0.224 $\pm$ 0.089	3.475	-0.029 $\pm$ 0.101	4.962

Table 4.1 shows the scaled Lasso estimates for r_{12}, r_{13}, and r_{14} based on 100 replications. In Example 1, $\widehat{r}_{jk}$ is quite accurate, as the condition of small $s_A(\log p)/\sqrt{n}$ holds well with values 0.8, 1.3, and 1.5 for the estimation of r_{12}, r_{13}, and r_{14} when $p = 200$, and with values 1.0, 1.6, and 1.9 when $p = 1000$. In Example 2, the scaled Lasso deteriorates as the condition $s_A(\log p)/\sqrt{n}$ starts to fail, with values 2.3, 2.8, and 3.4 for the estimation of r_{12}, r_{13}, and r_{14} when $p = 200$, and with values 2.8, 3.5, and 4.2 when $p = 1000$.

In addition, we would like to point out that another difficulty for this simulation study is the relatively small sample size $n = 100$. The distribution of r^* here tends to the normality very slowly, as this normality is usually applied for very large sample size in practice, say at least 500.

4.5 Proofs

In this section, we provide the proof of Theorem 4.1.

Proof of Theorem 4.1. For the simplicity, we denote $\boldsymbol{\beta}_{A^c,j}$ by $\boldsymbol{\beta}_j$ in this proof. Let $\sigma_j^* = \|\boldsymbol{X}_j - \boldsymbol{X}_{A^c}\boldsymbol{\beta}_j\|/\sqrt{n}$ for any $j \in A$. By the proof of Theorem 3.2, we have the following statement for any $j \in A$: in the event E_j

$$\max_{\ell \in A^c} \|\boldsymbol{X}_\ell\|_2^{-1} |\boldsymbol{X}_\ell'(\boldsymbol{X}_j - \boldsymbol{X}_{A^c}\boldsymbol{\beta}_j)|/(\sqrt{n}\sigma_j^*) \le \lambda, \tag{4.10}$$

it holds that

$$\left|\frac{\widehat{\sigma}_j}{\sigma_j^*} - 1\right| \le C_1\lambda^2 s_A, \quad \|\widehat{\boldsymbol{\beta}}_{A^c,j} - \boldsymbol{\beta}_{A^c,j}\|_1/\sigma_j^* \le C_2\lambda s_A, \tag{4.11}$$

where C_1 and C_2 are constants depending on M_0 only. Moreover, the event (4.10) holds with probability at least $1 - p^{-1/2}$.

Thus, for any $j, k \in A$, we may compare $\boldsymbol{z}_j^\top \boldsymbol{z}_k/n$ with $\boldsymbol{\varepsilon}_j^\top \boldsymbol{\varepsilon}_k/n$ as follows

$$\begin{aligned}
\boldsymbol{z}_j^\top \boldsymbol{z}_k/n - \boldsymbol{\varepsilon}_j^\top \boldsymbol{\varepsilon}_k/n &= \big\{\boldsymbol{\varepsilon}_j + \boldsymbol{X}_{A^c}(\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j)\big\}'\big\{\boldsymbol{\varepsilon}_k + \boldsymbol{X}_{A^c}(\boldsymbol{\beta}_k - \widehat{\boldsymbol{\beta}}_k)\big\}/n - \boldsymbol{\varepsilon}_j^\top \boldsymbol{\varepsilon}_k/n \\
&\leq \|\boldsymbol{X}'_{A^c}\boldsymbol{\varepsilon}_j/n\|_\infty \|\boldsymbol{\beta}_k - \widehat{\boldsymbol{\beta}}_k\|_1 + \|\boldsymbol{X}'_{A^c}\boldsymbol{\varepsilon}_k/n\|_\infty \|\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j\|_1 \\
&\quad + \|\boldsymbol{X}_{A^c}(\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j)\| \cdot \|\boldsymbol{X}_{A^c}(\boldsymbol{\beta}_k - \widehat{\boldsymbol{\beta}}_k)\|/n.
\end{aligned}$$

In the event $\cap_{j\in A} E_j$, it holds that

$$\boldsymbol{z}_j^\top \boldsymbol{z}_k/n - \boldsymbol{\varepsilon}_j^\top \boldsymbol{\varepsilon}_k/n \leq C_3 \sigma_j^* \sigma_k^* \lambda^2 s_A.$$

This and the property of Lipschitz function lead to (4.7).

□

Chapter 5

Statistical Methods for Real-time Blood Glucose Monitoring

5.1 Introduction

Diabetes mellitus, one of the leading causes of death in the world, is a metabolic disorder affecting the way that the human body uses digested food for growth and energy. Type 1 diabetes (T1D) is a form of diabetes mellitus that results from the loss of the insulin-producing beta cells of the islets of Langerhans in the pancreas, leading to insulin deficiency. It is also known as juvenile diabetes or insulin-dependent diabetes mellitus (IDDM). Nowadays injectable insulin is a widely-used treatment for this non-insulin producing type 1 diabetes. It is of great importance to control the insulin dose and thus the blood glucose level. A lack of insulin will result in the rise of blood glucose levels, while overdose of insulin injection will cause low blood sugar levels, or hypoglycemia, which may suddenly lead to ketoacidotic coma, an extremely dangerous situation. In order to avoid these circumstances, it is usually recommended to monitor the blood sugar levels continuously, especially before and after meals, and count the carbohydrate content of meals and snacks. However, the majority of diabetic patients fail to achieve the target glycated hemoglobin level (HbA1C) recommended by the Diabetes Control and Complications Trial (DCCT-Research-Group, 1994). For the effectiveness and convenience, people expect to develop a closed-loop artificial pancreas system without the requirement for patient intervention and action. Three major design elements are required for such system: 1) an insulin pump to accurately

deliver variable amounts of insulin, 2) a real-time continuous glucose monitor (CGM) to accurately determine ambient glucose levels, and 3) an effective algorithm to regulate insulin delivery rates based on real-time CGM outputs. This will be a truly transformational change in the treatment of patients with type 1 Diabetes.

