Lasso-Type Sparse Regression and High-Dimensional Gaussian Graphical Models

by

Xiaohui Chen

M.Sc., The University of British Columbia, 2008
B.Sc., Zhejiang University, 2006

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate Studies

(Electrical and Computer Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2012

Abstract

High-dimensional datasets, where the number of measured variables is larger than the sample size, are not uncommon in modern real-world applications such as functional Magnetic Resonance Imaging (fMRI) data. Conventional statistical signal processing tools and mathematical models could fail at handling those datasets. Therefore, developing statistically valid models and computationally efficient algorithms for high-dimensional situations are of great importance in tackling practical and scientific problems. This thesis mainly focuses on the following two issues: (1) recovery of sparse regression coefficients in linear systems; (2) estimation of high-dimensional covariance matrix and its inverse matrix, both subject to additional random noise.

In the first part, we focus on the Lasso-type sparse linear regression. We propose two improved versions of the Lasso estimator when the signal-to-noise ratio is low: (i) to leverage adaptive robust loss functions; (ii) to adopt a fully Bayesian modeling framework. In solution (i), we propose a robust Lasso with convex combined loss function and study its asymptotic behaviors. We further extend the asymptotic analysis to the Huberized Lasso, which is shown to be consistent even if the noise distribution is Cauchy. In solution (ii), we propose a fully Bayesian Lasso by unifying discrete prior on model size and continuous prior on regression coefficients in a single modeling framework. Since the proposed Bayesian Lasso has variable model sizes, we propose a reversible-jump MCMC algorithm to obtain its numeric estimates.

In the second part, we focus on the estimation of large covariance and precision matrices. In high-dimensional situations, the sample covariance is an inconsistent estimator. To address this concern, regularized estimation is needed. For the covariance matrix estimation, we propose a shrinkage-to-tapering estimator and show that it has attractive theoretic properties for estimating general and large covariance matrices. For the precision matrix estimation, we propose a computationally efficient algorithm that is based on the thresholding operator and Neumann series expansion. We prove that, the proposed estimator is consistent in several senses under the spectral norm. Moreover, we show that the proposed estimator is minimax in a class of precision matrices that are approximately inversely closed.

Preface

This thesis is written based on a collection of manuscripts, resulting from collaboration between several researchers. Chapter 2 is based on a journal paper published in *IEEE Transactions on Information Theory* [31], co-authored with Prof. Z. Jane Wang and Prof. Martin J. McKeown. Chapter 3 is based on a journal paper published in *Signal Processing* [32], co-authored with Prof. Z. Jane Wang and Prof. Martin J. McKeown. Chapter 4 is based on a submitted journal paper co-authored with Prof. Z. Jane Wang and Prof. Martin J. McKeown. Chapter 5 is based on a journal paper to appear in *IEEE Transactions on Signal Processing* [30], co-authored with Prof. Young-Heon, Kim and Prof. Z. Jane Wang. Chapter 6 is based on a conference paper appears in *2010 International Conference in Image Processing* [35], co-authored with Prof. Z. Jane Wang and Prof. Martin J. McKeown.

The research outline was designed jointly by the author, Prof. Z. Jane Wang and Prof. Martin J. McKeown. The majority of the research, including literature survey, model design, theorem proofs, numerical simulation, statistical data analysis and results report, was conducted by the author, with suggestions from Prof. Z. Jane Wang and Prof. Martin J. McKeown. The manuscripts were primarily drafted by the author, with helpful comments from Prof. Z. Jane Wang and Prof. Martin J. McKeown.

The biomedical application description in Section 5.6, Chapter 5, is written based on a grant proposal by Prof. Martin J. McKeown.

Table of Contents

Appendix

List of Tables

List of Figures

List of Acronyms

AIC Akaike Information Criterion

BIC Bayesian Information Criterion

BG Binomial-Gaussian model

CLIME Constrained ℓ_1-Minimization for Inverse Matrix Estimation

CMT Covariance Matrix Taper

CV Cross-Validation

DBN Dynamic Bayesian Networks

FBM Fractional Brownian Motion

fMRI functional Magnetic Resonance Imaging

iid independent and identically distributed

LAR(S) Least Angle Regression

Lasso Least Absolute Shrinkage and Selection Operator

LS Least Squares

MAP Maximum A Posterior estimator

mAR(r) multivariate Autoregressive model with order r

MCC Mathews Correlation Coefficient

MH Metropolis-Hastings algorithm/ratio

MLE Maximum Likelihood Estimator

MM Minorizatoin-Maximization algorithm

(M/R)MSE (Minimum/Root) Mean-Squared Error

OAS Oracle Approximating Shrinkage

PCA Principle Component Analysis

PD Parkinson's Disease

RIC Risk Inflation Criterion

RJ-MCMC Reversible-Jump Markov chain Monte Carlo algorithm

RLAD Regularized Least Absolute Deviation

ROI Region of Interest

RW Random Walk

SCAD Smoothly Clipped Absolute Deviation

SEM Structural Equation Modeling

SNR Signal-to-Noise Ratio

SPICE Sparse Permutation Invariant Covariance Estimator

STO Shrinkage-to-Tapering Oracle estimator

STOA Shrinkage-to-Tapering Oracle Approximating algorithm

TNR True Negative Rate

TPR True Positive Rate

Acknowledgements

First and foremost, I owe innumerable thanks to my PhD advisers, Prof. Z.Jane Wang and Prof. Martin J. McKeown, for being great mentors, both professionally and personally. This thesis would never be possible without their continuous support over the years. Many of their valuable and insightful suggestions not only encouraged me to constantly learn new things, but also taught me how to be an independent researcher. I am in particular indebted to them for generously allowing me with enough freedom for exploring new research topics of my own interests.

I would like also to express thanks to my wonderful collaborators, co-authors, and fellow graduate students at UBC. In particular, I would like to thank Prof. Young-Heon, Kim (Dept. Mathematics) for many stimulating discussions and personal encouragement.

The research work in this thesis has been supported by a Pacific Alzheimer's Research Foundation (PARF) Centre Grant Award and a Graduate Fellowship from UBC.

Chapter 1

Introduction

1.1 Challenges of High-Dimensional Modeling: An Overview

Statistical estimation in high-dimensional situations, where the number of measured variables p is substantially larger than the sample size n (a.k.a. *large-p-small-n*), is fundamentally different from the estimation problems in the classical settings where we have *small-p-large-n*. Since high-dimensional datasets are not uncommon in modern real-world applications, such as gene expression microarray data and functional Magnetic Resonance Imaging (fMRI) data, precise estimation of high-dimensional models is of great importance in tackling such practical and scientific problems. Generally speaking, learning salient information from relatively a few samples when many more variables are present is not possible without knowing special structures in the data.

To alleviate the ill-posed problem, it is natural to restrict our attention to subsets of all solutions with certain special structures or properties and meanwhile to incorporate the *regularization* ideas into estimation. *Sparsity* is one commonly hypothesized condition and it seems to be realistic for many real-world applications. There has been a surge in the literature, termed *compressed sensing* in signal processing literature and *Lasso* in statistical literature, on the recovery of sparse signals in under-determined linear systems [23, 24, 26–28, 42, 43, 115]. Many beautiful results on sparse representation, recovery conditions and algorithms have been reported in the literature. We remark that the literature on this topic is too extensive for us to give an exhaustive list. My PhD thesis mainly focuses on the following two issues: (1) the recovery of sparse regression coefficients in linear systems; (2) estimation of high-dimensional covariance matrix and its inverse matrix (a.k.a precision matrix, or the Gaussian graphical model in machine learning language), both subject to random noise. It is emphasized that, in my PhD work, these two problems are studied from both theoretic and algorithmic perspectives.

Although significant progress has been made on sparsity in high-dimensionality during the last decade, there are still a number of challenges attracting intensive research activities in the statistics, machine learning, and signal processing communities. These include:

C1 Minimal sampling: *What are the fundamental information-theoretic limits of the sample size in order to obtain theoretically guaranteed correct and stable estimates?*

C2 Computational tractability: *Can we develop concrete algorithms that are computationally feasible or even efficient for the seemingly daunting large-scale combinatorial problems in terms of computational cost?*

C3 Robustness: *How can we make the feature selection tools adaptive to data and protective against non-Gaussian noise?*

C4 Consistency (e.g. estimation and model selection consistency under random noise): *Can we guarantee the proposed algorithms and models work appropriately in theory, at least asymptotically?*

C5 Optimality: *Is it possible to improve the proposed models in terms of convergence rate?*

C1 has been relatively well-studied in the literature and it has close connections to the approximation theory. As mentioned earlier, a large volume of compressed sensing papers have made beautiful solutions to this [19, 23–26, 28, 41–43, 117].

For **C2**, different high-dimensional estimation problems may have different problem features and we will see that convex relaxation and certain simple matrix operations often achieve computational efficiency. **C3** is a practical concern since essentially all literature, with only a few exceptions [75, 107, 123], considers robust estimation procedures under the assumption of error distributions with heavy-tails. **C4** and **C5** together offer us an assurance to use the models from a theoretic perspective.

While there are many potential practical applications for this method, a motivating application of this work is brain effective connectivity modeling using fMRI data, where the goal is to infer the connectivity network between a large number of (spatially-fixed) brain regions-of-interests (ROIs). Studying brain connectivity is crucial in understanding brain functioning and can provide significant insight into the pathophysiology of neurological disorders such as Parkinson's disease. Based on prior neuroscience knowledge, the connections between brain regions generally can be considered *a priori* to form a sparse network. Several linear regression based formalisms have been popular for inferring brain connectivity using fMRI and we have recently employed the unified structural equation modeling (SEM) and multivariate autoregressive (mAR) models to capture both spatial and temporal brain connections. Moreover, it is well-known that fMRI data is typically very noisy. Therefore we formulate brain connectivity modeling as a problem of sparse linear regression under large variance noise [34].

1.2 Lasso-Type Sparse Linear Regression

1.2.1 Prior Arts

Recovering sparse signals from high-dimensional data sets has been attracting intensive research attention during the last decade. By sparse signals we mean that the underlying model generating all measured quantities can be approximated by a few numbers of the true signals and the approximation errors are due to (random) noise.

In this part of the thesis, we consider the problem of estimating the coefficient vector in a linear regression model, defined as

$$\mathbf{y} = X\boldsymbol{\beta} + \mathbf{e}, \tag{1.1}$$

where the random measurement error vector $\mathbf{e} = (e_1, \cdots, e_n)^*$ is assumed to be independent and identically distributed (iid) with zero mean and a constant finite second moment σ^2 for each component. We regard $\mathbf{e}$ as a column vector, and use $\mathbf{e}^*$ to denote its conjugate transpose. Here, X is the $n \times p$ design matrix which can either be non-stochastic or random. As usual, rows of X represent the p-dimensional observations and columns of X represent the predictors. $\mathbf{y}_{n\times 1}$ is the response vector and $\boldsymbol{\beta}_{p\times 1}$ is the coefficient vector to be estimated. We are interested in the setup where p is independent of n and fixed, but can be a large positive integer.

There are many variable selection models proposed in the literature from both frequentist and Bayesian perspectives. Nevertheless, we are interested in sparsity-promoting linear regression models based on the *least absolute shrinkage and selection operator*, i.e. the Lasso [115], because of its popularity and its attractive computational and theoretical properties [13, 16, 45, 75, 77, 86, 90, 94–96, 100, 101, 121, 128, 130, 132]. So far in the statistics community, the Lasso is probably the most popular variable selection tool used to estimate a sparse coefficient vector. In signal processing literature, the minimization of the ℓ_1 norm regularized linear model is often termed the *basis pursuit* [28]. Specifically, the Lasso estimator of Tibshirani [115] is defined as the following:

Definition 1.2.1. The Lasso estimator is defined as

$$\hat{\boldsymbol{\beta}}_n = \arg\min_{\mathbf{u}\in\mathbb{R}^p} \left(\frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^* \mathbf{u})^2 + \frac{\lambda_n}{n} \sum_{j=1}^{p} |u_j|^\gamma \right), \tag{1.2}$$

where $\mathbf{x}_i$ means the i^{th}-row of X and $\gamma = 1$.

Here, $\lambda_n \geq 0$ is a shrinkage tuning parameter. A larger λ_n yields a sparser linear sub-model whereas a smaller λ_n corresponds to a less-sparse one. In extreme cases, $\lambda_n = 0$ gives the unregularized model and $\lambda_n = \infty$ produces the null model consisting of no predictor. More generally, for $\gamma > 0$, (1.2) is called the *bridge regression estimator* by [55], and $\gamma = 2$ yields the ridge regression [68]. It is clear that (1.2) is convex for $\gamma \geq 1$, and that it can produce sparsity when $0 < \gamma \leq 1$, since the penalized objective function has a non-trivial mass at zero. Therefore, the Lasso can be viewed as a sparsity-promoting convexification of the ℓ_2 loss plus the ℓ_0 penalty so that standard convex optimization technologies can be applied to efficiently solve (1.2) for the Lasso [42, 117]. The popularity of the Lasso partially relies on the existence of fast and efficient implementation algorithms. For example, using the piecewise linearity of the Lasso estimator, a modification of the *Least Angle Regression* (LARS) algorithm can compute the whole optimal path (corresponding to all $\lambda_n \in [0, \infty]$) of the Lasso estimator on the same order as computational complexity of the least squares with size $n \times \min(n, p)$ [45]. A similar homotopy algorithm was proposed in [101]. These attractive algorithms allow the scalability of the Lasso to high-dimensional situations.

On the other hand, asymptotic properties of the Lasso estimator have also been extensively studied and analyzed. In a seminal work [77], Knight and Fu first derived the asymptotic distribution of the Lasso estimator (more generally the bridge estimator) and proved its estimation consistency under the shrinkage rate $\lambda_n = o(\sqrt{n})$ and $\lambda_n = o(n)$. More specifically, as long as errors are iid and possess a common finite second moment σ^2, the $\sqrt{n}$ scaled Lasso estimator with a sequence of properly tuned shrinkage parameters $\{\lambda_n\}_{n \in \mathbb{N}}$ has an asymptotic normal distribution with variance $\sigma^2 C^{-1}$, where $n^{-1} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^* \to C$ and C is a positive definite matrix. Later, [86] showed that there is a non-vanishing probability of the Lasso selecting wrong models with the optimal prediction criterion such as cross-validation (CV). [95] also discovered the conflict between model selection consistency and optimal prediction in the Gaussian graphical model setup. [130] found a sufficient and necessary condition required on the design matrix for the Lasso estimator to be model selection consistent, i.e. the *irrepresentable condition.* This condition was also observed by [132]. In graphical models, [95] obtained a similar set of conditions for the variable selection consistency of the Lasso estimator, namely the *neighborhood stability* assumptions.

These conditions are in general not easy to verify. Therefore, instead of requiring conditions on the design matrix for model selection consistency, there are also several variants of the original Lasso. For examples, the relaxed Lasso [94] uses two parameters to separately control the model shrinkage and selection; the adaptive Lasso [132] leverages a simple adaptation procedure to shrink the irrelevant predictors to 0 while keeping the relevant ones properly estimated; [96] suggested employ a two-stage hard thresholding rule, in the spirit of the Gauss-Dantzig selector [27], to set very small coefficients to 0.

Since the groundbreaking work of [27] which provided non-asymptotic upper bounds on the ℓ_2 estimation loss of the Dantzig selector with large probability, parallel ℓ_2 error bounds were found for the Lasso estimator by [96] under the *incoherent design* condition and by [13] under the *restricted eigenvalue* condition. In a previous work of [27], they showed that minimizing the ℓ_1 norm of the coefficient vector subject to the linear system constraint can exactly recover the sparse patterns, provided the *restricted isometry* condition holds and the support of the noise vector is not too large [26]. [19] tightened all previous error bounds for noiseless, bounded error and Gaussian noise cases. These bounds are nearly optimal in the sense that they achieve within a logarithmic factor the LS errors as if the true model were known (oracle property). [121] derived a set of sharp constraints on the dimensionality, sparsity of the model and the number of observations for the Lasso to correctly recover the true sparsity pattern. The ℓ_∞ convergence rate of the Lasso estimator was obtained by [90]. Other bounds for the sparsity oracle inequalities of the Lasso can be found in [16].

As we have mentioned earlier, there is a second view of variable selection approaches built on the Bayesian paradigm. Recent work has been proposed in the direction of Bayesian Lasso [102]. In [102], with a conditional Gaussian prior on $\boldsymbol{\beta}$ and the non-informative scale-invariant prior on the noise variance being assumed, a Bayesian Lasso model is proposed and a simple Gibbs sampler is implemented. It is shown that the Bayesian Lasso estimates in [102] are strikingly similar to those of the ordinary Lasso. Since this Bayesian Lasso in [102] involves the inversion of the covariance matrix of block coefficients at each iteration, the computational complexity prevents its practical application with, say, hundreds of variables. Moreover, similar to the regular Lasso, the Bayesian Lasso in [102] uses only one shrinkage parameter t to control model size and shrink estimates. Nonetheless, it is arguable whether the two effects can be simultaneously well-handled by a single tuning parameter [94]. To mitigate this non-separability problem, [99] proposed an extended Bayesian Lasso model by assigning a more flexible, covariate-adaptive penalization on top of the Bayesian Lasso in the context of Quantitative Trait Loci (QTL) mapping. Alternatively, introducing different sources of sparsity-promoting priors on both coefficients and their indicator variables have been studied, e.g. in [104], where a normal-Jeffrey scaled-mixture prior on coefficients and an independent Bernoulli prior with small success probability on the binary index vector are combined.