However, the accuracy of the CGM is a major obstacle, mainly due to the difficulties with the requirement of periodical recalibration of the device with individual patients. In fact, the current generation of real-time CGM devices is also far from optimal when used in an open-loop system, in which patients use the CGM device as a reference of their blood glucose level. For example, the recent Juvenile Diabetes Research Foundation randomized clinical trial (JDRF-CGM-Study-Group, 2008) showed that lowering of HbA1C levels 0.5% could be achieved in T1D patients who were 25 years of age or older, but no improvement was observed in patients who were 8-24 years of age, and the CGM failed to lower the risk of severe hypoglycemia, regardless of age.

The continuous glucose monitor measures the glucose level via an electrochemical glucose biosensor. It is a single electrode, coated with an enzyme, inserted into subcutaneous fat tissue, called interstitial space. When a glucose molecule in interstitial space comes in contact with the electrode, a current is generated in the sensor (Wang, 2008). By measuring this current, one obtains a surrogate for blood glucose density. The problem is to convert this current measurement (ISIG) into an accurate measure of blood glucose density (BG). Due to degradation by biofouling and other issues, the relationship between ISIG and BG is changing from time to time. The variability between sensors is also observed. Moreover, BG is measured by the finger stick (FS), which is medically regarded as the ground truth, but still has an random error. All these concerns present a statistical challenge.

In this chapter, we propose and implement statistical methods to produce more accurate and precise estimates for continuous glucose levels. Motivated by the

mechanism of glucose sensor and continuous blood glucose monitor, we design a statistical framework for modeling the dynamic relationship between the blood glucose level and interstitial signal. At the current stage, our Bayes model also incorporates the time series aspects of the data and the variability depending on sensor age. The methods have been tested and evaluated with an important large dataset, called "Star I", from Medtronic, Inc., including 137 subjects using blood glucose sensors for six months on average. The analysis shows that our blood glucose prediction outperforms the current measurement in the CGM device. This provides a possibility of upgrading the current continuous glucose monitor.

This chapter is organized as follows. In Section 5.2, we propose a general framework for continuous blood glucose estimation. In Section 5.3, we describe some methods for implementing nonparametric models under proper assumptions. Section 5.4 provides a statistical analysis of the STAR 1 dataset and compares our continuous blood glucose prediction with some existing measurements.

5.2 Modeling the Continuous Glucose Levels

As stated in the introduction part, the continuous glucose monitor tracks the glucose level via an electrochemical biosensor. In this section, we develop statistical methods for the continuous glucose level estimation according to two primary sources of information: previous finger stick results FS and interstitial signals ISIG. The current interstitial signal is also necessary to estimate the current glucose level.

The electrical current $\mathrm{ISIG}(t)$ is correlated with the glucose density in interstitial fluid near the sensor site at time t, which we denote by $\mathrm{IG}(t)$. Empirical and theoretical evidence suggests that there is an approximately linear relationship between $\mathrm{ISIG}(t)$

and $\mathrm{IG}(t)$. Our basic model relating $\mathrm{ISIG}(t)$ and $\mathrm{IG}(t)$ is

$$\mathrm{ISIG}(t) = \alpha(t)\mathrm{IG}(t), \tag{5.1}$$

where $\alpha(t)$ is a slowly varying stochastic process. $\mathrm{IG}(t)$ could be used as a surrogate for our primary target, blood glucose density $\mathrm{BG}(t)$. We may rewrite our model

$$\mathrm{BG}(t) = \beta(t)\mathrm{ISIG}(t), \tag{5.2}$$

where $\beta(t)$ is a new stochastic process. Although we never observe the true value of BG, one may use fingerstick measurements of blood glucose density $\mathrm{FS}(t)$ instead. Thus, our key model is

$$\mathrm{FS}(t) = \mathrm{BG}(t) + \epsilon(t) = \beta(t)\mathrm{ISIG}(t) + \epsilon(t), \tag{5.3}$$

where $\epsilon(t)$ is an error term with mean zero.

There is some debate about whether an intercept term should be included in (5.1). Also, a time lag between BG and IG has been discussed in the literature, because it takes time to diffuse blood glucose molecules into interstitial fluid. This suggests an extra term of derivative $\mathrm{IG}'(t)$ or $\mathrm{ISIG}'(t)$ may be added in (5.2). For the convenience, we do not include these possible terms in the following model description, while all the methods discussed in this section could be extended to models with extra terms.

More specifically, suppose fingerstick measurements are taken at time $t_k, k = 1, \ldots$ and the error terms $\epsilon(t_k)$ are iid normally distributed with standard deviation σ. Then the fingerstick measurements are related to ISIG values as follows:

$$\mathrm{FS}_{t_k} \Big| \mathscr{F}_{t_k} \sim N\Big(\beta(t_k)\mathrm{ISIG}(t_k), \sigma^2\Big), \tag{5.4}$$

where $\mathscr{F}_t = \sigma(\{ISIG(t_k);\ t_k \leq t\},\ \{FS(t_k); t_k < t\})$ is the σ-field generated by all the data up to time t, including $\text{ISIG}(t)$ but not $\text{FS}(t)$. The Bayes estimator of $\widehat{\text{BG}}(t)$ is

$$\widehat{\text{BG}}(t) = E\Big[\beta(t)\Big|\mathscr{F}_t\Big]\text{ISIG}(t). \tag{5.5}$$

Now the only thing that remains unclear is how to model the process $\beta(t)$. Due to the biofouling on the sensor and environmental changes in human body, the linear relation is unstable. Thus, we attempt to build up a dynamic model for $\beta(t)$ to address this change. Consider a single sensor with lifespan $[0, t^*)$. In order to incorporate sensor information into our estimation procedures, we consider a nonstationary discrete Markov process $\beta(t)$ with a constant transition intensity λ and independently generated new states according to distributions depending on sensor age. Such a process can be written as

$$\beta(t) = \beta(S_j), \quad S_j \leq t < S_{j+1},$$

where $S_1 < S_2 < \cdots$ are interarrival times of a Poisson process in $[0, \infty)$ with intensity λ, and at time $S_j = a$, $\beta(a)$ is generated according to a distribution $g(\cdot|a)$, independent of $\{\beta(t) : t \leq a\}$. Figure 5.1 provides an illustration of the discrete Markov model. The solid line indicates the transition times of the process $\beta(t)$ and the expected value of $g(\cdot)$ on each state. Our task is to estimate this expected value for each a, given the observed ratios FS/ISIG.

The transition probability is characterized as follows

$$p_{s,t}(b|b_0) = e^{-\lambda(t-s)}I\{b = b_0\} + \int_s^t g(b|a)\mathrm{d}e^{-\lambda(t-a)}.$$

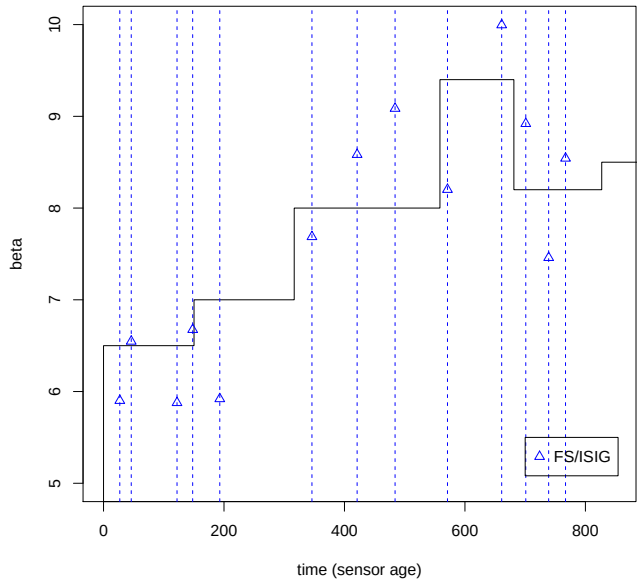

Figure 5.1: An illustration of the discrete Markov model. Triangles: the observed ratio of FS/ISIG; solid line: the expected value of BG/ISIG.

The first term is corresponding to the case of no transition from time s to time t, while the second term concerns that the last transition between (s,t) happens at time a for all a. Let $\pi(b|t^-)$ be the conditional probability mass function of $b(t)$ given $\mathscr{F}_t$. Due to the given transition probability, we have

$$\pi(b|t^-) = e^{-\lambda(t-t_k)}\pi(b|t_k) + \int_{t_k}^{t} g(b|a)\mathrm{d}e^{-\lambda(t-a)}, \tag{5.6}$$

for all $t \in (t_k, t_{k+1}]$. Since the data in (t_k, t_{k+1}) contains no information about the process $b(t)$, we do not update the density function for $t \in (t_k, t_{k+1})$, i.e. $\pi(b|t) = \pi(b|t^-)$. For $t = t_{k+1}$, the posterior can be updated by the new FS information at time τ_{k+1}. The updating rule is as follows:

$$\pi(b|t_{k+1}) \propto \pi(b|t_{k+1}^-)\exp\big[-\{\mathrm{FS}_{t_{k+1}} - \mathrm{ISIG}(t_{k+1})b\}^2/(2\sigma^2)\big]. \tag{5.7}$$

To sum up, the relation between FS and ISIG is characterized by (5.4), (5.6) and (5.7). The sequences of FS and ISIG are used to estimate the distribution of $\beta(t)$, leading to the blood glucose estimation via (5.5).

5.3 Nonparametric Bayes Methods

In the above formulation, the distribution $g(\cdot|a)$ remains unknown. Suppose $G(\cdot|a)$ are discrete distributions with a common support $B = \{b_i\}$ and that $G(\cdot|a) = G(\cdot|a_j)$ for $a_j \leq a < a_{j+1}, j \leq j^*$, $a_0 = 0$ and $a_{j^*+1} = \infty$. The density function $G(\cdot|a)$ is characterized by a matrix with elements $g_{ij} = G(\{b_i\}|a_j)$, $\sum_i g_{ij} = 1$. Specifically, the process $\{\beta(t), t \geq 0\}$ has the initial distribution $P\{\beta(0) = b_i\} = g_{i0}$ and transition probability is written in a discrete form as follows

$$p_{s,t}(b_i|x) = e^{-\lambda(t-s)} I\{b_i = x\} + \sum_{a_{j+1} > s} \left\{ e^{-\lambda(t-a_{j+1})_+} - e^{-\lambda(t-a_j \vee s)_+} \right\} g_{ij}.$$

The computation of the posterior still requires an estimation of g_{ij}. In this section, we introduce two methods for implementing our Markov model.

5.3.1 Implementation with MCMC methods

We may use Markov chain Monte Carlo (MCMC) methods to estimate the densities $\{g_{ij}\}$. Our algorithm chooses an equilibrium for $\{g_{ij}\}$ by iteratively estimating the blood glucose by the discrete Markov model and updating $\{g_{ij}\}$.

- Step 1. We estimate $\widehat{BG}$ based the given $\{g_{ij}\}$.

  The conditional probability mass function of $\beta(t)$ given $\mathscr{F}_t$ and the posterior

probability given new FS information are

$$\pi(b_i|t_k^-) = e^{-\lambda(t-t_k)}\pi(b_i|t_k) + \sum_{a_{j+1}>t_k} \{e^{-\lambda(t-a_{j+1})_+} - e^{-\lambda(t-a_j\vee t_k)_+}\}g_{ij},$$
$$\pi(b_i|t_k) = \frac{\pi(b_i|t_k^-)\exp\big(-\{\text{FS}_{t_k} - b_i\text{ISIG}(t_k)\}^2/(2\sigma^2)\big)}{\sum_j \pi(b_j|t_k^-)\exp\big(-\{\text{FS}_{t_k} - b_j\text{ISIG}(t_k)\}^2/(2\sigma^2)\big)}.$$

Then $\widehat{BG}(t) = \sum_i b_i\pi(b_i|t)\text{ISIG}(t)$.