Despite those appealing properties of the Lasso estimator and the advocacy of using the Lasso, the Lasso estimate is not guaranteed to provide a satisfactory estimation and detection performance, at least in some application scenarios. For instance, when the data are corrupted by some outliers or the noise is extremely heavy-tailed, the variance of the Lasso estimator can be quite large, usually become unacceptably large, even when the sample size approaches infinity [77]. Asymptotic analysis [77] and non-asymptotic error bounds on the estimation loss [13] both suggest that the performance of the Lasso linearly deteriorates with the increment of the noise power. A similar observation can sometimes be noted when the dimensionality of the linear model is very high while the data size is much smaller.

1.2.2 Our Contributions

In the above discussion, the distributions of measurement errors have not been specified. Prior literature has mainly been concerned with either *exact* recovery of sparse signals [23, 24, 26, 28, 42, 117] or *stable* recovery under iid noise with moderate variance, usually assumed to bounded or Gaussian (bounded in moments) [19, 25, 41, 43, 86]. In general, the error vector $\mathbf{e}$ is assumed to be iid Gaussian random variables with variance σ^2, i.e. $\mathbf{e}$ follows the distribution $N(\mathbf{0}, \sigma^2 I_{n\times n})$. As we have seen in the previous section, e.g. [77], it is clear that the accuracy of the Lasso estimator critically depends on σ^2 and the estimator is well suited for errors with a small or moderate variance σ^2. It is also noted that when σ^2 becomes larger, the Lasso estimator has a variance that is unbounded in the limiting case. This implies an undesirable property in real applications: instability. The reason for this poor performance with large σ^2 lies in the sensitivity of the ℓ_2 loss function to a few large errors which may arise from heavy-tailed distribution or outliers. This explains why empirical examples show that the Lasso estimator can behave poorly if data are contaminated [75].

The standard assumption of random errors with small variance σ^2, however, is unlikely to hold in many real applications, since outliers and/or heavy-tailed error distributions are commonly encountered in real situations. Therefore, the practical usage and efficiency of the Lasso can be limited. For example, it is typical for DNA microarray data to have a very low signal-to-noise ratio (SNR), meaning that σ^2 is large. Furthermore, in practice the number of observations we are able to afford can be less than the dimensionality of the assumed model. Therefore, a robust variable selection model is necessary to obtain a good estimator in terms of accuracy, at least asymptotically. By robustness, here we mean two things:

1. The estimate is asymptotically stable in the presence of large noise. More specifically, we hope that, with more and more data being collected, the variability of the estimate is acceptable even if the measurement errors (the errors in the responses) get larger and larger.

2. The estimate is robust against contamination of the assumed errors. More specifically, even when outliers are found in the responses, the estimation performance is comparable to the situations of having no outliers.

These two issues can be partially reflected in σ^2. The first scenario can be viewed as errors following a distribution with heavy tails (e.g. Student-t distribution, Cauchy distribution), while the second one can be modeled as errors and outliers together contributing to form a mixture model of distributions. In either of the two scenarios, the corresponding σ^2 can be very large or even infinity.

In the first part of my thesis (Chapter 2 and 3), robust Lasso-type regression models are considered when the noise has heavy-tails. More specifically, two solutions are proposed: (i) to leverage adaptive robust loss functions [31], as apposed to the Euclidean loss in standard Lasso; (ii) to adopt a fully Bayesian modeling framework [32]. Both solutions are aiming to obtain stabilized estimates.

In solution (i), we propose a robust version of the Lasso by adopting a convex combined loss function and derive the asymptotic theory of the proposed robust Lasso estimator. We show that the ordinary Lasso and the regularized least absolute deviation (RLAD) [123] are two special cases of the proposed robust Lasso model. Although the RLAD is a robust model selection tool, it has limitations in terms of *uniqueness*, *smoothness*, and *efficiency* of estimation. Specifically, since the objective function of the RLAD is purely piecewise linear (thus may not be strictly convex) in $\boldsymbol{\beta}$, its solution may not necessarily be unique in general [14]. Moreover, since the optimal path for the RLAD is discontinuous in λ_n, its estimator may have jumps with respect to (w.r.t.) a small amount of perturbation of the observed data even when the solution is unique. Finally, if the error distribution does not have many extreme values, then the RLAD is not an efficient estimator: the asymptotic efficiency of the RLAD estimator is just 63.7% compared with the Lasso estimator under the Gaussian error distribution. In contrast, the proposed robust Lasso model has advantages in terms of generality and flexibility. Combining ℓ_1 and ℓ_2 losses yields a robust solution and the combination weight can be tuned, either analytically or numerically estimated from data, to achieve the minimum asymptotic variance. Our asymptotic analysis also shows that under certain adaptation procedures and shrinkage conditions, the proposed approach is indeed model selection consistent. Meanwhile, for variables with non-zero coefficients, it will be shown that the proposed robust model has unbiased estimates and the variability of the estimates is stabilized compared with the ordinary Lasso estimator. Therefore, the oracle property in the sense of [51] is achieved for the proposed method.

We further derive a parallel asymptotic analysis of an alternative robust version of the Lasso with the Huber loss function, a.k.a. the Huberized Lasso. To the best of our knowledge, currently there is no asymptotic theory for the Huberized Lasso, although [107] empirically studied its performance. For the Huberized Lasso, asymptotic normality and model selection consistency are established under much weaker conditions on the error distribution, i.e. no finite moment assumption is required for preserving similar asymptotic results as in the convex combined case. Thus, the Huberized Lasso estimator is well-behaved in the limiting situation when the error follows a Cauchy distribution, which has infinite first and second moments. The analysis result obtained for the non-stochastic design is extended to the random design case with additional mild regularity assumptions. These assumptions are typically satisfied for auto-regressive models.

In solution (ii), we introduce two parameters in the proposed Bayesian Lasso model to separately control the model selection and estimation shrinkage in the spirit of [94] and [127]. In particular, we propose a Poisson prior on the model size and the Laplace prior on $\boldsymbol{\beta}$ to identify the sparsity pattern. Since the proposed joint posterior distribution is highly nonstandard and a standard MCMC is not applicable, we employ a reversible-jump MCMC (RJ-MCMC) algorithm to obtain the proposed Bayesian Lasso estimates by simultaneously performing model averaging and parameter estimation. It is worth emphasizing that, though RJ-MCMC algorithms have been developed in the literature before model selection and estimation purposes (e.g. [4] proposed a hierarchical Bayesian model and developed an RJ-MCMC algorithm for joint Bayesian model selection and estimation of noisy sinusoids; similarly [111] proposed an accelerated truncated Poisson process model for Bayesian QTL mapping), these methods are not intended for promoting sparse models whereas our model utilizes sparsity promoting priors in conjunction with the discrete prior on the model size. One advantage of the proposed model is that it requires no cross-validation for parameter tuning, which is computationally intensive and inevitable in the Lasso to determine the optimal parameters.

1.3 High-Dimensional Covariance and Precision Matrix Estimation

1.3.1 Challenges

In the second part of my thesis, I focus on the estimation of large covariance Σ and precision matrices $\Omega = \Sigma^{-1}$, when the number of observations is far fewer than the number of parameters in the matrix (Chapter 4 and 5). Estimation of the covariance and precision matrices for high-dimensional datasets is attracting increasing recent attention [10, 11, 20–22, 36, 47, 80]. It is challenging because: (i) there are $p(p+1)/2$ unknown parameters to estimate from n observations when $p \gg n$; (ii) Σ (hence Ω) is intrinsically positive definite. The estimation problem was partially motivated by many modern high-throughput devices that make huge scientific data available to us. While there are many potential practical applications for this method, a motivating application of this work is brain effective connectivity modeling using fMRI data, where the goal is to infer the connectivity networks, represented by non-zero entries in Ω, of a large number of (spatially-fixed) brain regions-of-interests (ROIs). In particular, some of the proposed models and algorithms have been further applied to learn the brain connectivity networks for Parkinson's disease (PD) [35], the second most common neuro-degenerative disorder in Canada. Studying brain connectivity is crucial in understanding brain functioning and can provide significant insight into the pathophysiology of neurological disorders such as PD. Based on prior neuroscience knowledge, the connections between brain regions generally can be considered *a priori* to form a sparse network.

Conventional statistical signal processing tools and mathematical models could fail at handling those huge datasets, due to either theoretical or algorithmic reasons. Example applications of covariance estimation for a large number of variables include, and of course are not limited to, array signal processing [1, 64], hyperspectral image classification [9], construction of genetic regulatory networks from microarray profiles [39] and brain functional MRI networks [91]. Suppose we have n data points $X_{n\times p} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}^T$ that are iid from a zero-mean, p-dimensional multivariate Gaussian $N(\mathbf{0}, \Sigma)$. The most natural estimator of the covariance matrix Σ is the *unstructured sample covariance matrix*[1]

$$\Sigma_n^\star = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \in \mathbb{R}^{p\times p 2}. \tag{1.3}$$

It has been well-known from the classical normal distribution theory that $\Sigma_n^\star$ is a "good" estimator of Σ when p is fixed and $n \to \infty$. Please see [3] for a thorough and recent discussion on this subject.

Unfortunately, the tools and results from the classical theory fail to work when the dimensionality p grows as the data size increases, a well-known fact called the *curse-of-dimensionality.* For instance, from the eigen-structure perspective, random matrix theory predicts that a recentred and rescaled version of the largest eigenvalue

[1] We assume in this definition that the sample mean vector $\bar{\mathbf{x}} = \mathbf{0}$. Statistical literature often uses $\Sigma_n^\star = (n-1)^{-1}\sum_i(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T = (n-1)^{-1}(X-\bar{X})^T(X-\bar{X})$ where $\bar{X}$ is the matrix stacking $\bar{\mathbf{x}}$ n-times. These two definitions, however, are asymptotically equivalent by noting that they have the same limiting spectral law and $X - \bar{X} = (I - n^{-1}\mathbf{1}\mathbf{1}^T)X$ which implies that $\|X - \bar{X}\| \leq \|I - n^{-1}\mathbf{1}\mathbf{1}^T\|\|X\| = \|X\|$ since the largest singular value of $(I - n^{-1}\mathbf{1}\mathbf{1}^T)$ is 1.

[2] In Chapter 4 and 5, we assume the samples take values in $\mathbb{R}^p$ and thus use $\mathbf{x}_i^T$ instead of the conjugate transpose $\mathbf{x}_i^*$. Nonetheless, we shall see from the concluding remarks of Chapter 5 that nothing prevents the obtained results extending to $\mathbb{C}^p$.

of $\Sigma_n^\star$ for a certain class of Σ has a Tracy-Widom limiting law, when $p/n \leq 1$ as n and p both go to infinity [46]. Therefore in particular, it is suggested that the vanilla principle component analysis (PCA) is not suitable when a large number of variables are projected to lower dimensional orthogonal subspaces based on a limited number of observations [72, 73]. Consider, for example, the identity covariance matrix Σ with all eigenvalues being equal to 1. Asymptotic random matrix theory roughly states that the largest eigenvalue is $\lambda_{\max}(\Sigma_n^\star) \cong (1 + \sqrt{p/n})^2$ and the smallest eigenvalue is $\lambda_{\min}(\Sigma_n^\star) \cong (1 - \sqrt{p/n})^2$, for n/p approaching to some positive ratio with n and p both going to infinity. In this case, the curse-of-dimensionality is phenomenal in the sense that the spectrum of the sample covariance matrix is more spread than the spectrum of Σ, the Dirac δ mass at 1.

A natural solution to mitigate this difficulty is to restrict our attention to subsets of covariance matrices with certain special structures or properties and meanwhile incorporate the *regularization* ideas into estimation. *Sparsity* is one commonly hypothesized condition and it seems to be realistic for many real-world applications. Considering certain sparse covariance matrices, simple banding [10], tapering [20], and thresholding [11] on $\Sigma_n^\star$ are shown to be consistent estimators for Σ. Surprisingly, some of these conceptually and computationally simple estimators are even shown to be optimal in the sense of minimax risk [20–22].

1.3.2 Estimating Covariance Matrix

Prior Arts

Significant recent progress has been made in both theory and methodology development for estimating large covariance matrices. Regularization has been widely employed. Broadly speaking, regularized estimation of large covariance matrices can be classified into two major categories. The first category includes Steinian shrinkage-type estimators that shrink the covariance matrix to some well-conditioned matrices under different performance measures. For instances, Lediot and Wolf (LW) [82] proposed a shrinkage estimator by using a convex combination between $\Sigma_n^\star$ and $p^{-1}\text{Tr}(\Sigma_n^\star)I$ and provided a procedure for estimating the optimal combination weight that minimizes the mean-squared errors and that is distribution-free. Chen et. al. [36] further extended the idea and improved the LW estimator through two strategies: one is based on the Rao-Blackwellization idea to condition on the sufficient statistics $\Sigma_n^\star$ and the other is to approximate the oracle by an iterative algorithm. Closed-form expressions of both estimators were given in [36]. More recently, Fisher and Sun [53] proposed using $\text{diag}(\Sigma_n^\star)$ as the shrinkage target with possibly unequal variances. These shrinkage estimators are amenable for general covariance matrices with "moderate-dimensionality". Here, by moderate-dimensionality, we mean that p grows nicely as n increases, e.g. $p \to \infty$ and $p = O(n^k)$ for some $0 < k \leq 1$.

Estimators in the second category directly operate on the covariance matrix through operators such as thresholding [10], banding [11], and tapering [20]. Banding and tapering estimators are suitable for estimating covariance matrices where a natural ordering exists in the variables such as covariance structures in time-series. Banding simply sets the entries far away from the main diagonal to be zeros and keeps the entries within the band unchanged. Tapering is similar to banding, with the difference in that it gradually shrinks the off-diagonal entries within the band to 0. We can view banding as a hard-thresholding rule while tapering is a soft-thresholding rule, up to a certain unknown permutation [12]. In contrast, thresholding can deal with general permutation-invariant covariance matrices and introduce sparsity without requiring additional structures. These estimators are statistically consistent if certain *sparsity* is assumed and the dimensionality p grows at any sub-exponential rate of n, which allows much larger covariance matrices be estimable. In fact, it is further known that tapering and thresholding estimators are minimax [20–22]. The rate-optimality under the operator norm is not true for the banding estimator in [11] and it was shown that tapering is generally preferred to banding [20]. However, it is worth mentioning that, when the assumed sparsity condition is invalid, all the above estimators in the second category become sub-optimal.

Our Contributions

Despite recent progress on large covariance matrix estimation, there has been relatively little fundamental theoretical study on comparing the shrinkage-category and tapering-category estimators. To fill this gap, we first study the risks of shrinkage estimators and provide a comparison of risk bounds between shrinkage and tapering estimators. Further, motivated by the observed advantages and disadvantages of shrinkage and tapering estimators under different situations, to properly estimate *general* and *high-dimensional* covariance matrices, we propose a shrinkage-to-tapering estimator that combines the strengths of both shrinkage and tapering approaches. The proposed estimator has the form of a general shrinkage estimator with the crucial difference that the shrinkage target matrix is a tapered version of $\Sigma_n^\star$. By adaptively combining $\Sigma_n^\star$ and a tapered $\Sigma_n^\star$, the proposed shrinkage-to-tapering oracle (STO) estimator inherits the optimality in the minimax sense when sparsity is present (e.g. AR(1)) and in the minimum mean-squared error (MMSE) sense when sparsity is absent (e.g. fractional Brownian motion). Therefore, the proposed estimator improves upon both shrinkage and tapering estimators. A closed-form of the optimal combination weight is given and a STO approximating (STOA) algorithm is proposed to determine the oracle estimator.

1.3.3 Estimating Precision Matrix

Prior Arts

Estimation of Ω is a more difficult task than estimating Σ because of the lack of natural and pivotal estimators as $\Sigma_n^\star$ when $p > n$. Nonetheless, accurately estimating Ω has important statistical meanings. For example, in Gaussian graphical models, a zero entry in the precision matrix implies the conditional independence between the corresponding two variables. Further, there are additional concerns in estimating Ω

beyond those we have already seen in estimating large covariance matrices.

First, since $\Sigma_n^\star$ is a natural estimator of Σ, so is $(\Sigma_n^\star)^{-1}$ of Ω. However it is obvious that $\Sigma_n^\star$ is invertible only if $p < n$. Even worse, assuming $\Sigma_n^\star$ is invertible, $(\Sigma_n^\star)^{-1}$ still does not converge to Ω in the sense of eigen-structure when $p/n \to c > 0$ [46, 72].

Secondly, it is known that, under mild hypotheses and for a certain class of sparse matrices, applying simple hard thresholding to $\Sigma_n^\star$ yields a consistent [11] and optimal estimator of Σ in the sense of minimax risk ([21, 22]). Therefore, $[T_t(\Sigma_n^\star)]^{-1}$, where T_t is the thresholding operator with cutoff t, is a natural estimator of Ω. Indeed, [21] has showed that this estimator is rate-optimal under the matrix L^1 norm when minimaxity is considered. Nonetheless, it is possible that $[T_t(\Sigma_n^\star)]^{-1}$ fails to preserve sparsity (including the sparsity measure in terms of strong ℓ_q-balls, see definition in (5.2)), because a sparse matrix does not necessarily have a sparse inverse which plays a central role in Gaussian graphical models. Hence, the natural estimator $[T_t(\Sigma_n^\star)]^{-1}$ of Ω proposed in [21] can be unsatisfactory.