- Step 2. We update the density function $\{g_{ij}\}$ given FS and ISIG.

  Let $X(t_k) = \{\text{FS}_{t_\ell}, \text{ISIG}(t_\ell), \ell \le k\}$. We begin with the last finger-stick time t_m for this sensor. Let τ be the time of the last transition of $\beta(t)$ before t_m,

$$\tau = \max\Big\{t < t_m : t = 0 \text{ or } \beta(t) \neq \beta(t_m)\Big\}.$$

  Given $X(t_m)$ and $\tau \in (t_k, t_{k+1}]\cap[a_j, a_{j+1})$, $\beta(t_m)$ has the conditional probability mass function

$$p_{j,k}(b_i) = C_{j,k}^{-1} g_{ij} \prod_{\ell=k+1}^{m} \Big\{\sigma^{-2}\exp\big(-\{\text{FS}_{t_\ell} - b_i\text{ISIG}(t_\ell)\}^2/(2\sigma^2)\big)\Big\}$$

  with $C_{j,k} = \sum_{b_i\in B} g_{ij}\prod_{\ell=k+1}^{m}\sigma^{-2}\exp\big(-\{\text{FS}_{t_\ell} - b_i\text{ISIG}(t_\ell)\}^2/(2\sigma^2)\big)$. Let $\mathcal{C}_1, \ldots, \mathcal{C}_{j^*}$ be the collections of density functions for various sensor ages. Now we need to generate the transition time to determine which category this density function belongs to.

  1. Generate (J, K) with $P\{(J, K) = (j, k)\} = q_{j,k}$, where

$$\begin{aligned} q_{j,k} &= P\Big\{\tau \in (t_k, t_{k+1}] \cap [a_j, a_{j+1})\Big|X(t_m)\Big\} \\ &\propto C_{j,k}\int_{(t_k,t_{k+1}]\cap[a_j,a_{j+1})} e^{-\lambda t}\lambda dt. \end{aligned}$$

2. Add $p_{j,k}(\cdot)$ to collection $\mathcal{C}_j$;

3. Let $m = K$.

We run this procedure until $K = 0$ or $t_K = 0$. The training set can be repeatedly used until each category reaches a certain size. The empirical distribution from $\mathcal{C}_j$ gives a new estimate of $\{g_{ij}, b_i \in B\}$. A weighted average of the new and old estimators can be used to generate the estimators

The MCMC methods always requires many iterations to compute equilibrium distributions, although we may choose short burning periods after having a warm initial. Therefore, the computational cost of this method is very expensive.

5.3.2 Implementation with an empirical method

For the computational simplicity, we may implement our Bayes model in an approximate way. Assume that the last transition time is at time t. Then the approximate density function is

$$\pi(b|t^-) = e^{-\lambda(t-\tau_k)}\pi(b|\tau_k) + (1 - e^{-\lambda(t-\tau_k)})g(b|t).$$

There is also an alternative explanation for this prior density function. It is a linear combination of two sources of prior information: the recent body environment $\pi(b|\tau_k)$ and the aging of sensor $g(b|t)$. The weights are defined adaptively, depending on the time since the last FS. When the last fingerstick measurement is taken very recently, the prior distribution will more rely on the posterior information based on that fingerstick test $\pi(b|\tau_k)$; vice versa.

The coefficient distribution $g(\cdot|a)$ is estimated by an empirical distribution as follows

$$g(\cdot|a) \approx \text{Average}(\pi(\cdot|a)), \tag{5.8}$$

where $\pi(b|t)$ are the posteriors for earlier sensors at the same age. In practice, we treat the age of sensor as a discrete variable in units of days and record the average of posteriors for all possible ages. When the posterior is updated due to a new FS at time t_k, we add this posterior to the corresponding collection of density functions and compute a new approximation for $g(\cdot|a)$ by (5.8).

5.4 Analysis of the STAR 1 Dataset

In this section, we describe the STAR 1 dataset and present our numerical results for continuous blood glucose level estimation. The estimation are evaluated from several aspects, including overall estimation accuracy, the detection of hypoglycemia and hyperglycemia, etc.

5.4.1 Descriptive analysis

In the STAR 1 study, 137 subjects using blood glucose sensors were monitored for periods of time spanning 6 days to 948 days (mean 188 days, SD 148). For each patient in the study, an electrical current measurement ISIG (in nano amps, nA) from the blood glucose sensor was recorded every 5 minutes with limited exceptions. Less frequently – approximately every 6 hours, on average – patients in the study recorded a more accurate measure of blood glucose density, obtained via fingerstick FS. Fingerstick measurements are considered as the ground truth of the blood glucose levels in this chapter. Each individual in the study regularly replaced their blood glucose biosensors, on average every 2.72 days. These measurements are entered into the CGM system and are essential for calibrating CGM algorithms. The dataset also includes the blood glucose level estimation by the CGM device, which provides a benchmark for estimation accuracy.

Figure 5.2 shows a typical dataset for a representative subject in a period of about

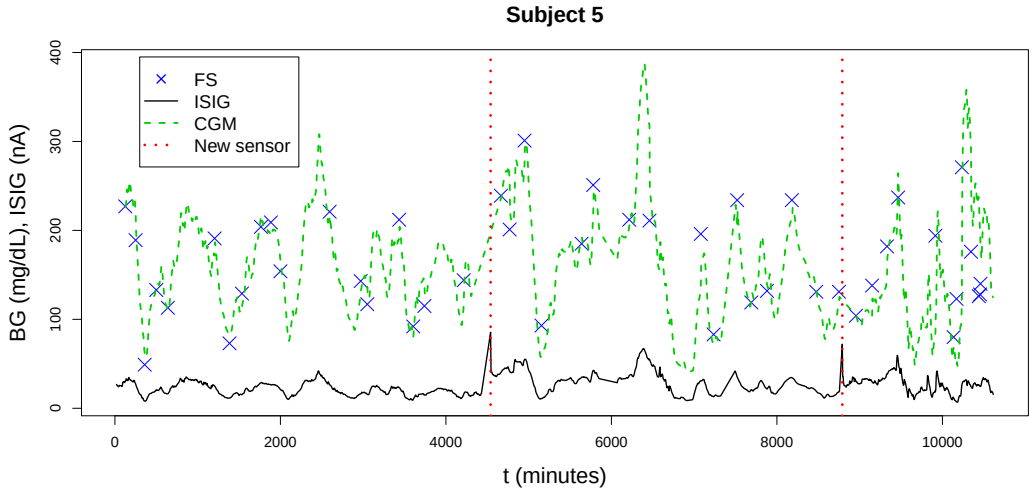

Figure 5.2: FS(t), ISIG(t), CGM(t), and sensor replacement times for Subject 5 in Star 1 dataset.