Thirdly, state-of-the-art precision matrix estimation procedures are essentially based on penalized likelihood maximization or constrained error minimization approaches, e.g. CLIME [18], SPICE [108], graphical Lasso [56] and variants [7], adaptive graphical Lasso [50], SCAD [80], and neighborhood selection [95, 126]. They are optimization algorithms with different objective functions. They have, however, a common structural feature in the objective functions: one term is the goodness-of-fit and the other term measures the model size which is often formulated by sparsity promoting penalties such as matrix 1-norm, SCAD, etc. The interior point method is standard for solving the optimization problems; but it is computationally infeasible when the dimensionality is large. Moreover, its high computational cost can be magnified by the parameter tuning procedure such as cross-validation. It is worth mentioning that the graphical Lasso [56] can be solved in a looped LAR fashion [45] and thereby its computational cost to estimate $\Omega \in S_k$ (see Eqn. (5.1) for definition) is equivalent to solving a sequence of p least squares problems, each of which has the complexity of $O(p^3)$ in terms of basic algebraic operations over some rings, e.g. the real or complex numbers. Therefore, the computational complexity of the graphical Lasso is $O(p^4)$. Moreover, since the graphical Lasso has an outer loop for sweeping over the p columns, it is empirically observed that the graphical Lasso is problem-dependent and its computational cost can be prohibitively high when the true precision matrix is not extremely sparse.

Our Contributions

In light of these challenges in estimating the precision matrix when $p \gg n$, we propose a new easy-to-implement estimator with attractive theoretic properties and computational efficiency. The proposed estimator is constructed on the idea of the finite Neumann series approximation and constitutes merely matrix multiplication and addition operations. The proposed estimator has a computational complexity of $O(\log(n)p^3)$ for problems with p variables and n observations, representing a significant improvement upon the aforementioned optimization methods. The proposed estimator is promising for ultra high-dimensional real-world applications such as gene microarray network modeling.

We prove that, for the class of approximately inversely closed precision matrices, the proposed estimator is consistent in probability and in L^2 under the spectral norm. Moreover, its convergence is shown to be rate-optimal in the sense of minimax risk. We further prove that the proposed estimator is model selection consistent by establishing a convergence result under the entry-wise ∞-norm.

1.4 Thesis Outline

Now, we outline the structure of the rest of this thesis.

Chapter 2 [31] presents our research on the robust Lasso models and their asymptotic properties. More specifically, we propose a robust version of the Lasso and derive the limiting distribution of its estimator, from which the estimation consistency can be immediately established. We further prove the model selection consistency of the proposed robust Lasso under an adaptation procedure for the penalty weight. Meanwhile, a parallel asymptotic analysis is performed for the Huberized Lasso, a previously proposed robust Lasso [107]. We show that the Huberized Lasso estimator preserves similar asymptotics even with a Cauchy error distribution. Therefore, our analysis shows that the asymptotic variances of the two robust Lasso estimators are stabilized in the presence of large variance noise, compared with the unbounded asymptotic variance of the ordinary Lasso estimator. Finally, the asymptotic analysis from the non-stochastic design is extended to the case of random design.

Chapter 3 [32] presents our research on the Bayesian Lasso model. In this work, we first utilize the Bayesian interpretation of the Lasso model and propose several hyper-priors to extend the Lasso to the fully Bayesian paradigm. Since the proposed Bayesian Lasso contains discrete and continuous (hyper-)parameters that simultaneously control the model size and parameter shrinkage, we construct a provably convergent reversible-jump MCMC algorithm to obtain its numeric estimates. We use simulations to show the improved performance of the proposed Bayesian Lasso model in terms of estimation error and pattern recovery.

Chapter 4 [33] presents our research on the estimation of high-dimensional covariance matrices based on a small number of iid Gaussian samples. In this chapter, we first study the asymptotic risk of the MMSE shrinkage estimator proposed by [36] and show that this estimator is statistically inconsistent for a typical class of sparse matrices that often appear as the covariance of auto-regressions. We then propose a shrinkage-to-tapering oracle estimator that improves upon both shrinkage and tapering estimators. We further develop an implementable approximating algorithm for the proposed estimator.

Chapter 5 [31] presents our research on the estimation of high-dimensional precision matrices. In this research, we propose an efficient algorithm involving only matrix multiplication and addition operations based on the truncated Neumann series representation. The proposed algorithm has a computational complexity of $O(\log(n)p^3)$ in terms of basic algebraic operations over real or complex numbers. We prove that, for the class of approximately inversely closed precision matrices, the proposed estimator is consistent in probability and in L^2 under the spectral norm. Moreover, its convergence is shown to be rate-optimal in the sense of minimax risk. We further prove that the proposed estimator is model selection consistent by establishing a convergence result under the entry-wise ∞-norm. Finally, we apply the proposed method to learn functional brain connectivity from frontal cortex directly to the subthalamic nucleus based on fMRI data.

Chapter 6 [35] presents an application of the group robust Lasso model to an fMRI group analysis. In this work, we consider incorporating sparsity into brain connectivity modeling to make models more biologically realistic and performing group analysis to deal with inter-subject variability. To this end, we propose a group robust Lasso model by combining advantages of the group Lasso and robust Lasso model developed in Chapter 2. The group robust Lasso is applied to a real fMRI data set for brain connectivity study in Parkinson's disease, resulting in biologically plausible networks.

Chapter 7 briefly summarizes the major contributions of the thesis and provides concluding remarks. A number of future topics are discussed.

Appendices include the notations used in the thesis, proofs of lemmas and theorems stated in the thesis body.

Chapter 2

Robust Lassos and Their Asymptotic Properties

2.1 Introduction

2.1.1 Sparsity-Promoting Linear Models

In this chapter, we consider the problem of estimating the coefficient vector in a linear regression model, defined as

$$\mathbf{y} = X\boldsymbol{\beta} + \mathbf{e}. \tag{2.1}$$

Here X is the $n \times p$ design matrix which can either be non-stochastic or random. As per convention, rows of X represent the p-dimensional observations and columns of X represent the predictors. $\mathbf{y}$ is the response vector and $\boldsymbol{\beta}$ is the coefficient vector to be estimated. We regard $\mathbf{e}$ as a column vector, and use $\mathbf{e}^*$ to denote its conjugate transpose. The random measurement error vector $\mathbf{e} = (e_1, \cdots, e_n)^*$ is assumed to be iid with zero mean. Here we do not have to generally assume that the error possesses a finite second moment σ^2 for each component. Recovering the true sparse coefficients in the presence of random errors with large variability is of primary interest of this chapter.

For a sparse regression problem where p is large but the number of true predictors with non-zero coefficients is small, the traditional least squares (LS) estimator for the full model in (2.1) may not be feasible. Even if the LS estimator exists and is unique, it can have an unacceptably high variance since including unnecessary predictors can significantly degrade estimation accuracy. More importantly, the LS estimator cannot be naturally interpreted as extracting (sparse) signals from the ambient linear system, which can make subsequent inferences difficult. In fact, for the case of p being large, overfitting is a serious problem for all statistical inferences based on the full model since the data always prefer models with a greater number of coefficients. Therefore, sparse representation for a large linear system is crucial for extracting true signals and improving prediction performance. Robust model selection for recovering sparse representations is a problem of interest for both theory and applications.

Depending on the particular research interest and application, the problem of recovering sparse representations can be formulated accordingly to different scenarios depending upon the relative magnitudes of p and n. One scenario is the over-determined case, i.e. $p < n$, such as p being fixed and $n \to \infty$. Another scenario of great interest is the under-determined case, i.e. $p > n$. For instance, the typical setup in compressed sensing is $p \gg n$ with n being fixed for a deterministic X [26]. In this chapter, following [51, 77, 122, 123], we assume the classical over-determined case, such that $p < n$ and p is fixed, and formulate the sparse linear regression model accordingly. In particular, to study the asymptotic performance, we use the setting as in [77, 122] that $n \to \infty$ and p is presumably large.

Though differently formulated, these two scenarios are in fact related, and theoretical results from one scenario have implications for the other. Wainwright showed that asymptotic results of the Lasso estimator (defined in (2.2)) in the classical scenario [77] continue to be true in the double-asymptotic scenario (both n and p approach to infinity) [121]. Connections between these two scenarios can also be found in [130], and both scenarios are areas of active research. There has been extensive development in the theory and application for under-determined systems [23, 24, 26–28, 42, 43]. For instance, sparse recovery for under-determined linear systems is vital to the area of compressed sensing [24–26, 42]. Additionally, identifying non-zero coefficients in a large over-determined linear system subject to random errors is also pertinent to the signal and image processing and machine learning communities. For examples, face recognition using the sparse PCA, where contaminated face images are recovered by only a few principal components representing different facial features [123]. In [69], face images can be represented by learnt features that are basis vectors with sparse coefficients in the matrix factorized domain. In [119], sparse mAR model with a fixed number of brain ROIs whose intensity signals are measured by magnetic resonance scans over time is used to model brain connectivity. The fused Lasso [57], which is closely related to the total variation denoising in signal processing, is used to reconstruct images with sparse coefficients and gradients.

We note that our approach differs from the typical scenario in compressed sensing since they serve different purposes. Compressed sensing research mainly focuses on the theory and algorithms required to (exactly or approximately) discover a sparse representation [24–26, 42], e.g. determining the value of n to exactly recover sparse signals in the absence of noise. Since error bounds on the estimates suggest that the approximation quality deteriorates linearly with the dispersion of errors [41], any sparse approximation, $\hat{\boldsymbol{\beta}} \approx \boldsymbol{\beta}$, will be inaccurate when the underlying error distribution has heavy tails. In this paper, we concentrate on developing robust estimators to recover the true sparse coefficients in the presence of large noise. We also study the properties of the proposed estimators such as estimation and model selection consistency. Therefore, our analysis is put in an asymptotic framework.

For the purpose of variable/model selection in linear regression problems, a variety of methodologies have been proposed. Though different methods employ different selection criteria, all approaches share a common feature: penalizing models with a larger number of coefficients. There is no general agreement on which cost function and penalization criterion type is optimal. However, prior approaches can be classified according to the choice of loss/cost and penalization functions:

1. *ℓ_2 loss with various functional forms of the penalty:* ℓ_2 loss (also referred as LS cost function, the squared error loss) is widely employed. Early approaches

used the ℓ_2 loss coupled with penalties proportional to the model size, i.e. the ℓ_0 norm of the coefficient vector. For example, the Akaike Information Criterion (AIC) [2], Bayesian Information Criterion (BIC) [109], and Risk Inflation Criterion (RIC) [54] are popular model selection criteria of this type. Since these penalized LS cost functions are combinatorial in nature, estimating and comparing all possible models may become computationally impractical, particularly with high dimensional models. Therefore, more efficient and practical model-shrinkage tools have been proposed, such as ridge regression which combines the LS estimator with the ℓ_2 penalty for the coefficients [68]. Although ridge regression can reduce the model size in terms of the numeric magnitudes of estimates, it shrinks all coefficients and thus cannot perform model selection and parameter estimation at the same time. The *least absolute shrinkage and selection operator* (Lasso) proposed by [115] is a popular and useful technique for simultaneously performing variable/model selection and estimation. Specifically, the Lasso estimator of Tibshirani is defined as

$$\hat{\boldsymbol{\beta}}_n = \arg\min_{\mathbf{u}\in\mathbb{R}^p} \left(\frac{1}{n}\sum_{i=1}^{n} (y_i - \mathbf{x}_i^*\mathbf{u})^2 + \frac{\lambda_n}{n}\sum_{j=1}^{p} |u_j|^\gamma \right), \tag{2.2}$$

where $\mathbf{x}_i$ is the i^{th}-row of X and $\gamma = 1$. This is the same definition (1.2) used in Chapter 1 and the properties of the Lasso estimator has been thoroughly discussed therein. In addition, we remark that [51] proposed a non-convex penalized LS cost function, the *Smoothly Clipped Absolute Deviation* (SCAD) model, which can avoid the over-penalization problem of the Lasso.

2. *ℓ_∞ loss with the ℓ_1 penalty:* As an alternative to the ℓ_2 cost function, the *Dantzig selector* [27] combines the ℓ_∞ error measurement criterion and the ℓ_1 penalty:

$$\min_{\mathbf{u}\in\mathbb{R}^p} \|\mathbf{u}\|_{\ell_1} \quad \text{subject to} \quad \|X^*(\mathbf{y} - X\mathbf{u})\|_{\ell_\infty} \leq \lambda_n. \tag{2.3}$$

A small λ_n ensures a good approximation and the minimization yields a maximized sparsity. The Dantzig selector can be efficiently solved by recasting as a linear programming problem.

3. *Robust losses with the ℓ_1 penalty:* The lack of robustness of the ℓ_2 loss is well-known. Since the ℓ_1 loss function is more robust to outliers, the corresponding regression model leads to a robustified version of the LS regression. This regression model is called the *Least Absolute Deviation* regression (LAD) in the literature. Unfortunately, in the context of linear model selection, robustness has not received much attention compared to say, the Lasso. This is largely due to the difficulty of handling a non-differentiable ℓ_1 loss function. To our best knowledge, there are only a few studies considering robust losses. The *regularized LAD* (RLAD) model for robust variable selection has been proposed which can be recasted as a linear program [123]. The RLAD adopts the ℓ_1 loss coupled with the ℓ_1 penalty. An alternative robust version of Lasso can be formed by using the Huber loss with the ℓ_1 penalty to create a *Huberized Lasso* which is robust to contamination [107].

2.1.2 Summary of Present Contributions

We shall propose a robust Lasso and mainly focus on the asymptotic theory of its estimator since there is no general theory that guarantees the consistency of a selection criterion. Our asymptotic analysis put forth here shows that under certain adaptation procedure and shrinkage conditions, the proposed estimator is model selection consistent. Meanwhile, for variables with non-zero coefficients, it will be shown that the proposed robust model has unbiased estimates and the variability of the estimates is stabilized compared with the ordinary Lasso estimator. Therefore, the oracle property [51] is achieved for the proposed method. Now, we summarize our contribution as follows:

1. We propose using a convex combined loss of ℓ_1 (LAD) and ℓ_2 (LS), rather than the pure LS cost function, coupled with the ℓ_1 penalty to produce a robust

version of the Lasso. Asymptotic normality is established, and we show that the variance of the asymptotic normal distribution is stabilized. Estimation consistency is proved at different shrinkage rates for $\{\lambda_n\}$ and further proved by a non-asymptotic analysis for the noiseless case.

2. Under a simple adaptation procedure, we show that the proposed robust Lasso is model selection consistent (defined in (2.14)), i.e. the probability of the selected model to be the true model approaches to 1.

3. As an extension of the asymptotic analysis of our proposed robust Lasso, we study an alternative robust version of the Lasso with the Huber loss function, the Huberized Lasso. To the best of our knowledge, currently there is no asymptotic theory for the Huberized Lasso, although [107] empirically studied its performance. For the Huberized Lasso, asymptotic normality and model selection consistency are established under much weaker conditions on the error distribution, i.e. no finite moment assumption is required for preserving similar asymptotic results as in the convex combined case. Thus, the Huberized Lasso estimator is well-behaved in the limiting situation when the error follows a Cauchy distribution, which has infinite first and second moments.

4. The analysis result obtained for the non-stochastic design is extended to the random design case with additional mild regularity assumptions. These assumptions are typically satisfied by auto-regressive models.

2.1.3 Organization of the Chapter

The rest of the chapter is organized as follows. We introduce the proposed robust version of the Lasso with convex combined loss in Section 2.2. Its asymptotic behavior is then studied and compared with the Lasso. Section 2.3 defines an adaptive robust Lasso and its model selection consistency is proved. Section 2.4 concerns the Lasso with the Huber loss function and its asymptotic behavior is analyzed. Section 2.5

extends the analysis results from the non-stochastic design to the random design under additional mild regularity conditions. In Section 2.6, a simulation study is used to support the theoretical results found in previous sections.

2.2 A Convex Combined Robust Lasso

2.2.1 A Robust Lasso with the Convex Combined Loss

As discussed earlier, the ℓ_2 loss in the Lasso model is not robust to heavy-tailed error distributions and/or outliers. This indicates that the Lasso is not an ideal goodness-of-fit measure criterion in the presence of noise with large variance. In order to build a robust, sparsity-promoting model, we propose a flexible robust version of the Lasso, where the estimator $\hat{\boldsymbol{\beta}}_n$ is defined as

$$\hat{\boldsymbol{\beta}}_n = \arg\min_{\mathbf{u}\in\mathbb{R}^p} \left(\frac{1}{n} \sum_{i=1}^{n} L(\mathbf{u}; y_i, \mathbf{x}_i) + \frac{\lambda_n}{n} \|\mathbf{u}\|_{\ell_1} \right) \tag{2.4}$$

where the cost function,

$$L(u; y_i, \mathbf{x}_i) = \delta \left(y_i - \mathbf{u}^*\mathbf{x}_i\right)^2 + (1-\delta)\left|y_i - \mathbf{u}^*\mathbf{x}_i\right|,$$

is a convex combination of the ℓ_2 and ℓ_1 losses of $y_i - \mathbf{u}^*\mathbf{x}_i$, and $\delta \in [0,1]$. Note that this reduces to the traditional Lasso if $\delta = 1$, and it reduces to the RLAD if $\delta = 0$.

2.2.2 Asymptotic Normality and Estimation Consistency

In order to ensure that there is no strong dependency among the predictors (columns of X), namely model identifiability, we need a regularity assumption on the design matrix. Here, we assume the following classical conditions:

Non-stochastic design assumptions

1. *(Design matrix assumption)* The Gram matrix $C_n = n^{-1}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^*$ converges

to a positive definite matrix C as $n \to \infty$.

2. *(Error assumption)*

   (a) The error has a symmetric common distribution w.r.t. the origin. So $Ee_i = 0$ and the median of e_i is 0.

   (b) e_i has a continuous, positive probability density function (p.d.f.) f w.r.t. the Lebesgue measure in a neighborhood of 0.

   (c) e_i possesses a finite second moment σ^2.

Remark 1. [77] and [130] made an additional assumption on the design matrix for the fixed p case:

$$n^{-1} \max_{1 \le i \le n} \mathbf{x}_i^* \mathbf{x}_i \to 0 \tag{2.5}$$

as $n \to \infty$, i.e. $\max_{1 \le i \le n} |\mathbf{x}_i| = o(\sqrt{n})$. However, we shall show that this regularity condition is unnecessary and it is actually a direct consequence of assumption 1. Please refer to Lemma A.2.3 for a proof. Note that (2.5) has already been observed by [105].

Our first main theorem described below is to establish the asymptotic normality of the robust Lasso estimator. Since the loss function is non-differentiable, the usual Taylor expansion argument fails and a more subtle argument is required.

Theorem 2.2.1. *Under the above assumptions 1 and 2, if $\lambda_n/\sqrt{n} \to \lambda_0 \ge 0$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \Rightarrow \arg\min(V)$ where*

$$\begin{aligned} V(\mathbf{u}) &= (\delta + (1-\delta) f(0))\, \mathbf{u}^* C \mathbf{u} + \mathbf{u}^* \mathbf{W} \\ &+ \lambda_0 \sum_{j=1}^{p} [u_j sgn(\beta_j)\, 1(\beta_j \neq 0) + |u_j| 1(\beta_j = 0)] \end{aligned}$$

and

$$\mathbf{W} \sim N(\mathbf{0}, \big((1-\delta)^2 + 4\delta^2\sigma^2 + 4\delta(1-\delta)M_{10}\big)\, C).$$

An immediate consequence of Theorem 2.2.1 is:

Corollary 2.2.2. *If $\lambda_n = o(\sqrt{n})$, then $\hat{\boldsymbol{\beta}}_n$ is $\sqrt{n}$-consistent.*

A couple of observations can be made from Corollary 2.2.2. If $\delta = 0$, then we have an asymptotic variance of $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ equal to $\frac{1}{4f(0)^2}C^{-1}$, which is the reduced case of using a pure ℓ_1 loss without penalization [105]. When compared with the asymptotic variance of the ordinary Lasso $\sigma^2 C^{-1}$ (c.f. Theorem 2 in [77]) which is unbounded when σ^2 goes to infinity, our estimator has a finite asymptotic variance when δ is chosen carefully. As long as the value of $\delta^2\sigma^2$ is well controlled, the corresponding estimator can be stabilized asymptotically. Hence, it is desirable to seek a $\delta \in [0,1]$ which yields the minimum of the asymptotic variance. Assume for now that we know the error distribution and let us consider the asymptotic variance in (10). Let $x = \delta^{-1} - 1 \geq 0$ for $0 < \delta \leq 1$, and define

$$v(x) = \frac{x^2 + 4\sigma^2 + 4M_{10}x}{4(1 + f(0)x)^2} = \frac{1}{4} \times \frac{x + 4\sigma^2 x^{-1} + 4M_{10}}{f(0)^2 x + x^{-1} + 2f(0)}. \tag{2.6}$$

Ignoring the terms $4M_{10}$ and $2f(0)$ for a moment, it is easy to observe from the arithmetic-geometric mean inequality that the numerator of $v(x)$ is minimized at 2σ and the denominator at $f(0)^{-1}$. If $2\sigma > 1/f(0)$, then the numerator of $v(x)$ will dominate its denominator. Hence, $v(x)$ is minimized when $x \to \infty$, i.e. $\delta \to 0$. In another word, the convex combined robust Lasso is reduced to the RLAD for the case of having noise with large variance, to achieve the optimal asymptotic variance. Similarly, if $2\sigma < 1/f(0)$, the denominator dominates the numerator as $x \to 0$, i.e. $\delta \to 1$. The optimal weight of the robust Lasso corresponds to the special case of the ordinary Lasso when the noise has a moderate variance. Nevertheless, taking also the terms $4M_{10}$ and $2f(0)$ into account, the optimal δ may lie in the interval of $(0,1)$. Hence, our robust Lasso provides better flexibility.

In practice, the error distribution is usually unknown and thus the analytical form of the optimal weight is unavailable. Fortunately, we can still estimate the convex combined weight from the data by allowing the weight to be data dependent

$\delta_n = \delta(\{\mathbf{x}_i, y_i\}_{i\in\{1,\cdots,n\}})$. For example, an intuitive choice of measuring the spreadness of empirical errors is to use a renormalized quantity such as

$$\delta_n = (\hat{\sigma}^2 + 1)^{-1/2}, \tag{2.7}$$

where $\hat{\sigma}^2 = n - p^{-1} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^* \hat{\boldsymbol{\beta}}_{LS}\right)^2$ and $\hat{\boldsymbol{\beta}}_{LS}$ is the LS estimator for the linear model. If the noise variance is large, then δ_n is likely to concentrate within a small neighborhood of zero, and thus the robust Lasso behaves more like the RLAD. On the contrary, the ℓ_2 component can dominate the ℓ_1 if the error distribution has a small variance. When σ^2 is large, the robust Lasso estimator with the data-driven weight in (2.7) has an asymptotic variance which is not larger than $\frac{9}{4f(0)^2}C^{-1}$, as shown in the following Corollary 2.2.3.

Corollary 2.2.3. *Suppose $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$ and choose δ_n as in (2.7). Then $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \Rightarrow \arg\min(V)$ where*

$$V(u) = (\delta + (1-\delta)f(0))\,\mathbf{u}^* C\mathbf{u} + \mathbf{u}^* W + \lambda_0 \sum_{j=1}^{p} [u_j sgn\,(\beta_j)\,1(\beta_j \neq 0) + |u_j| 1(\beta_j = 0)]$$

and

$$W \sim N\left(\mathbf{0}, v_\delta C\right),$$

with

$$v_\delta = \left(1 - \sqrt{\frac{1}{\sigma^2+1}}\right)^2 + \frac{4\sigma^2}{\sigma^2+1} + 4\sqrt{\frac{1}{\sigma^2+1}}\left(1 - \sqrt{\frac{1}{\sigma^2+1}}\right) M_{10}.$$

Remark 2. By Jensen's inequality, we have

$$M_{10}^2 = (E|e_i|)^2 \leq Ee_i^2 = \sigma^2.$$

So

$$v_\delta \leq \left[\left(1 - \sqrt{\frac{1}{\sigma^2+1}}\right) + \frac{2\sigma}{\sqrt{\sigma^2+1}}\right]^2 \leq 9.$$

In light of the $\sqrt{n}$-rate convergence, we actually allow the sequence of $\{\lambda_n\}$ to grow faster while meantime preserving the estimation consistency, as demonstrated in the following Theorem 2.2.4.

Theorem 2.2.4. *Under the assumptions 1 and 2, if* $\lambda_n/n \to \lambda_0 \geq 0$*, then* $\hat{\boldsymbol{\beta}}_n \xrightarrow{P} \arg\min(Z)$ *where*

$$Z(\mathbf{u}) = \delta(\mathbf{u}-\beta)^* C(\mathbf{u}-\beta) + \delta\sigma^2 + (1-\delta)r + \lambda_0 \|\mathbf{u}\|_{\ell_1}, \tag{2.8}$$

and $r = \lim_{n\to\infty} n^{-1}\sum_{i=1}^n E|y_i - \mathbf{u}^*\mathbf{x}_i| < \infty$*. In particular, if* $\lambda_n = o(n)$*, then* $\arg\min(Z) = \boldsymbol{\beta}$ *so that* $\hat{\boldsymbol{\beta}}_n$ *is a consistent estimator of* $\boldsymbol{\beta}$*.*

2.2.3 A Bound on MSE for the Noiseless Case

We have seen the asymptotic variance of the robust Lasso estimator, which does not necessarily hold for the case of having a finite sample size n. A more interesting question is that, given a fixed design matrix $X_{n\times p}$ and the assumed linear model, how accurately we can recover the true $\boldsymbol{\beta}$ using the robust Lasso? In this section, we would like to answer this question under a simpler scenario, i.e. in the noiseless case. This explicit estimation error bound due to the bias provides an implication on the asymptotic behavior of the robust Lasso estimator in the presence of noise. Indeed, statistical common sense tells us that the variance of the robust Lasso estimator can be smaller than the unpenalized case because of the bias-variance trade-off. Therefore, the mean squared estimation loss is expected to be controlled at the order of the LS+LAD, provided the bias term is small.

More specifically, our observation is that, under the shrinkage rate $\lambda_n = o(n)$ and certain assumptions on X, the proposed robust Lasso can accurately estimate $\boldsymbol{\beta}$ in terms of ℓ_2 loss in the absence of noise. Assume that $\boldsymbol{\beta}$ is S-sparse, i.e. $|\text{supp}\,(\boldsymbol{\beta})| =$

$|A| = S$. Let

$$\phi_{\min}(S) = \min_{T \le S} \inf_{|\mathbf{u}| \le T, \mathbf{u} \neq \mathbf{0}} \frac{\mathbf{u}^* C_n \mathbf{u}}{\mathbf{u}^* \mathbf{u}} \tag{2.9}$$

$$\phi_{\max}(S) = \max_{T \le S} \sup_{|\mathbf{u}| \le T, \mathbf{u} \neq \mathbf{0}} \frac{\mathbf{u}^* C_n \mathbf{u}}{\mathbf{u}^* \mathbf{u}} \tag{2.10}$$

be the restricted extreme eigenvalues of the submatrices of $C_n$with the number of columns being less than or equal to S. We assume a similar *incoherent design* condition as in [96]. That is, we assume there is a positive integer S_0 such that

$$\frac{S_0 \phi_{\min}(S_0)}{S \phi_{\max}(p - S_0)} > 16. \tag{2.11}$$

This condition measures the linear independency among restricted sets of the columns of X. A large value of the LHS in (2.11) prevents degeneracy of the restricted columns of X. With this incoherent design hypothesis, we can show that the ℓ_2 estimation loss decays to 0 if $\{\lambda_n\}$ grow at a proper rate.

Proposition 2.2.5. *Assume $\sigma^2 = 0$ and $\boldsymbol{\beta}$ is S-sparse. Suppose that the incoherent design condition (2.11) holds. Then the robust Lasso estimator $\hat{\boldsymbol{\beta}}_n$ for $\delta \in (0, 1]$ defined in (2.4) satisfies*

$$\left\| \hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} \right\|_{\ell_2} \le \frac{\lambda_n}{n} \frac{\sqrt{S}}{\delta D_0} - \frac{1 - \delta}{\delta \sqrt{n D_0}}, \tag{2.12}$$

where

$$D_0 = 1 - 4\sqrt{\frac{S \phi_{\max}(p - S_0)}{S_0 \phi_{\min}(S_0)}}. \tag{2.13}$$

Remark 3. In the noiseless case, Proposition 2.2.5 suggests that if $\lim_{n \to \infty} \lambda_n / n = 0$, then $\left\| \hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} \right\|_{\ell_2} \to 0$ as $n \to \infty$. This condition on the shrinkage rate is exactly the one assumed in Theorem 2.2.4. Hence, both the asymptotic and non-asymptotic analysis show that $\lambda_n = o(n)$ is sufficient for the conclusion that $\hat{\boldsymbol{\beta}}_n$ is consistent.

2.3 The Adaptive Robust Lasso and Its Model Selection Consistency

We have established the estimation consistency of the robust Lasso so far. However, in many scenarios, it is also desirable to have the model selection consistency, defined as

$$P\left(\text{supp}\left(\hat{\boldsymbol{\beta}}_n\right) = \text{supp}\left(\boldsymbol{\beta}\right)\right) \to 1 \tag{2.14}$$

as $n \to \infty$. Note that neither estimation consistency nor consistency in the ℓ_2 norm necessarily implies the model selection consistency. Consider, for a counterexample, that $\hat{\beta}_n[j] = n^{-1}$ for $\beta[j] = 0$.

In terms of choosing a sequence of shrinkage tuning parameters $\{\lambda_n\}_{n\in\mathbb{N}}$, [86] and [95] showed that the ordinary Lasso has a conflict between the consistency for model selection and optimal prediction. As a solution to achieve both estimation and model selection consistency, the adaptive Lasso [132] was proposed and its model selection consistency and asymptotic normality under certain rate of shrinkage were proved. To extend the idea of the adaptive Lasso [132] to our proposed robust Lasso, we define the adaptive robust Lasso as

$$\hat{\boldsymbol{\beta}}_n = \arg\min_{\mathbf{u}\in\mathbb{R}^p} \left(\frac{1}{n}\sum_{i=1}^{n} L(\mathbf{u}; y_i, \mathbf{x}_i) + \frac{\lambda_n}{n}\sum_{j=1}^{p} \hat{w}_j \left|u_j\right| \right), \tag{2.15}$$

where $\hat{\mathbf{w}} = (\hat{w}_1, \cdots, \hat{w}_p)^*$ is a vector of adaptive weights, which allow unequal penalties for the coefficients. For example, we can take $\hat{\mathbf{w}} = 1/\left|\hat{\boldsymbol{\beta}}_{LS}\right|^{\gamma}$ for some $\gamma > 0$. Let $A = \{j : \beta_j \neq 0\}$ and $A_n = \{j : \hat{\beta}_n[j] \neq 0\}$. By definition, the estimator $\{\hat{\boldsymbol{\beta}}_n\}$ is said to be consistent for model selection if and only if $P(A_n = A) \to 1$ as $n \to \infty$. Now following the similar argument as in [132], we have the following theorem showing the model selection consistency of the adaptive robust Lasso.

Theorem 2.3.1. *Suppose assumption 1 and 2 are satisfied. Let $\lambda_n = o(\sqrt{n})$ and $\lambda_n n^{(\gamma-1)/2} \to \infty$ for some $\gamma > 0$. Then the adaptive robust Lasso defined in (2.15)* *with* $\hat{\mathbf{w}} = 1/\left|\hat{\boldsymbol{\beta}}_{LS}\right|^{\gamma}$ *has the following properties:*

1. *Asymptotic normality for the true non-zero coefficients, i.e.*

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n[A] - \boldsymbol{\beta}[A]\right) \Rightarrow N\left(\mathbf{0}, \frac{(1-\delta)^2 + 4\delta^2\sigma^2 + 4\delta(1-\delta)M_{10}}{4[\delta + (1-\delta)f(0)]^2} C_{11}^{-1}\right).$$

2. *Model selection consistency if* $\|\mathbf{x}^j\|_{\ell_1} = O(\sqrt{n})$ *for all* $j \notin A$.

Remark 4. The additional condition, $\|\mathbf{x}^j\|_{\ell_1} \leq K\sqrt{n}$, for the model selection consistency is not trivial. It can be implied, for example, by $\sum_{i=1}^n \left|\frac{\mathbf{x}_i}{\sqrt{n}}\right| 1\left(a_n \left|\frac{\mathbf{x}_i}{\sqrt{n}}\right| > \tau\right) \to 0$ for every $\tau > 0$ and $a_n = \left(\sum_{i=1}^n \left|\frac{\mathbf{x}_i}{\sqrt{n}}\right|^3\right)^{-1/2}$.

2.4 The Huberized Lasso

For the robustness purpose, as an alternative for using a convex combination of ℓ_1 and ℓ_2 losses, we can use the Huber loss function and thus the corresponding ℓ_1-penalized model is called *the Huberized Lasso* [107]. The Huberized Lasso is defined as

$$\hat{\boldsymbol{\beta}}_n^H = \arg\min_{\mathbf{u}\in\mathbb{R}} \left(\frac{1}{n}\sum_{i=1}^{p} L(\mathbf{u}; y_i, \mathbf{x}_i) + \frac{\lambda_n}{n}\|\mathbf{u}\|_{\ell_1}\right), \tag{2.16}$$

where

$$L(\mathbf{u}; y_i, \mathbf{x}_i) = \begin{cases} (y_i - \mathbf{u}^*\mathbf{x}_i)^2 & \text{if } |y_i - \mathbf{u}^*\mathbf{x}_i| \leq \delta, \\ 2\delta|y_i - \mathbf{u}^*\mathbf{x}_i| - \delta^2 & \text{if } |y_i - \mathbf{u}^*\mathbf{x}_i| > \delta. \end{cases}$$

The Huberized Lasso enjoys the everywhere differentiability which is not true for the convex combined loss. Although the Huberized Lasso has already been used as a robustified version of the Lasso, currently there is no asymptotic theory for it. Here, we expect the Huberized Lasso to have similar asymptotic properties to the case of the convex combination loss. We first establish the asymptotic normality of the Huberized Lasso. It is worth mentioning that the proof details are considerably more complicate than the convex combined loss case.

Remarkably, as shown in Theorem 2.4.1 below, we note that no condition is required on the finiteness of the variance or even the first moment for the error distribution in order to achieve the asymptotic normality (and model selection consistency) for the Huberzied Lasso estimator (and its adaptive version). In other words, assumption 2(c) is not required, and the minimal set of assumptions only include the symmetry of error distribution and the continuity of its p.d.f. around the transition points $\pm\delta$. Therefore, the asymptotic normality and model selection consistency results are still valid for the Cauchy errors, whose first and second moments are infinite.