one week. The measurements FS(t), ISIG(t), and CGM(t) are plotted along with indicators for when sensors were replaced.

In Section 5.2, we have discussed the linear relationship (5.2) between the blood glucose levels and the interstitial current signals (ISIG). Figure 5.3 supports this aspect of our statistical framework.

When modeling the relation between the blood glucose level and the interstitial signal, one expects that the sensitivity of the sensor decreases over time (due to biofouling and other causes) and, hence, that $\beta(S_j + t)$ is larger that $\beta(S_j)$, for $t > 0$. This is confirmed by Figure 5.4, where the ratios FS/ISIG are plotted for different sensor ages. Note that the ratio tends to increase with sensor age.

5.4.2 Continuous blood glucose estimation

We implement our nonparametric methods described in Sections 5.2 and 5.3 and study its estimation performances from various aspects. Three methods are going to be compared in this section:

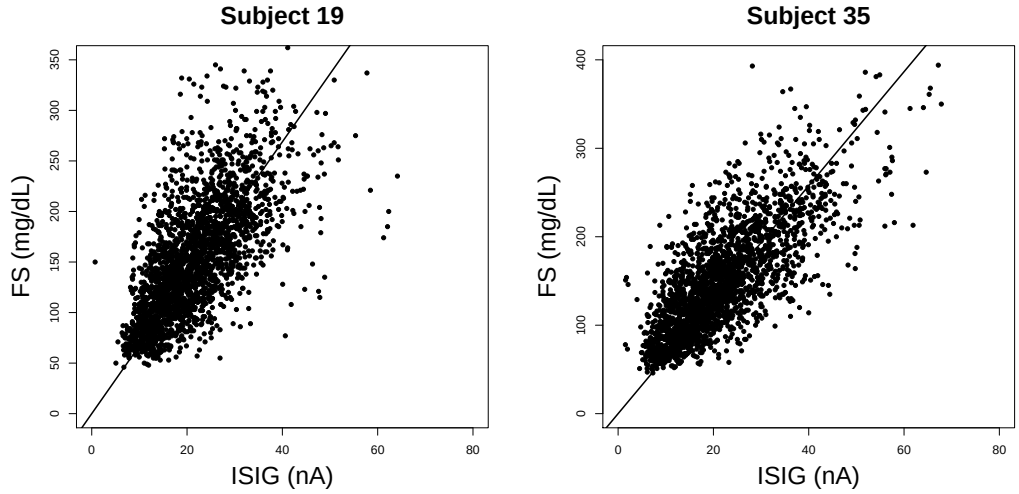

Figure 5.3: FS vs. ISIG for two subjects in the STAR 1 dataset.

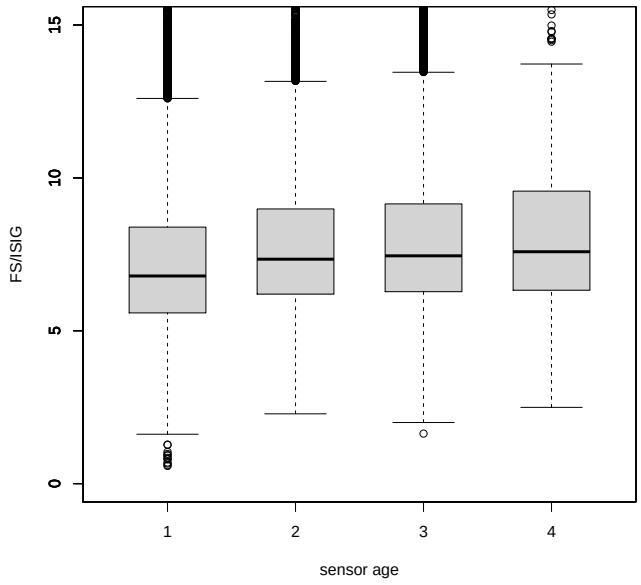

Figure 5.4: The boxplots of the ratios of FS/ISIG by sensor age. 1: Sensor less than 1 day old. 2: Sensor between 1 and 2 days old. 3: Sensor between 2 and 3 days old. 4: Sensor at least 3 days old.

1. The nonparametric Bayes method with the implementation in Section 5.3.2.

2. The Kalman filter, also known as dynamic linear model, from Dicker et al. (2012).

3. The original measurement in the CGM device.

The overall performance of each estimator was measured by its mean absolute relative difference,

$$\text{MARD}(\widehat{\text{BG}}) = \frac{1}{\#\{t\}} \sum_t \left\{ \frac{|\widehat{\text{BG}}(t) - \text{FS}(t)|}{\text{FS}(t)} \right\}, \tag{5.9}$$

where the sum on the right-hand side above is taken over all FS times t_k. MARD is widely used as the fundamental metric for comparing the performance of these methods. The first 50 patients in the Star 1 dataset were used for training the tuning parameters for these methods, while the remaining 87 patients were used for validation.

For the nonparametric method, we still need to specify some necessary (tuning) parameters as follows.

- The noise level σ for fingerstick measurement is first estimated by a ballpark interval $[10, 25]$ according to Brunner et al. (1998), which provides the percentage of measurements within a defined range of the reference values according to different glycemic ranges for various blood glucose meters.