Theorem 2.4.1. *Under the assumptions 1, 2(a), and 2(b), if $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$ and f is continuous at $\pm\delta$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_n^H - \boldsymbol{\beta}) \Rightarrow \arg\min(V)$ where*

$$\begin{aligned} V(\mathbf{u}) &= K_{0\delta}\mathbf{u}^* C\mathbf{u} + 2\mathbf{u}^*\mathbf{W} \\ &+ \lambda_0 \sum_{j=1}^{p} [u_j sgn\,(\beta_j)\, 1(\beta_j \neq 0) + |u_j| 1(\beta_j = 0)] \end{aligned} \tag{2.17}$$

and

$$\mathbf{W} \sim N\left(\mathbf{0}, (\delta^2 M_{0\delta} + K_{2\delta})C\right).$$

Here, the assumption 2(b) is understood as the continuity of f around $\pm\delta$.

Corollary 2.4.2. *If $\lambda_n = o(\sqrt{n})$, then $\hat{\boldsymbol{\beta}}_n^H$ is $\sqrt{n}$-consistent.*

Proof. For $\lambda_0 = 0$, $V(\mathbf{u}) = K_{0\delta}\mathbf{u}^* C\mathbf{u} + 2\mathbf{u}^*\mathbf{W}$ is minimized at

$$\arg\min(V) = -\frac{C^{-1}\mathbf{W}}{K_{0\delta}} \sim N\left(\mathbf{0}, \frac{\delta^2 M_{0\delta} + K_{2\delta}}{K_{0\delta}^2} C^{-1}\right).$$

□

We can give a numerical example of the stabilized asymptotic variance of the Huberized estimator when the error is Cauchy distributed with zero mean and scale parameter s:

$$f(e_i) = \frac{s}{\pi(s^2 + e_i^2)}.$$

Take $\delta = 1$, then $s = 3$ gives the asymptotic variance 20.53 while $s = 1$ stabilizes the variance to 2.55!

Similarly, the adaptive Huberized Lasso is defined as in (2.15) with the change of the loss function. The following Theorem 2.4.3 shows that the adaptive Huberized Lasso is model selection consistent. Since the proof almost follows the same line of Theorem 2.3.1, we omit the details here.

Theorem 2.4.3. *Suppose assumptions 1, 2(a), and 2(b) are satisfied. Let* $\lambda_n = o(\sqrt{n})$ *and* $\lambda_n n^{(\gamma-1)/2} \to \infty$ *for some* $\gamma > 0$. *Then the adaptive Huberized Lasso defined in (2.15) with* $\hat{\mathbf{w}} = 1/\left|\hat{\boldsymbol{\beta}}_{LS}\right|^\gamma$ *has the following properties:*

1. *Asymptotic normality for the true non-zero coefficients, i.e.*

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n^H[A] - \boldsymbol{\beta}[A]\right) \Rightarrow N\left(\mathbf{0}, \frac{\delta^2 M_{0\delta} + K_{2\delta}}{K_{0\delta}^2} C^{-1}\right). \tag{2.18}$$

2. *Model selection consistency.*

Remark 5. The adaptive weight used in the Huberized Lasso needs to be adjusted to $\hat{\boldsymbol{\beta}}_{LAD}$ in the case that the least squares estimator is not guaranteed to be a consistent estimator, for instance, when the error is Cauchy. Then the theorem continues to be true due to the fact that $\hat{\boldsymbol{\beta}}_{LAD}$ is also a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}$ under assumptions 1), 2a), and 2b) [105].

Remark 6. The additional assumption for the model selection consistency of the robust Lasso with the convex combined loss, i.e. $\|\mathbf{x}^j\|_{\ell_1} = O(\sqrt{n})$ for all $j \notin A$, is not required for that of the Huberized Lasso. The difference lies in the fact that the Huberized Lasso objective function is differentiable everywhere. Its derivative agrees with the derivative of the Lasso on $[-\delta, \delta]$ and is less than that of the Lasso otherwise.

2.5 Random Designs

Until now, we have discussed the limiting behaviors of the non-stochastic design matrix case. In practice, since no infinitely precise measurement device exists, there are also measurement errors in the predictors. This is also the situation for autoregression models. It would be interesting to ask the question at what extent and under what assumptions the previous results of the non-stochastic design case still hold for the random design case.

Let $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}_{n\in\mathbb{N}}, P)$ be a filtered stochastic process, that is $\mathcal{F}_n$ is an increasing sequence of sub-σ-fields of $\mathcal{F}$ and $\mathcal{F}_0 = \{\emptyset, \Omega\}$. Let $\sigma(e_i)$ be the σ-field generated by r.v. e_i.

Random design assumptions

1. *(Random design matrix assumption)* $C_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^* \xrightarrow{P} C$ where C is a positive definite matrix.

2. *(Measurability assumption)* $\mathbf{x}_i$ is $\mathcal{F}_{i-1}$-measurable for all $i \in \mathbb{N}$.

3. *(Error assumption)*

   (a) The error has a symmetric common distribution w.r.t. the origin. So $Ee_i = 0$ and median of e_i is 0.

   (b) e_i has a continuous, positive p.d.f. f w.r.t. the Lebesgue measure in a neighborhood of 0.

   (c) e_i possesses a finite second moment σ^2.

   (d) $\sigma(e_i)$ is independent of $\mathcal{F}_{i-1}$ for all $i \in \mathbb{N}$.

Example 2.5.1. Consider the auto-regression model

$$\mathbf{x}_i = \boldsymbol{\beta}^*\mathbf{x}_{i-1} + \mathbf{e}_{i-1}, \tag{2.19}$$

where $\{\mathbf{e}_i\}$ are assumed to be i.i.d. Let $\mathcal{F}_n = \sigma(\{\mathbf{e}_i\}_{i\in\{0,\cdots,n\}})$. Then the measurability assumption and part d) of the error assumption of the random design are satisfied.

Theorem 2.5.1. *If X and $\{e_n\}_{n\in\mathbb{N}}$ obey the set of random design assumptions and $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$, then the robust Lasso estimator $\hat{\boldsymbol{\beta}}_n$ defined in (2.4) satisfies $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \Rightarrow \arg\min(V)$ where*

$$\begin{aligned} V(\mathbf{u}) &= (\delta + (1-\delta)f(0))\, \mathbf{u}^* C\mathbf{u} + \mathbf{u}^*\mathbf{W} \\ &+ \lambda_0 \sum_{j=1}^{p} [u_j sgn\,(\beta_j)\, 1(\beta_j \neq 0) + |u_j| 1(\beta_j = 0)] \end{aligned}$$

and

$$\mathbf{W} \sim N(\mathbf{0}, \big((1-\delta)^2 + 4\delta^2\sigma^2 + 4\delta(1-\delta)M_{10}\big)\, C).$$

As a special case of random design, we now consider the Gaussian random matrix.

Corollary 2.5.2. *Suppose X is an $n \times p$ Gaussian random matrix obeying the random design matrix and measurability assumptions and $\{e_n\}_{n\in\mathbb{N}}$ obeys the error assumptions. If $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$, then the robust Lasso estimator $\hat{\boldsymbol{\beta}}_n$ defined in (2.4) satisfies $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \Rightarrow \arg\min(V)$ where*

$$\begin{aligned} V(\mathbf{u}) &= (\delta + (1-\delta)f(0))\, \mathbf{u}^* C\mathbf{u} + \mathbf{u}^*\mathbf{W} \\ &+ \lambda_0 \sum_{j=1}^{p} [u_j sgn\,(\beta_j)\, 1(\beta_j \neq 0) + |u_j| 1(\beta_j = 0)] \end{aligned}$$

and

$$\mathbf{W} \sim N(\mathbf{0}, \big((1-\delta)^2 + 4\delta^2\sigma^2 + 4\delta(1-\delta)M_{10}\big)\, C).$$

Proof. The corollary follows easily from Theorem 2.5.1 and the fact that the smallest singular value of X, $\sigma_{\min}(X) \to 1$ P-a.s. by the strong law of large numbers or [5]. □

Remark 7. Using the same conditioning argument, we can show the asymptotic normality of the Huberized Lasso for the random design case as well.

2.6 Numeric Examples

Since the Huberized loss is differentiable, piecewise quadratic and the penalty is piecewise linear in $\boldsymbol{\beta}$, it follows that $\hat{\boldsymbol{\beta}}_n^H(\lambda)$ is piecewise linear in λ and hence the whole path of shrinkage can be efficiently computed with the LARS-Lasso algorithm [107]. In contrast, since the convex combined loss is not differentiable at $y_i - \mathbf{u}^*\mathbf{x}_i = 0$, it is not guaranteed that the solution path is piecewise linear in λ. Nonetheless, since the objective function in this case is convex, we can still solve it with an unconstrained convex optimization procedure. Here for a fair comparison, we used `CVX`, a package for specifying and solving convex programs [62]. The underlying model we assume is as follows:

$$y_i = \mathbf{x}_i^*\boldsymbol{\beta} + e_i, \tag{2.20}$$

where $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^*$. X is realized from a Gaussian random matrix with zero mean and unit variance. So we have $C = I_{8\times 8}$. The errors are generated based on two different mechanisms, with more details given shortly. The intercept term is not considered since it can always be estimated by the mean of $\mathbf{y}$. Therefore, the response $\mathbf{y}$ is centered before applying any shrinkage model. The shrinkage tuning parameter λ_n is chosen to be $n^{1/3}$ for all Lasso and robust Lasso models such that they have both parameter estimation and model selection consistency for adaptive weight $\hat{\mathbf{w}} = 1/|\hat{\boldsymbol{\beta}}_n^{LS}|^\gamma$ with $\gamma = 1$. Note that the shrinkage sequence chosen here is not universal and optimal in terms of prediction. It is used merely to demonstrate the validity of the derived theory. Practical determination of $\{\lambda_n\}$ is usually by the BIC, cross-validation procedure, etc. The theoretic variances of asymptotic normality can be numerically computed. We set $\delta = 0.1$ for the convex combined loss, and $\delta = 1$ for the Huber loss. All simulations are averaged and reported over 100 simulated date sets, each of which contains $n = 1,000$ data points. The following two error distributions are considered:

Error distribution	Model	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
Gaussian Mixture $\mu = 5, \sigma = 1$	Lasso	13.50 (13.84)	13.50 (8.74)	13.50 (11.43)	13.50 (13.94)	13.50 (12.22)	13.50 (14.12)	13.50 (11.87)	13.50 (16.00)
	RLASSO	7.66 (8.05)	7.66 (6.44)	7.66 (5.71)	7.66 (5.84)	7.66 (6.94)	7.66 (6.71)	7.66 (4.80)	7.66 (7.41)
	HLASSO	5.65 (7.03)	5.65 (5.61)	5.65 (4.99)	5.65 (5.56)	5.65 (5.87)	5.65 (6.39)	5.65 (4.10)	5.65 (5.79)
	adaLASSO	13.5 (14.52)	13.5 (13.81)	0 (5.33)	0 (8.32)	13.5 (13.71)	0 (5.50)	0 (9.63)	0 (11.15)
	adaRLASSO	7.66 (8.02)	7.66 (7.94)	0 (0.51)	0 (1.28)	7.66 (9.19)	0 (0.54)	0 (1.75)	0 (1.76)
	adaHLASSO	5.65 (6.57)	5.65 (6.57)	0 (0.58)	0 (1.13)	5.65 (7.95)	0 (0.54)	0 (1.35)	0 (1.94)
Student-t $\nu = 3, \sigma = 3$	Lasso	27 (28.77)	27 (28.13)	27 (26.50)	27 (26.87)	27 (30.53)	27 (25.52)	27 (23.58)	27 (23.96)
	RLASSO	12.87 (17.27)	12.87 (17.54)	12.87 (15.27)	12.87 (14.64)	12.87 (20.33)	12.87 (13.25)	12.87 (13.15)	12.87 (11.85)
	HLASSO	13.45 (14.92)	13.45 (15.08)	13.45 (12.42)	13.45 (11.06)	13.45 (16.43)	13.45 (11.88)	13.45 (13.08)	13.45 (10.00)
	adaLASSO	27 (28.46)	27 (27.22)	0 (19.96)	0 (23.74)	27 (25.89)	0 (18.38)	0 (14.90)	0 (15.48)
	adaRLASSO	12.87 (19.59)	12.87 (16.19)	0 (4.59)	0 (4.98)	12.87 (16.25)	0 (3.05)	0 (2.36)	0 (2.61)
	adaHLASSO	13.45 (16.08)	13.45 (12.95)	0 (3.31)	0 (2.94)	13.45 (13.89)	0 (1.89)	0 (1.35)	0 (1.98)

Table 2.1: Theoretic and empirical (shown in the parentheses) asymptotic variances of $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ for the mixture of Gaussian and Student-t error distributions, respectively. RLASSO is the robust Lasso with the convex combined loss function and HLASSO is the Huberized Lasso. The adaLASSO, adaRLASSO, and adaHLASSO are the corresponding adaptive versions. β_1 to β_8 are the coefficients of eight predictors.

1. *Symmetric Gaussian mixture with three components.* The errors are simulated from a Gaussian mixture distribution with symmetric two-side outliers, i.e. the error is assumed to have the following p.d.f.

$$f(e_i) = \frac{1}{4}N(e_i; -\mu, \sigma^2) + \frac{1}{2}N(e_i; 0, \sigma^2) + \frac{1}{4}N(e_i; \mu, \sigma^2), \tag{2.21}$$

where $N(e_i; \mu, \sigma^2)$ denotes the p.d.f. of a normal random variable with mean μ and variance σ^2. It is clear that f satisfies the error assumption. The results are reported for $\mu = 5$ and $\sigma = 1$. Figure 2.1 shows the histograms of $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ for the non-adaptive models and Figure 2.2 shows those of adaptive models. The theoretic and empirical variances of the limiting distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ are shown in Table 2.1. Several conclusions can be drawn here:

(a) The variances of $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ based on simulations are quite close to the theoretic asymptotic variances (see Figure 2.1).

(b) The variances of the scaled convex combined robust Lasso and the Huberized Lasso estimators are smaller than that of the Lasso estimator, as expected (see Table 2.1).

(c) Although the adaptive Lasso has been proved to be model selection consistent, the simulation study shows that the adaptive Lasso performs poorly when the noise variance σ^2 is large, at least for a relatively large sample size $n = 1,000$. In contrast, the two adaptive robust Lassos show significant performance improvements over the ordinary Lasso (see Figure 2.2 for the zero coefficients).

(d) Based on the simulations results, it is observed that the non-adaptive (robust) Lassos do not seem to be model selection consistent even though the irrepresentable condition of [130] is met in our simulation setup. Closer examination reveals that the particular shrinkage sequence we chose is not a model selection consistency one for the non-adaptive cases because

$n^{-1/6-c} \nrightarrow \infty$ for any $1 > c \geq 0$ as given by Theorem 1 in [130].

2. *Student-t errors with heavy tails.* The setup is the same as the Gaussian mixture case, except that the errors are generated from a Student-t distribution with the degree of freedom $\nu = 3$ and $\sigma = 3$. The theoretic and empirical values of the variances of the asymptotic distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ are given in Table 2.1. Based on the histogram results, same observations can be noted as in the Gaussian mixture case. Therefore, due to the space concern, we do not report the figures here.

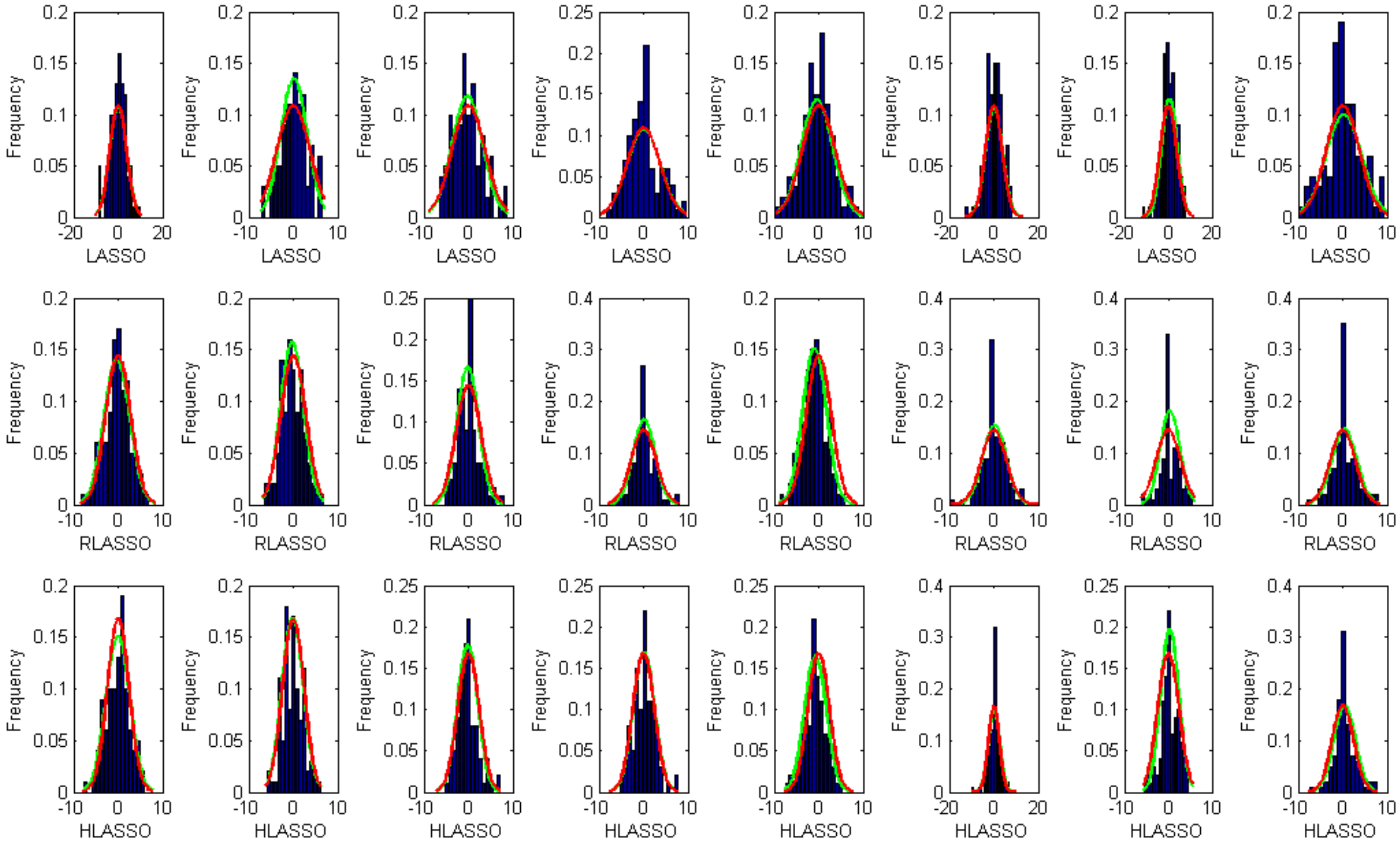


Figure 2.1: Histograms of $\arg\min(V_n) = \sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ for the Gaussian mixture with $\mu = 5$. Green curve is the fitted normal distribution to the estimated values of $\arg\min(V_n)$ from data over 100 simulations. Red curve is its theoretic asymptotic normal distribution of $\arg\min(V)$. Three rows represent the Lasso, convex combined robust Lasso, and Huberized Lasso models in the order of top-down. Columns represent the eight predictors in the order of $(3, 1.5, 0, 0, 2, 0, 0, 0)$.

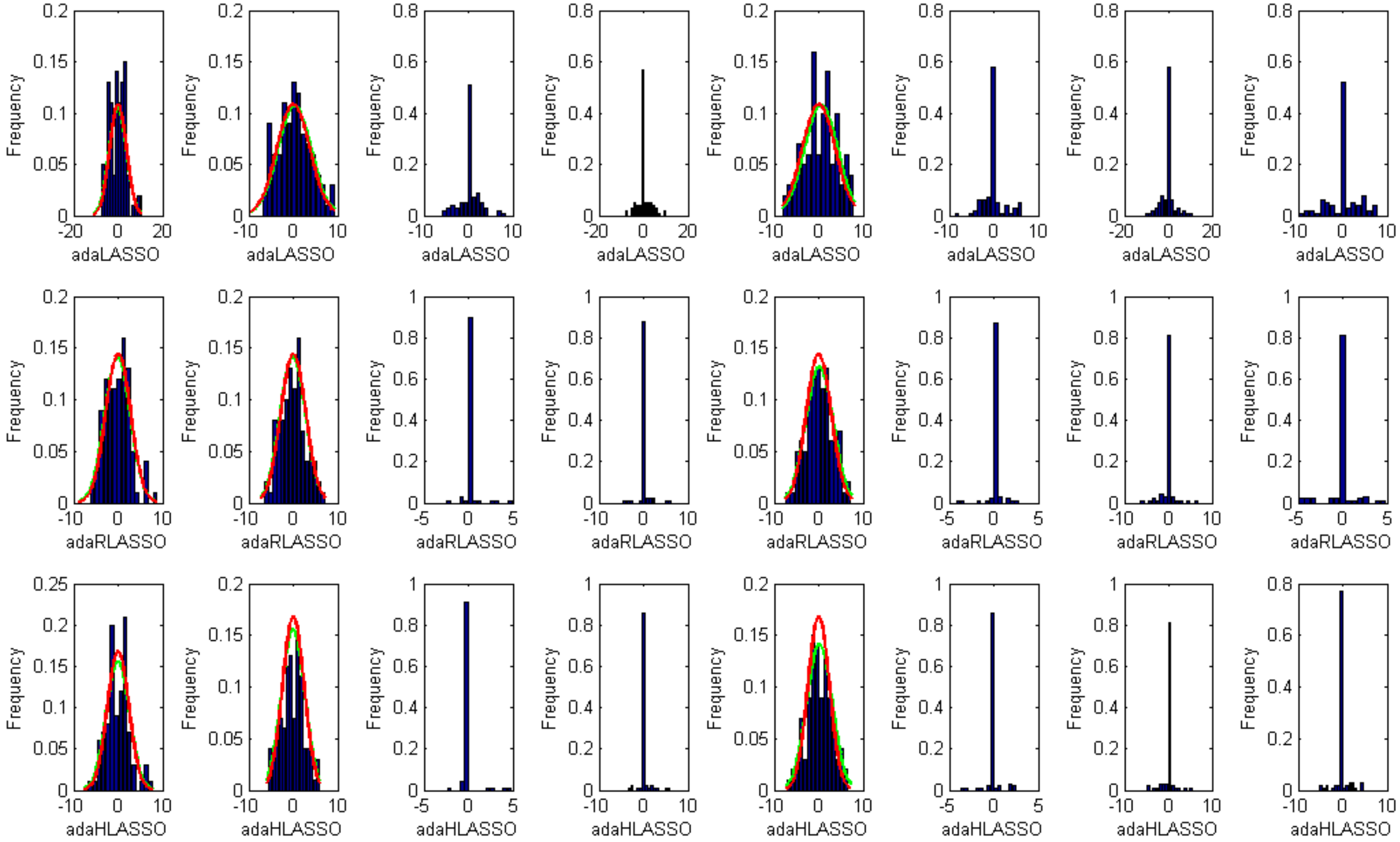


Figure 2.2: Histograms of $\arg\min(V_n) = \sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ for the Gaussian mixture with $\mu = 5$ with adaptation. Green curve is the fitted normal distribution to the estimated values of $\arg\min(V_n)$ from data over 100 simulations. Red curve is its theoretic asymptotic normal distribution of $\arg\min(V)$. Three rows represent the Lasso, convex combined robust Lasso, and Huberized Lasso models in the order of top-down. Columns represent the eight predictors in the order of $(3, 1.5, 0, 0, 2, 0, 0, 0)$.

2.7 Conclusion and Discussion

In the presence of noise with large variance, the standard Lasso may behave poorly in estimating the true regression coefficients. We propose a flexible, robust version of the Lasso, which combines the advantages of both the ℓ_1 and ℓ_2 losses. The asymptotic normality and model selection consistency are established at certain shrinkage rates. The limiting behavior of the Huberized Lasso, another robust Lasso, is also studied. Analysis results derived from the non-stochastic design case are extended to the random design, for which auto-regression models can be suitably handled.

The asymptotic analysis framework presented in this chapter provides an appropriate starting point for future, more general analysis on such robust models, and we hope that the current finite-dimensional asymptotic results can provide certain implication and insight into more challenging settings. For instance, the asymptotic analysis could shed light into non-asymptotic analysis since finite sample size results can be closely related to the asymptotic ones. We have derived a finite sample size ℓ_2 estimation error bound for the noiseless case (Proposition 2.2.5) where in fact p is allowed to be greater than n as long as the "incoherent design" condition is valid, and its conclusion agrees with that of the finite-dimensional asymptotic analysis. A future direction is to derive non-asymptotic estimation error bounds in the presence of noise for the situations where $p > n$ and $p \to \infty$. This is challenging, since analyzing the penalized robust losses is much more complicated than the regular Lasso. However, those error bounds can provide great insights into the finite sample size behavior of the robust estimator and could be useful to many research areas such as compressed sensing.

As mentioned in the beginning of this chapter, the motivating biomedical application of this work is brain effective connectivity modeling using fMRI data. Since biomedical research is usually conducted at a group level, means to address within-group, inter-subject variability are required. The proposed robust Lassos can be easily extended to group versions by minimizing the summation of block Euclidean norms objective/regularized function [48, 128]. Therefore, sparse features can be learned at the group level and we have already shown some promising preliminary group analysis results regarding brain connectivity in fMRI [35]. The results are presented in Chapter 6.

Chapter 3

A Bayesian Lasso via Reversible-Jump MCMC

3.1 Introduction

3.1.1 Sparse Linear Models

This chapter considers the same multivariate linear regression model (2.1) in Chapter 2. Here, we give further motivations of the present chapter, which aims to obtain more stable estimates in a fully Bayesian paradigm.

As we have seen in Chapter 1, there have been many variable/model selection methods proposed in the literature from both frequentist and Bayesian perspectives, with a good overview given in [74, 106]. With such a wealth of methods, it is difficult to argue which model is universally preferable. Among different methods, Bayesian approaches using Markov chain Monte Carlo (MCMC) have recently become popular [40]. For instance, the "spike and slab" priors on the regression coefficients were proposed in [98]. Using the Laplace or Bernoulli-Gaussian mixture prior on the regression coefficients and introducing latent variables to identify subsets have been the popular choice. Assuming a hierarchical Bayes Gaussian mixture model with latent variables to identify subsets, a Gibbs sampling approach was presented in [58]. The work in [58] was further explored in [79] by embedding the priors jointly. Several MCMC methods have been compared in [40] for selecting the regression coefficients. In this paper, we are interested in the general category of MCMC-based Bayesian approaches; however, as motivated by the success of Lasso model which is to be discussed shortly, we will employ a Bayesian model different from the above mixture models and propose a fully Bayesian Lasso framework with the RJ-MCMC approach (not the regular MCMC approaches as in the above work).

The penalized likelihood approach in (1.2) has an alternative Bayesian interpretation. As noted in [115], Lasso estimates can be interpreted as *Maximum A Posteriori* (MAP) estimates with the regression coefficients possessing independent Laplace (a.k.a. double-exponential) priors. More recently, [70] proposed a more general optimization approach, the Minorize-Maximize (MM) algorithm [81], to transfer the problem of maximizing the posterior function w.r.t. the Laplace prior to sequentially maximizing its quadratic surrogate functions. Motivated by this connection between Lasso estimates and the Bayesian interpretation for the Laplace prior, several Laplace-like priors have been recently proposed for promoting sparsity, e.g. a mixture of delta-mass at 0 and the Laplace prior was studied in [127] and Jeffrey's non-informative mixing distribution on the prior of $\boldsymbol{\beta}$ in [52]. The popularity and the good performance of the Lasso model motivates us to employ a Laplace prior for the regression coefficients in our proposed RJ-MCMC based Bayesian approach, a Bayesian Lasso estimator. The observations in [102] actually suggested potential advantages of the Laplace prior over a Gaussian (or a Student-t) prior.

3.1.2 Related Work and Our Contributions

It must be emphasized that the non-Bayesian Lasso and Lasso-like approaches have one aspect in common: they are optimization methods with the goal of determining the model parameters that maximize some objective function. Meanwhile, the number of variables set to be zero in these methods critically depends on the tuning shrinkage parameter, where its value can generally be selected to minimize the generalized cross-validation errors. In this paper, apart from those Lasso-like optimization methods, we propose a new fully Bayesian framework to deal with the Lasso objective function. Such a fully Bayesian approach with the Laplace prior on $\boldsymbol{\beta}$, referred as *a Bayesian Lasso*, does not require cross-validation (CV) type methods to determine the optimum shrinkage parameter as in Lasso, since a non-informative prior is given for the shrinkage controlling parameter and its posterior distribution is completely derived from the observed data. By integrating parameters w.r.t. their posterior distributions, the proposed Bayesian Lasso estimator has a different posterior distribution from the ordinary Lasso (Laplace prior), and can yield more robust estimates.

Very recently, work has been proposed in the direction of Bayesian Lasso [102]. In [102], with a conditional Gaussian prior on $\boldsymbol{\beta}$ and the non-informative scale-invariant prior on the noise variance being assumed, a Bayesian Lasso model is proposed and a simple Gibbs sampler is implemented. It is shown that the Bayesian Lasso estimates in [102] are strikingly similar to those from the ordinary Lasso. Since this Bayesian Lasso in [102] involves the inversion of the covariance matrix of a block coefficients at each iteration, the computationally complexity prevents its practical application with, say, hundreds of variables. Moreover, similar to the regular Lasso, the Bayesian Lasso in [102] uses only one shrinkage parameter t to both control model size and shrink estimates. Nonetheless, it is arguable whether the two simultaneous effects can be well-handled by a single tuning parameter [94]. To mitigate this non-separability problem, [99] proposed an extended Bayesian Lasso model by assigning a more flexible, covariate-adaptive penalization on top of the Bayesian Lasso in the context of Quantitative Trait Loci (QTL) mapping. Alternatively, introducing different sources of sparsity-promoting priors on both coefficients and their indicator variables have been studied, e.g. in [104], where a normal-Jeffery scaled-mixture prior on coefficients and an independent Bernoulli prior with small success probability on the binary index vector are combined. Motivated by this observation, we introduce two parameters in the proposed Bayesian Lasso model to separately control the model selection and estimation shrinkage issues in the spirit of [94] and [127], and propose a Poisson prior on the model size together with the Laplace prior on $\boldsymbol{\beta}$ to identify the sparsity pattern. Since the proposed joint posterior distribution is highly nonstandard and a standard MCMC is not applicable, we employ a reversible-jump MCMC (RJ-MCMC) to obtain the proposed Bayesian Lasso estimates by simultaneously performing model averaging and parameter estimation. It is worth emphasizing that, though RJ-MCMC algorithms have been developed in the literature before model selection and estimation purposes (e.g. [4] proposed a hierarchical Bayesian model and developed an RJ-MCMC algorithm for joint Bayesian model selection and estimation of noisy sinusoids; similarly [111] proposed an accelerated truncated Poisson process model for Bayesian QTL mapping), these methods are not intended for promoting sparse models whereas our model utilizes sparsity promoting priors in conjunction with the discrete prior on the model size.

As we show later, the proposed fully Bayesian Lasso framework provides estimation performance improvements when compared with Lasso, the Gibbs sampler-based Bayesian Lasso in [102] and the Binomial-Gaussian model in [59]. When handling the nearly singular case ($p \approx n$), the performance improvement from the proposed Bayesian Lasso estimate is even more significant. As a side benefit, we also extend the proposed RJ-MCMC estimation framework to the Binomial-Gaussian model in [59], and the developed BG-MCMC approach yields significant performance improvements over the original non-Bayesian approach in [59].

3.1.3 Notations

We now define some notations to be used throughout the chapter. Let $\boldsymbol{\gamma}$ be a p-length binary vector where ones denote non-zero coefficients and zeros denote zero coefficients. Equivalently, position of ones in $\boldsymbol{\gamma}$ can be thought as the *active set* or *support* of a linear regression model while position of zeros is called the *inactive set.* $|\boldsymbol{\gamma}|$ is used to denote the number of non-zeros in $\boldsymbol{\gamma}$, meaning the cardinality of the support of $\boldsymbol{\gamma}$. Some special functions and probability density functions are listed in Table 3.1.

Table 3.1: Special functions and probability density functions used in the chapter.

Name	Functional form
Gamma function	$\Gamma(a) = \int_0^\infty t^{a-1} \exp(-t)\, dt$
Beta function	$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$
Gamma distribution	$\mathrm{Ga}(x; a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)$
Inverse Gamma distribution	$\mathrm{IG}(x; a, b) = \frac{b^a}{\Gamma(a)} x^{-(a+1)} \exp(-\frac{b}{x})$
Beta distribution	$\mathrm{Beta}(x; a, b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$

The rest of the chapter is organized as follows. In Section 3.2, we first describe the new hierarchical, fully Bayesian Lasso model, and then propose an RJ-MCMC algorithm to simulate this posterior distribution for computing the unbiased minimum variance estimator of the regression coefficient vector. Simulations are carried out in Section 3.4 to evaluate the performance of the proposed approach. Section 3.5 presents the results on a diabetes data set. In Section 3.6, we apply the proposed method to real fMRI data to demonstrate the applicability of the proposed Bayesian Lasso method.

3.2 A Fully Bayesian Lasso Model

3.2.1 Prior Specification

The proposed fully Bayesian model has the basic structure of a standard linear regression model in (2.1), with the addition of priors over the parameters to be estimated. Objective hyper-priors are chosen on the sparsity promoting priors.

Firstly, to achieve the parsimonious estimation goal, we assign sparsity promoting priors on each of the non-zero coefficients β_j's. Here, independent Laplace priors are assumed for $\boldsymbol{\beta}$, i.e. each component β_j in the active set $\boldsymbol{\gamma}$ with $|\boldsymbol{\gamma}| = k$ follows the distribution

$$\pi(\beta_j|\tau, \boldsymbol{\gamma}) = \frac{1}{2\tau} \exp\left(-\frac{|\beta_j|}{\tau}\right), \tag{3.1}$$

where τ is a shrinkage tuning parameter for β_j with $j \in \boldsymbol{\gamma}$. Otherwise, $\beta_j \equiv 0$ for $j \notin \boldsymbol{\gamma}$.