- The tuning parameter e^λ is a weight for the effect of recent performance and thus between $(0, 1)$.

- Initial distributions $g(\cdot|a)$ should be given for each age category.

Various parameters are tested on the training dataset and the optimal values are selected in terms of mean absolute relative difference (MARD) that will be defined later. We took $\sigma = 20$, $e^\lambda = 0.6$ and let initial distributions $g(\cdot|a)$ be uniform $\mathcal{U}[0, 14]$ for each

any age. In practice, it is noticed that the choice of initial distributions does not affect the performance and the results are pretty stable when $e^{\lambda} \in [0.5, 0.7]$ and $\sigma \in [20, 25]$.

In Table 5.1, we report the estimation performances of three methods in terms of several loss functions: mean, standard deviation and median of absolute relative difference. For nonparametric Bayes method and Kalman filter, we also report their improvements over the original CGM and the number of subjects for which the subject-level MARD of the specified method is smaller than MARD(CGM). It is concluded that both methods outperform the original CGM.

Table 5.1: Summary statistics for analysis of Star 1 dataset.

		MARD (SD)	MedARD	ΔMARD	N_{MARD} (N)
Train.	NP-Bayes	0.1539 (0.1515)	0.1153	0.0111	47 (50)
	Kalman filter	0.1538 (0.1504)	0.1152	0.0112	46 (50)
	CGM	0.1650 (0.1675)	0.1225		
Valid.	NP-Bayes	0.1552 (0.1705)	0.1118	0.0087	77 (87)
	Kalman filter	0.1536 (0.1653)	0.1105	0.0103	79 (87)
	CGM	0.1639 (0.1831)	0.1168		

MARD, mean absolute relative difference; MedARD, median absolute relative difference; ΔMARD, the difference between the MARD of the indicated method and $\text{MARD}(\text{CGM})$; N_{MARD}, the number of subjects for which the subject-level MARD of the specified method is smaller than MARD(CGM); N, the total number of subjects.

5.4.3 More results on the estimation performances

In addition to the overall accuracy of blood glucose level estimation, we are interested in other performance measurements: the estimation accuracy in different ranges, the reliability of detecting hypoglycemia and hyperglycemia, etc. All the performance results are based on the test dataset of 87 patients.

Figure 5.5 shows the estimation accuracy for different FS values in terms of mean

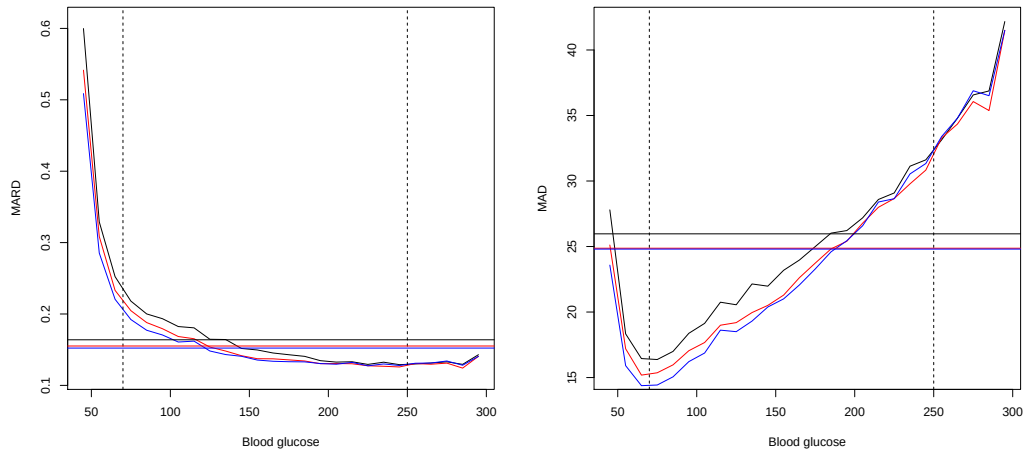

Figure 5.5: Mean absolute (relative) differences. CGM (black); NP-Bayes (red); Kalman filter (blue). Horizontal lines: overall MARD (MAD). Vertical lines: Thresholds for hypoglycemia and hyperglycemia.

absolute relative difference (5.9) and mean absolute difference

$$\mathrm{MAD}(\widehat{\mathrm{BG}}) = \frac{1}{\#\{t\}} \sum_{t} \left\{ |\widehat{\mathrm{BG}}(t) - \mathrm{FS}(t)| \right\}, \tag{5.10}$$

where the sum on the right-hand side above is taken over all FS times t_k. We can see that the absolute relative difference is large in the low range of blood glucose levels, while the absolute difference is large in the high range of blood glucose levels.

Another criteria for for CGM algorithms is to reliably detect hypoglycemia (low blood glucose density) and hyperglycemia (high blood glucose density). Threshold rules are a simple class of rules for detecting hypoglycemia or hyperglycemia based on an estimate. Following Bode et al. (2004), any timepoint with FS less than 70 mg/dL is defined to be a hypoglycemic period and any timepoint with FS greater than 250 mg/dL is defined to be a hyperglycemic period. We may use a strict threshold $\widehat{\mathrm{BG}} \leq 70$ to detect hypoglycemia or increase the threshold $\widehat{\mathrm{BG}} \leq 90$ to reduce the risk of missing hypoglycemic episodes. Table 5.2 provides sensitivity, specificity, positive predictive

value and negative predictive value for hypoglycemia. For a threshold $\widehat{\text{BG}} \leq S$, these measurements are defined as follows

- Sensitivity (SENS): $P(\widehat{\text{BG}} < S | \text{FS} < 70)$
- Specificity (SPEC): $P(\widehat{\text{BG}} \geq S | \text{FS} \geq 70)$
- Positive predictive value (PPV): $P(\text{FS} < 70 | \widehat{\text{BG}} < S)$
- Negative predictive value (NPV): $P(\text{FS} \geq 70 | \widehat{\text{BG}} \geq S)$

Table 5.2: Hypoglycemia detection with different threshold rules.

Threshold	Method	SENS	SPEC	PPV	NPV
≤ 70	NP-Bayes	0.4431	0.9730	0.5048	0.9657
	Kalman filter	0.4788	0.9693	0.4922	0.9677
	CGM	0.4273	0.9696	0.4662	0.9646
≤ 90	NP-Bayes	0.8550	0.9007	0.3484	0.9901
	Kalman filter	0.8772	0.8913	0.3338	0.9915
	CGM	0.8388	0.8981	0.3384	0.9890

SENS, sensitivity; SPEC, specificity; PPV, positive predictive value; NPV, negative predictive value.

It is noticed that the sensitivity of hypoglycemia with $S = 70$ is always below 0.5. It may be very dangerous if we fail to detect the hypoglycemia. Thus, it is necessary to have a higher threshold. However, the higher threshold reduces the positive predictive value, which means patients may be bothered by many false warnings. We can see from Table 5.2 that both of the nonparametric Bayes method and the Kalman filter outperform the original CGM in the sensitivity and positive predictive value. The Kalman filter is better in terms of sensitivity, while the nonparametric Bayes method is better in terms of positive predictive value. Moreover, the ROC curve for the nonparametric Bayes and Kalman filtering methods appear to dominate that for CGM across the entire plotted range in the left of Figure 5.6.

As for the hyperglycemia detection, a strict threshold is $\widehat{\text{BG}} \geq 250$ and a relaxed threshold is $\widehat{\text{BG}} \geq 220$ for lowering the risk of missing hyperglycemic episodes. Table

5.3 provides sensitivity, specificity, positive predictive value and negative predictive value for hyperglycemia that are defined with a threshold $\widehat{\text{BG}} \geq S$ as follows

- Sensitivity (SENS): $P(\widehat{\text{BG}} > S | \text{FS} > 250)$
- Specificity (SPEC): $P(\widehat{\text{BG}} \leq S | \text{FS} \leq 250)$
- Positive predictive value (PPV): $P(\text{FS} > 250 | \widehat{\text{BG}} > S)$
- Negative predictive value (NPV): $P(\text{FS} \leq 250 | \widehat{\text{BG}} \leq S)$

Table 5.3 suggests that the original CGM performs the best in terms of the sensitivity for the hyperglycemia detection, followed by the nonparametric Bayes method and then the Kalman filter. The ROC curve for three methods are roughly mixed together.

Table 5.3: Hyperglycemia detection with different threshold rules.

Threshold	Method	SENS	SPEC	PPV	NPV
≥ 250	NP-Bayes	0.6315	0.9641	0.7819	0.9277
	Kalman filter	0.6145	0.9699	0.8065	0.9250
	CGM	0.6374	0.9617	0.7724	0.9286
≥ 220	NP-Bayes	0.8405	0.8921	0.6137	0.9648
	Kalman filter	0.8213	0.9013	0.6293	0.9612
	CGM	0.8432	0.8876	0.6047	0.9652

SENS, sensitivity; SPEC, specificity; PPV, positive predictive value; NPV, negative predictive value.

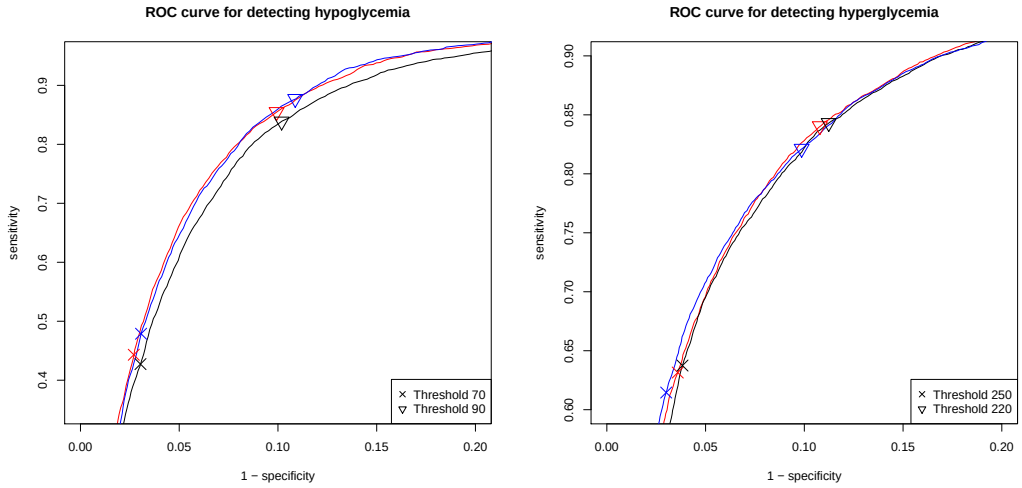

Figure 5.6: Partial ROC curves of detecting hypoglycemia and hyperglycemia for the validation data. CGM (black); NP-Bayes (red); Kalman filter (blue).

Bibliography

ANTONIADIS, A. (2010). Comments on: ℓ_1-penalization for mixture regression models by N. Städler, P. Bühlmann and S. van de Geer. *Test* **19**, 257–258.

BANERJEE, O., EL GHAOUI, L. & D'ASPREMONT, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research* **9**, 485–516.

BELLONI, A., CHERNOZHUKOV, V. & WANG, L. (2011). Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* **98**, 791–806.

BICKEL, P., RITOV, Y. & TSYBAKOV, A. (2009). Simultaneous analysis of lasso and Dantzig selector. *Annals of Statistics* **37**, 1705–1732.

BODE, B., GROSS, K., RIKALO, N., SCHWARTZ, S., WAHL, T., PAGE, C., GROSS, T. & MASTROTOTARO, J. (2004). Alarms based on real-time sensor glucose values alert patients to hypo-and hyperglycemia: the guardian continuous monitoring system. *Diabetes Technology and Therapeutics* **6**, 105–113.