Since in practice we may not have prior knowledge regarding how much shrinkage amount should be put on the coefficients before performing any experiment, it is reasonable to also assign a non-informative prior on the hyper-parameter τ. A non-informative prior for the noise variance σ^2 is given based on the same rationale. Together, traditional non-informative priors are put on the higher level as:

$$p(\tau, \sigma^2) \propto (\tau\sigma)^{-1}. \tag{3.2}$$

Due to Lindley's paradox [89], one needs to be very careful to assign improper priors for τ and σ^2 when performing model selection/averaging. Since we do not allow the null model with no predictor, τ and σ^2 are the common parameters for all possible sub-models. Hence, the underdetermined proportional constants are the same for all sub-models and therefore do not affect the model comparisons based on the posteriors.

Further, prior probabilities are also assigned to each possible sub-model. Specifically, a sub-model containing k predictors follows a right-truncated Poisson distribution at p,

$$p(k|\lambda) = \frac{e^{-\lambda}\lambda^k}{Ck!},$$

where C is a normalization constant, and $k = 1, ..., p$. Within the set of sub-models having k predictors, each sub-model is assumed to be with equal prior probability. To complete the prior specification, the parameter λ is also assumed to follow a non-informative prior as

$$p(\lambda) \propto \lambda^{-1}.$$

Therefore, the posterior distribution of the parameter vector $(\boldsymbol{\beta}, \tau, \sigma^2, \boldsymbol{\gamma}, \lambda)$ can be expressed as (up to a multiplicative constant)

$$p(\boldsymbol{\beta}, \tau, \sigma^2, \boldsymbol{\gamma}, \lambda|\mathbf{y}, X) \propto \frac{e^{-\lambda}\lambda^{k-1}}{\binom{p}{k}k!}\tau^{-(k+1)}\sigma^{-(n+1)} \exp\left(-\frac{||\boldsymbol{\beta}||_1}{\tau} - \frac{||\mathbf{y} - X\boldsymbol{\beta}||_2^2}{2\sigma^2}\right). \tag{3.3}$$

The derivation detail of (3.3) can be found in the Appendix A.3.1. It is further noted that, to reduce the computational cost, the parameters $\tau, \sigma^2, \text{and}, \lambda$ can be analytically integrated out. More specifically, by noting the normalization constants of the respective Gamma and inverse Gamma p.d.f.'s (Table 3.1), we can show that

$$\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}|\mathbf{y}, X) \propto \Gamma(k)B(k, p-k+1) \left\|\boldsymbol{\beta}\right\|_1^{-k} \left\|\mathbf{y} - X\boldsymbol{\beta}\right\|_2^{-(n-1)}. \tag{3.4}$$

Comparing (3.4) with the posterior distribution of the standard Lasso reveals interesting observations. (3.4) comprises two parts: $\left\|\boldsymbol{\beta}\right\|_1^{-k}$ is the prior information and $\left\|\mathbf{y} - X\boldsymbol{\beta}\right\|_2^{-(n-1)}$ is the likelihood. These two parts are linked through polynomial terms, representing a different weighting scheme from Lasso. By integrating out the nuisance parameters analytically, a more stable estimator is possible, in addition to the advantage of requiring fewer computations.

It is obvious that this joint posterior distribution is in a non-standard form and there is no closed-form analytic expression of $\mathbb{E}(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X})$ w.r.t. its posterior distribution. Hence, we must resort to simulation-based approaches to compute the numeric estimates. In this paper, we will develop a Markov chain Monte Carlo (MCMC) type of estimation method.

Remark 8. For the standard Lasso approach, there is one shrinkage parameter t in (1.2) to both control model size and shrink estimates, but for some applications, it is desirable to separate those two effects. Here, in the proposed framework, we have two parameters (i.e. τ and λ) to separately control the model selection and estimation shrinkage issues. Roughly speaking, a Poisson prior on the size of active set k controls the size of the expected number of selected predictors and a Laplace prior on $\boldsymbol{\beta}$ recovers the non-zero coefficients which can best represent the full model conditioned on k.

Algorithm 1: The proposed RJ-MCMC based Bayesian Lasso

```
Input: The number of iterations $T$. Random walk step size $\epsilon$.
Data: $X$ and $\mathbf{y}$.
Output: $\left\{\boldsymbol{\theta}^{(t)} = (\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)}) \middle| t \in \{0, \cdots, T\}\right\}$.
begin
    Initialization: set $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\beta}^{(0)}, \boldsymbol{\gamma}^{(0)})$ and $t = 1$.
    repeat
        if $k^{(t-1)} = 1$ then
            $k^{(t)} \leftarrow k^{(t-1)} + U(\{0, 1\})$[3].
        else if $k^{(t-1)} = p$ then
            $k^{(t)} \leftarrow k^{(t-1)} - U(\{0, 1\})$.
        else
            $k^{(t)} \leftarrow k^{(t-1)} + U(\{-1, 0, 1\})$.
        end
        Sample $s \sim N(0, \epsilon^2)$.
        $K \leftarrow \boldsymbol{\gamma}^{(t-1)}$ and $K^c \leftarrow \{1, \cdots, p\} \setminus K$.
        if $k^{(t-1)} = k^{(t)}$ then
            Sample $j \sim U(K)$.
            Update $\beta_j^{(t)} \leftarrow \beta_j^{(t-1)} + s$ with an MH step, details in Section 3.3.2.
        else if $k^{(t)} = k^{(t-1)} + 1$ then
            Sample $j \sim U(K^c)$.
            Perform a "birth" move and update $\beta_j^{(t-1)}$, details in Section 3.3.3.
        else
            Sample $j \sim U(K)$.
            Perform a "death" move and update $\beta_j^{(t-1)}$, details in Section 3.3.3.
        end
        $t \leftarrow t + 1$.
    until $t = T$.
end
```

3.3 Bayesian Computation

Since the joint posterior distribution in (3.4) contains both discrete and continuous parameters, a closed-form solution of the unbiased minimum variance estimator is infeasible: the posterior expectation of coefficients $\mathbb{E}(\boldsymbol{\beta}|X, \mathbf{y})$ and the posterior probability of the inclusion of coefficients $\mathbb{E}(\boldsymbol{\gamma}|X, \mathbf{y})$. Moreover, standard MCMC is not applicable in this case since the model dimension is not fixed. To address this difficulty, we propose a hybridized MCMC sampler to simultaneously perform model av-

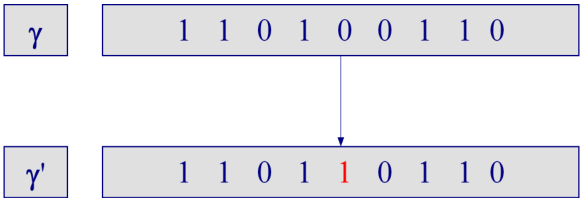


Figure 3.1: An illustration of model jumping from $\boldsymbol{\gamma} \to \boldsymbol{\gamma}'$ with $|\boldsymbol{\gamma}| = 5$ and $|\boldsymbol{\gamma}'| = 6$. New predictors (in red) are created from current model. A position with 1 indicates a non-zero coefficient, 0 denotes current model excludes this coefficient.

eraging and parameter estimation. Our proposed algorithm falls into the RJ-MCMC umbrella [63]. RJ-MCMC is a powerful prototype that creates MCMC algorithms for variable dimensional models and may be better than separate within-model MCMC runs if we aim at making joint inference about the models and their parameters. Moreover, running separate MCMC for each model is computationally prohibitive for large scale problems. The proposed algorithm is summarized in Algorithm 1, with more details given in the following paragraphs.

3.3.1 Design of Model Transition

It is clear that the distribution of $\boldsymbol{\beta}$ depends on the model dimensionality. For example, deleting predictors will force their corresponding index set to zeros. Here, we propose three types of model moves as following:

$$1.\ \boldsymbol{\gamma} \to \boldsymbol{\gamma}; \quad 2.\ \boldsymbol{\gamma} \to \boldsymbol{\gamma}'; \quad 3.\ \boldsymbol{\gamma}' \to \boldsymbol{\gamma}.$$

To smoothly move between models and allow fast mixing, we design local jumps. At each sampling step, the current model is only allowed to stay in the same dimension or move to its neighboring models. As illustrated in Figure 3.1, the proposed model has either the same dimension as the previous one, or one predictor added or deleted from the current model. Moreover, we assign each possibility with equal probabilities, as

$$\begin{aligned} p(\boldsymbol{\gamma} \to \boldsymbol{\gamma}) &= \frac{1}{3}, \\ p(\boldsymbol{\gamma} \to \boldsymbol{\gamma}') &= \frac{1}{3(p-k)}, \\ p(\boldsymbol{\gamma}' \to \boldsymbol{\gamma}) &= \frac{1}{3(k+1)}, \end{aligned}$$

for $k = 2, \cdots, p-2$ and $|\boldsymbol{\gamma}| = k$ and $|\boldsymbol{\gamma}'| = k+1$. The boundary models need slightly different probabilities. For $|\boldsymbol{\gamma}| = 1$, we do not allow the null model with no predictor at all. Hence for a model with just one predictor, it can only stay in one-dimensional or move to two-dimensional, each with probability $\frac{1}{2}$. Namely, for $|\boldsymbol{\gamma}| = 1$ and $|\boldsymbol{\gamma}'| = 2$

$$\begin{aligned} p(\boldsymbol{\gamma} \to \boldsymbol{\gamma}) &= \frac{1}{2}, \\ p(\boldsymbol{\gamma} \to \boldsymbol{\gamma}') &= \frac{1}{2(p-1)}. \end{aligned}$$

Similarly, for $|\boldsymbol{\gamma}| = p$ and $|\boldsymbol{\gamma}'| = p-1$, we have

$$\begin{aligned} p(\boldsymbol{\gamma} \to \boldsymbol{\gamma}) &= \frac{1}{2}, \\ p(\boldsymbol{\gamma} \to \boldsymbol{\gamma}') &= \frac{1}{2p}. \end{aligned}$$

3.3.2 A Usual Metropolis-Hastings Update for Unchanged Model Dimension

For the models with unchanged dimension, the standard Metropolis-Hastings (MH) algorithm is used to update $\boldsymbol{\beta}$ [67]. Specifically at iteration t, a predictor at position $j \in \boldsymbol{\gamma}$ to be updated is randomly selected from current non-zero coefficients. Then, a proposal distribution $q(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta'})$ is chosen to update this predictor, where $\boldsymbol{\theta}^{(t)}$ is current parameter estimate and $\boldsymbol{\theta'}$ is the proposed parameter. Here, a Gaussian random walk (RW) is used as our proposal, $N(0, \epsilon^2)$, with some fixed small step size ϵ. Set $\boldsymbol{\beta'} = \boldsymbol{\beta} + u\mathbf{e}_j$ where $\mathbf{e}_j$ is the j^{th} standard Euclidean basis. Then the acceptance probability becomes

$$\min\left\{\left(\frac{\|\boldsymbol{\beta}'\|_1}{\|\boldsymbol{\beta}\|_1}\right)^{-k}\times\left(\frac{\|\mathbf{y}-X\boldsymbol{\beta}'\|_2}{\|\mathbf{y}-X\boldsymbol{\beta}\|_2}\right)^{-(n-1)},1\right\}. \tag{3.5}$$

3.3.3 A Birth-and-Death Strategy for Changed Model Dimension

Since there is no concept of metric structure and compactness as in the Euclidean space for trans-dimensional jumps in $\boldsymbol{\Theta}$, designing an optimal or even a valid proposal is not an easy task. The standard MCMC optimal *scaling* proposal has no analogue for reversible jump moves [15]. For the model jumping moves, there are two commonly used proposals: i) *birth-and-death* and ii) *split-and-merge*. The birth-and-death is a simple form of model transformation: In the birth step, a new predictor is added to the current model, by generating parameters of a new predictor from a prior distribution; in the death step, a predictor is removed from current model, and the reversibility constraint must be satisfied according to the *detailed balance* equation. [63] showed that if there exists a σ-finite symmetric measure μ, with respect to which $\pi(d\boldsymbol{\theta})q(\boldsymbol{\theta}, d\boldsymbol{\theta'})$ is absolutely continuous with π is our target posterior distribution in (3.4), then the detailed balance condition holds for all Borel subsets $B, B' \subset \mathcal{B}(\boldsymbol{\Theta})$,

$$\int_{B\times B'}\alpha(\boldsymbol{\theta},\boldsymbol{\theta'})\rho(\boldsymbol{\theta},\boldsymbol{\theta'})\mu(d\boldsymbol{\theta},d\boldsymbol{\theta'}) = \int_{B'\times B}\alpha(\boldsymbol{\theta'},\boldsymbol{\theta})\rho(\boldsymbol{\theta'},\boldsymbol{\theta})\mu(d\boldsymbol{\theta'},d\boldsymbol{\theta}), \tag{3.6}$$

if the acceptance ratio $\alpha(\boldsymbol{\theta},\boldsymbol{\theta'})$ is chosen to be

$$\alpha(\boldsymbol{\theta},\boldsymbol{\theta'}) = \frac{\rho(\boldsymbol{\theta'},\boldsymbol{\theta})}{\rho(\boldsymbol{\theta},\boldsymbol{\theta'})}\wedge 1, \tag{3.7}$$

where the extended μ-integrable function ρ is the Radon-Nikodym derivative of $\pi \times q$ with respect to μ. It is easy to see that the unchanged model dimension update $\boldsymbol{\gamma} \rightarrow \boldsymbol{\gamma}$ is just a special case by taking μ as the Lebesgue measure on $\mathcal{B}(\mathbb{R}^k)\otimes\mathcal{B}(\mathbb{R}^k)$, the product Borel σ-algebra on $\mathbb{R}^k$.

For a birth move, a new predictor is created by random generation from the inactive set. The proposed predictors are accepted with the probability given by the *generalized Metropolis-Hasting ratio*. More specifically, a coefficient is randomly generated outside the current support set (i.e. generating a $j \in \boldsymbol{\gamma}^c$ and setting the value of this coefficient with a zero mean Gaussian realization u). Since the model dimension is augmented by generating an additional variable u, there is a Jacobian term for the acceptance probability of the birth move which is 1 in this case.

By putting the posterior ratio computed from (3.4) and model transition probabilities together into (3.7), the acceptance probability is given by

$$\min\left\{\frac{k^2}{p-k} \times \frac{\|\boldsymbol{\beta}'\|_1^{-(k+1)}\,\|\mathbf{y}-X\boldsymbol{\beta}'\|_2^{-(n-1)}}{\|\boldsymbol{\beta}\|_1^{-k}\,\|\mathbf{y}-X\boldsymbol{\beta}\|_2^{-(n-1)}} \times \frac{p(\boldsymbol{\gamma}' \to \boldsymbol{\gamma})}{p(\boldsymbol{\gamma} \to \boldsymbol{\gamma}')} \times N(u;0,\epsilon^2)^{-1}, 1\right\} \tag{3.8}$$

where $N(u;\mu,\epsilon^2)$ means the Gaussian density $N(\mu,\epsilon^2)$ evaluated at u.

Similarly, the death move is simply the reverse of the birth move. The acceptance probability for $\boldsymbol{\gamma} \to \boldsymbol{\gamma}'$ with $|\boldsymbol{\gamma}| = k$ and $|\boldsymbol{\gamma}'| = k-1$ is given by

$$\min\left\{\frac{p-(k-1)}{(k-1)^2} \times \frac{\|\boldsymbol{\beta}'\|_1^{-(k-1)}\,\|\mathbf{y}-X\boldsymbol{\beta}'\|_2^{-(n-1)}}{\|\boldsymbol{\beta}\|_1^{-k}\,\|\mathbf{y}-X\boldsymbol{\beta}\|_2^{-(n-1)}} \times \frac{p(\boldsymbol{\gamma}' \to \boldsymbol{\gamma})}{p(\boldsymbol{\gamma} \to \boldsymbol{\gamma}')} \times N(u;0,\epsilon^2), 1\right\}. \tag{3.9}$$

(3.8) and (3.9) ensure the reversibility of the constructed Markov chain by (3.6).

Remark 9. If one is also interested in the nuisance parameters τ, σ^2 and λ, it is easy to extend the current algorithm to include embedded Gibbs samplers. Since the full conditional distributions of τ, σ^2, and λ can be given in closed forms, the Gibbs sampler [61] is used to simulate these parameters from their posterior distributions,

$$\begin{aligned} \tau|\cdot &\sim \mathrm{IG}\left(k, \sum_{j\in\boldsymbol{\gamma}} |\beta_j|\right) && (3.10)\\ \sigma^2|\cdot &\sim \mathrm{IG}\left(\frac{n-1}{2}, \frac{\|\mathbf{y}-X\boldsymbol{\beta}\|_2^2}{2}\right) && (3.11)\\ \lambda|\cdot &\sim \mathrm{Ga}(k,1). && (3.12) \end{aligned}$$

Here $|\cdot$ means the conditional distribution given everything else, including data and other parameters.