BRUNNER, G., ELLMERER, M., SENDLHOFER, G., WUTTE, A., TRAJANOSKI, Z., SCHAUPP, L., QUEHENBERGER, F., WACH, P., KREJS, G. & PIEBER, T. (1998). Validation of home blood glucose meters with respect to clinical and analytical approaches. *Diabetes Care* **21**, 585–590.

BUNEA, F., TSYBAKOV, A. & WEGKAMP, M. H. (2007). Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics* **1**, 169–194.

CAI, T., LIU, W. & LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106**, 594–607.

CANDES, E. & TAO, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *Annals of Statistics* **35**, 2313–2404.

DCCT-RESEARCH-GROUP (1994). The effects of intensive diabetes treatment on the development and progression of long-term complications in adolescents with insulin-dependent diabetes mellitus: the diabetes control and complications trial. *J. Pediatr.* **124**, 177–188.

DICKER, L., SUN, T., ZHANG, C.-H., KEENAN, D. B. & SHEPP, L. (2012). Continuous blood glucose monitoring: a Bayes-hidden markov approach .

EFRON, B., HASTIE, T., JOHNSTONE, I. & TIBSHIRANI, R. (2004). Least angle regression (with discussion). *Annals of Statistics* **32**, 407–499.

FAN, J., GUO, S. & HAO, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of Royal Statistical Society* **74**, 37–65.

FAN, J. & LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

FAN, J. & PENG, H. (2004). On non-concave penalized likelihood with diverging number of parameters. *Annals of Statistics* **32**, 928–961.

FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441.

GREENSHTEIN, E. (2006). Best subset selection, persistence in high-dimensional statistical learning and optimization under ℓ_1 constraint. *Annals of Statistics* **34**, 2367–2386.

GREENSHTEIN, E. & RITOV, Y. (2004). Persistence in high–dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10**, 971–988.

HUANG, J., MA, S. & ZHANG, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica* **18**, 1603–1618.

HUBER, P. J. & RONCHETTI, E. M. (2009). *Robust Statistics*. Wiley, 2nd ed., pp. 172–175.

JDRF-CGM-STUDY-GROUP (2008). Continuous glucose monitoring and intensive treatment of type 1 diabetes. *N. Eng. J. Med.* **359**, 1464–1476.

KOLTCHINSKII, V., LOUNICI, K. & TSYBAKOV, A. B. (2011). Nuclear norm penalization and optimal rates for noisy low rank matrix completion. *Annals of Statistics* **39**, 2302–2329.

LAM, C. & FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrices estimation. *Annals of Statistics* **37**, 4254–4278.

MEINSHAUSEN, N. & BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Annals of Statistics* **34**, 1436–1462.

MEINSHAUSEN, N. & BÜHLMANN, P. (2010). Stability selection (with discussion). *Journal of the Royal Statistical Society: Series B* **72**, 417–473.

MEINSHAUSEN, N. & YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics* **37**, 246–270.

NEGAHBAN, S., RAVIKUMAR, P., WAINWRIGHT, M. J. & YU, B. (2009). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *In Advances in Neural Information Processing Systems (NIPS)* **22**.

OSBORNE, M., PRESNELL, B. & TURLACH, B. (2000a). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis* **20**, 389–404.

OSBORNE, M., PRESNELL, B. & TURLACH, B. (2000b). On the lasso and its dual. *Journal of Computational and Graphical Statistics* **9**, 319–337.

RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. & YU, B. (2008). Model selection in gaussian graphical models: High-dimensional consistency of l1-regularized mle. *In Advances in Neural Information Processing Systems (NIPS)* **21**.

ROTHMAN, A., BICKEL, P., LEVINA, E. & ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2**, 494–515.

SCHEETZ, T. E., KIM, K.-Y. A., SWIDERSKI, R. E., PHILP, A. R., BRAUN, T. A., KNUDTSON, K. L., DORRANCE, A. M., DIBONA, G. F., HUANG, J., CASAVANT, T. L., SHEFFIELD, V. C. & STONE, E. M. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proc. Nat. Acad. Sci* **103**, 14429–14434.

STÄDLER, N., BÜHLMANN, P. & VAN DE GEER, S. (2010). ℓ_1-penalization for mixture regression models (with discussion). *Test* **19**, 209–285.

SUN, T. & ZHANG, C.-H. (2010). Comments on: ℓ_1-penalization for mixture regression models by N. Städler, P. Bühlmann and S. van de Geer. *Test* **19**, 270–275.

VAN DE GEER, S. (2007). The deterministic lasso. Tech. Rep. 140, ETH Zurich, Switzerland.

VAN DE GEER, S. (2008). High–dimensional generalized linear models and the lasso. *Annals of Statistics* **36**, 614–645.

VAN DE GEER, S. & BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics* **3**, 1360–1392.

WANG, J. (2008). Electrochemical glucose biosensors. *Chem. Rev.* **108**, 814–825.

YANG, S. & KOLACZYK, E. D. (2010). Target detection via network filtering. *IEEE Transactions on Information Theory* **56**, 2502–2515.

YE, F. & ZHANG, C.-H. (2010). Rate minimaxity of the lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *Journal of Machine Learning Research* **11**, 3481–3502.

YUAN, M. (2010). Sparse inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research* **11**, 2261–2286.

YUAN, M. & LIN, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika* **94**, 19–35.

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* **38**, 894–942.

ZHANG, C.-H. & HUANG, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics* **36**, 1567–1594.

ZHANG, C.-H. & ZHANG, S. (2011). Confidence intervals for low-dimensional parameters with high-dimensional data .

ZHANG, T. (2009). Some sharp performance bounds for least squares regression with L_1 regularization. *Ann. Statist.* **37**, 2109–2144.

ZHAO, P. & YU, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research* **7**, 2541–2567.

Vita

Tingni Sun

2012 Ph.D. in Statistics, Rutgers, The State University of New Jersey, New Brunswick, New Jersey, USA.

2007 B.S. in Mathematics and Applied Mathematics, Peking University, Beijing, China.