3.4 Simulations

3.4.1 Setup

Data sets of 100 data points are simulated, with a series of different model dimensions. The value of p is set to be 15, 45, and 90. Each setting has a specific sparse structure of the coefficients, and we denote a particular setting by a format such as 30/90, meaning that 30 out of the total 90 predictor coefficients are non-zeros. For instance, for the 90 dimensional case, we set the coefficients to be

$$(\underbrace{3,\cdots,3}_{\text{10 times}},\underbrace{0,\cdots,0}_{\text{20 times}},\underbrace{1.5,\cdots,1.5}_{\text{10 times}},\underbrace{0,\cdots,0}_{\text{20 times}},\underbrace{2,\cdots,2}_{\text{10 times}},\underbrace{0,\cdots,0}_{\text{20 times}}).$$

Although independent sparsity promoting priors are used (3.1), correlations between predictors are also introduced and compared with uncorrelated data and their influence on the performance of various algorithms is explored. We set the correlation level to be 0.5 in our simulations. Several standard linear model selection approaches are compared, including the Lasso [115] (and its variant *Gauss-Lasso* for extending Lasso to accommodate both model selection and regression objectives, where Lasso is used as model selection followed by a least squares estimate based on the selected model), Lar [45] (and its variant *Gauss-Lar*, where Lar is used as model selection followed by a least squares estimate based on the selected model), a Gibbs sampler based Bayesian Lasso [102] and the Binomial-Gaussian (BG) model [59]. To examine the effects on the overall performance of the proposed Poisson-Laplace model and the proposed MCMC estimation approach, we also extend the BG model in [59] to the fully Bayesian framework, and develop a corresponding MCMC algorithm, referred as *BG-MCMC*, with details given in Appendix A.3.2. We shall see that the proposed

Table 3.2: RMSEs averaged over 100 simulations for $n = 100$. The number in the bracket is the standard deviations of the estimated RMSEs. Methods under comparison are the Lasso [115], Gauss-Lasso, Lar [45], Gauss-Lar, Gibbs sampler based Bayesian Lasso [102], Binomial-Gaussian (BG) [59], proposed BG-MCMC and proposed Bayesian Lasso (BLasso) with RJ-MCMC algorithm.

	No correlation			**Correlation=0.5**		
	3/15	15/45	30/90	3/15	15/45	30/90
Lasso [115]	0.728 (0.146)	1.023 (0.158)	7.968 (2.441)	0.561 (0.116)	0.859 (0.163)	9.895 (2.360)
Gauss-Lasso	0.153 (0.070)	0.559 (0.125)	5.751 (3.059)	0.280 (0.137)	0.765 (0.145)	9.848 (5.073)
Lar [45]	0.728 (0.146)	1.021 (0.159)	6.250 (2.349)	0.560 (0.116)	0.868 (0.172)	8.251 (2.668)
Gauss-Lar	0.153 (0.070)	0.561 (0.123)	3.602 (2.700)	0.280 (0.137)	0.815 (0.155)	6.732 (3.503)
Gibbs sampler [102]	0.376 (0.087)	0.832 (0.121)	1.837 (0.294)	0.499 (0.112)	1.126 (0.170)	2.422 (0.375)
BG [59]	0.165 (0.077)	0.429 (0.103)	10.362 (3.254)	0.199 (0.103)	0.604 (0.130)	11.537 (4.437)
BG-MCMC	0.180 (0.072)	0.435 (0.106)	0.704 (0.129)	0.220 (0.107)	0.577 (0.147)	0.909 (0.149)
Proposed BLasso	0.157 (0.068)	0.417 (0.080)	0.708 (0.235)	0.199 (0.092)	0.577 (0.111)	1.497 (0.959)

BG-MCMC method provides significant performance improvements over the original non-Bayesian method in [59]. The shrinkage tuning parameter t for the Lasso and Lar is determined by 10-fold CV with the minimal prediction errors. The proposed BG-MCMC and the Bayesian Lasso estimators are initialized at the LS estimate of the full model and run for 100,000 iterations with the first half runs being discarded as warm-up. The step size ϵ of RW is 0.05. The performances of the coefficient estimates are measured by the Root Mean Squared Errors (RMSEs) and all results are averaged over 100 simulations.

3.4.2 Empirical Performance Comparisons

The RMSE performances of the eight models are summarized in Table 3.2. Several observations can be summarized as follows:

First, performances of the Lasso and Lar are similar to each other; this is in particular pronounced in lower dimension cases. This is because the Lasso (implemented in a modified Lar algorithm [45]) rarely drops variables from the active set and hence is very similar to the Lar. Further, compared with the Lasso and Lar, the proposed Bayesian Lasso with RJ-MCMC algorithm consistently yields much smaller RMSE and smaller estimation variability. This observation is even more significant for the $p \approx n$ cases where the MLE can be highly unstable. The reduced estimation variability is likely due to the fact that the parameters are estimated based on averaged models, rather than conditioning on a single *best* plausible model given by the penalized MLE principle. Moreover, since the Lasso uses one tuning parameter to simultaneously select variables and shrink estimates, the estimation may be shrunk along with the decreased model size. However, for our proposed approach, since two parameters (τ and λ) incorporate together to control these two effects, it is more flexible and likely to obtain an unbiased estimate. To support this claim, we reported the RMSE results of Gauss-Lasso and Gauss-Lar in Table 3.2 and also reported the estimated sparsity patterns in Lasso and Lar in Table 3.3. These two tables together show that the major source of RMSEs of Lasso and Lar comes from the errors made in the model selection stage.

The proposed RJ-MCMC based Bayesian Lasso also consistently yields better estimation accuracy than the Gibbs sampler-based Bayesian Lasso. The Gibbs sampler method yields the MSEs between the proposed RJ-MCMC method and Lasso/Lar.

The BG method in [59] has comparable, slightly worse performance over the proposed Bayesian Lasso in lower dimension regimes. However, when the model size increases comparable to the data size (e.g. $p = 90$ and $n = 100$), the performance of the BG method substantially degrades. The major advantage of the BG model is that the marginal likelihood of the model can be given in closed-form, conditioned on the known active set for a model. However, since comparing marginal likelihood of all possible models requires the enumeration of the 2^p possible models, an exhaustive search is not computationally feasible so a common means to approximate the exact solution is to adopt a stepwise searching strategy. As used in [59], forward selection is used to traverse between models, however false predictors introduced in earlier stages of the algorithm cannot be eliminated at a later stage. Degraded performance of BG is especially pronounced when predictors are highly correlated or the sample size is not large enough. Hence, for medium/large scale problems with no special structure (e.g. orthogonality among predictors), the BG method is not a good choice for model selection in practice. In contrast, the stochastic search algorithm proposed in this paper successfully avoids being trapped by sub-optima with the price of increased computational cost. Empirically, we observe that the proposed fully Bayesian algorithms can accurately estimate the model and associated parameters with a reasonable sampling size.

By extending the proposed fully Bayesian framework to the BG model [59], we further confirm the strength of the MCMC approach. We derive an MCMC algorithm for the BG model, as in Appendix A.3.2. Many parameters can be integrated out because of the Gaussianality. The main interesting quantity is the model index parameter, the posterior of which can be viewed as the posterior probability of including the corresponding coefficient. It is seen from Table 3.2 that the derived BG-MCMC achieves similar performances as the proposed RJ-MCMC based Bayesian Lasso.

We note empirically that there is not much room for improvement for the two fully Bayesian approaches, the proposed BLasso and BG-MCMC. With the assumption that we are given an oracle revealing the location of the true non-zero coefficients, it is easy to see that the optimal least squares estimator has an RMSE converging to $\sqrt{\frac{k}{n}}\sigma$ for large n. In our setup, with $\sigma^2 = 1$, the fundamental information-theoretic limits of RMSEs in Table 3.2 are 0.173, 0.387, 0.548, respectively. In our simulations, we empirically observe that the performances of the proposed RJ-MCMC based Bayesian Lasso and BG-MCMC approximate the lower estimation error bound. Therefore, we suggest that there is a substantial advantage over greedy search and optimization based methods by using the proposed fully Bayesian framework coupled with the stochastic search.

We also examine the sparsity patterns recovered by different algorithms under consideration. For BG, Lasso and Lar, sparsity patterns are characterized by the locations of non-zero coefficients. For BG-MCMC and the proposed Bayesian Lasso with RJ-MCMC, a non-zero coefficient is declared when its posterior probability passes the threshold 0.5. For the Bayesian Lasso based on the Gibbs sampler, since the distribution of $\boldsymbol{\beta}$ is assumed to have a conditional zero-mean normal distribution, we calculate its two-side tail probabilities that exceed the posterior estimate $\left|\hat{\boldsymbol{\beta}}_j\right|$ and then compare the tail probabilities with a significance level 0.1. To compare the algorithms, we evaluate the performances of support recovery in terms of their F-scores which combine the precision and recall (true positive rate) measures. More specifically, the precision P is the fraction of detected true positive among all identified positives whereas the recall R is the identified true positive ratio to the total number of true positives. The F-score is defined as the harmonic mean of the precision and recall, i.e.

$$F = \frac{2PR}{P+R}. \tag{3.13}$$

The closer F-score of an algorithm is to 1, the better performance it has.

It is clear from Table 3.3 that the proposed BLasso and BG-MCMC yield the best and stablest performances, with the proposed BLasso slightly outperms BG-MCMC. Lasso, Lar, Gibbs sampler, BG have comparable support detecting capability in lower dimensional problems; however, their performances significantly degrade as the problem size gets larger.

3.4.3 Convergence Analysis

We now prove that the proposed RJ-MCMC framework in Algorithm 1 converges to the posterior distribution of $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ given in (3.4). The proof is based on the standard argument, e.g. see [97]. Let $M = \left\{\left(\boldsymbol{\gamma}^{(i)}, \boldsymbol{\beta}^{(i)}\right)\right\}_{i\in\mathbb{N}}$ be the Markov chain constructed by Algorithm 1 such that the detailed balance condition (3.6) implies $\pi(\boldsymbol{\gamma}, \boldsymbol{\beta}|\mathbf{y})$ is an

Table 3.3: F-scores of estimated sparsity patterns averaged with standard deviations in brackets over 100 simulations for $n = 100$. Methods under comparison are the Lasso [115], Lar [45], Gibbs sampler based Bayesian Lasso [102], Binomial-Gaussian (BG) [59], the proposed BG-MCMC and the proposed Bayesian Lasso (BLasso) with RJ-MCMC algorithm.

	No correlation			**Correlation=0.5**		
	3/15	15/45	30/90	3/15	15/45	30/90
Lasso [115]	0.996 (0.025)	0.888 (0.062)	0.713 (0.179)	0.927 (0.099)	0.802 (0.049)	0.599 (0.133)
Lar [45]	0.996 (0.025)	0.887 (0.063)	0.798 (0.119)	0.927 (0.099)	0.763 (0.046)	0.673 (0.099)
Gibbs sampler [102]	0.964 (0.070)	0.612 (0.066)	0.528 (0.037)	0.929 (0.109)	0.572 (0.062)	0.517 (0.041)
BG [59]	0.850 (0.097)	0.505 (0.066)	0.230 (0.017)	0.791 (0.048)	0.393 (0.086)	0.141 (0.045)
BG-MCMC	0.993 (0.031)	0.993 (0.014)	0.991 (0.012)	0.994 (0.028)	0.998 (0.008)	0.999 (0.004)
Proposed BLasso	1.000 (0.000)	1.000 (0.003)	1.000 (0.004)	1.000 (0.000)	1.000 (0.000)	0.979 (0.070)

invariant distribution for M. With this target distribution, it suffices to show that M is ergodic with respect to $\pi(\boldsymbol{\gamma}, \boldsymbol{\beta}|\mathbf{y})$ [116]. This is equivalent to show that M is $\pi(\boldsymbol{\gamma}, \boldsymbol{\beta}|\mathbf{y})$-irreducible and aperiodic. Aperiodicity is obvious and the only part we need to argue is the $\pi(\boldsymbol{\gamma}, \boldsymbol{\beta}|\mathbf{y})$-irreducibility. The idea of showing the irreducibility of M is to find a particular path that monotonically shrinks the model size to one with certain positive probability. We can assume without loss of generality that the only predictor in the destination model is the first one. Let $K(\boldsymbol{\gamma}, \boldsymbol{\beta}; \boldsymbol{\gamma}', d\boldsymbol{\beta}')$ be the transition kernel of M:

$$P\left(\boldsymbol{\gamma}', \boldsymbol{\beta}' \in B|\boldsymbol{\gamma}, \boldsymbol{\beta}\right) = \int_B K(\boldsymbol{\gamma}, \boldsymbol{\beta}; \boldsymbol{\gamma}', d\boldsymbol{\beta}') \tag{3.14}$$

for all $B \in \mathcal{B}(\mathbb{R}^{|\boldsymbol{\gamma}'|})$. In order to prove that M is $p(\boldsymbol{\gamma}, \boldsymbol{\beta}|\mathbf{y})$-irreducible, it suffices to establish a μ-irreducibility of M for some σ-finite measure μ defined on the measurable space $\left([p] \times \mathbb{R}^p, 2^{[p]} \otimes \mathcal{B}(\mathbb{R}^p)\right)$, where $2^{[p]}$ denotes the power set of $[p]$ and $[p] = \{1, \cdots, p\}$, see [116]. Taking $\mu(\boldsymbol{\gamma}, \boldsymbol{\beta}) = p^{-1}\mathbb{I}_{\{1\}}(|\boldsymbol{\gamma}|)N(0, 1)$, we want to show that, for any $\boldsymbol{\gamma} \in 2^{[p]}$ and $\boldsymbol{\beta} \in \mathbb{R}^{|\boldsymbol{\gamma}|}$, there is a non-vanishing probability for the state

$(\boldsymbol{\gamma}, \boldsymbol{\beta})$ to commute $(\{1,0,0,\cdots\}, B)$ for every $\mu(\{1,0,0,\cdots\} \times B) > 0$. Considering the one-step transition kernel (3.14) for the event that a death occurs, we have by construction

$$K(\boldsymbol{\gamma}, \boldsymbol{\beta}; \boldsymbol{\gamma}', d\boldsymbol{\beta}') = \frac{1}{3k} \min\{A, 1\} \mathbb{I}_{S_{\boldsymbol{\gamma},\boldsymbol{\beta}}}(\boldsymbol{\beta}') d\boldsymbol{\beta}', \tag{3.15}$$

where A is the first term in the death probability ratio in (3.9) and

$$S_{\boldsymbol{\gamma},\boldsymbol{\beta}} = \left\{ \boldsymbol{\beta}' \in \mathbb{R}^{k-1} \,|\exists j \text{ s.t. } \boldsymbol{\gamma} = \{j\} \cup \operatorname{supp}(\boldsymbol{\beta}') \right\}.$$

Note that $\|\boldsymbol{\beta}\|_1 = \|\boldsymbol{\beta}'\|_1 + |u| \geq \|\boldsymbol{\beta}'\|_1$. Let $C_k = \frac{\|\mathbf{x}^j\|_2^2 u^2 + 2\langle \mathbf{y} - X\boldsymbol{\beta}, \mathbf{x}^j \rangle u}{\|\mathbf{y} - X\boldsymbol{\beta}\|_2^2} < \infty$ and $\alpha_k = (1 + C_k)^{-1}$. The Cauchy-Schwartz inequality implies that $\alpha_k > 0$. We have $A \geq C_k' \alpha_k > 0$, where $C_k' = \frac{p-(k-1)}{(k-1)^2} \frac{p(\boldsymbol{\gamma}' \to \boldsymbol{\gamma})}{p(\boldsymbol{\gamma} \to \boldsymbol{\gamma}')} N(u; 0, \epsilon^2) > 0$. So we deduce that

$$K(\boldsymbol{\gamma}, \boldsymbol{\beta}; \boldsymbol{\gamma}', d\boldsymbol{\beta}') \geq \frac{C_k' \alpha_k \mathbb{I}_{S_{\boldsymbol{\gamma},\boldsymbol{\beta}}}(\boldsymbol{\beta}') d\boldsymbol{\beta}'}{3p}.$$

Iterating this process $k-1$ times, we can obtain

$$\begin{aligned} P\left(\{1,0,0,\cdots\} \times B | \boldsymbol{\gamma}, \boldsymbol{\beta}\right) &\geq \int_B \prod_{i=2}^{k} K(\boldsymbol{\gamma}_i, \boldsymbol{\beta}_i; \boldsymbol{\gamma}_{i-1}, d\boldsymbol{\beta}_{i-1})\, d\boldsymbol{\beta}_1 \\ &\geq (3p)^{-(k-1)} \mu(\{1,0,0,\cdots\} \times B) \prod_{i=2}^{k} (C_i' \alpha_i) > 0. \end{aligned}$$

The last step may be complemented by a standard MH-step (3.5) to update the coefficient with the same dimension. This shows that we can reach the state $(\{1,0,0,\cdots\}, B)$ with a strictly positive probability. In summary, the above facts lead to the following convergence theorem.

Theorem 3.4.1. *Let* $\left(\boldsymbol{\gamma}^{(i)}, \boldsymbol{\beta}^{(i)}\right)$ *be the Markov chain with transition kernel given by the proposed RJ-MCMC algorithm in Algorithm 1. This Markov chain converges to the posterior probability distribution* $\pi(\boldsymbol{\gamma}, \boldsymbol{\beta}|\mathbf{y})$ *in (3.4), regardless of the initialization* *of the algorithm, i.e.*

$$\left\|\pi^{(i)}(\boldsymbol{\gamma}, \boldsymbol{\beta}) - \pi(\boldsymbol{\gamma}, \boldsymbol{\beta}|\mathbf{y})\right\|_{TV} \to 0, \tag{3.16}$$

where $\pi^{(i)}(\boldsymbol{\gamma}, \boldsymbol{\beta})$ *means the empirical distribution of* $\left(\boldsymbol{\gamma}^{(i)}, \boldsymbol{\beta}^{(i)}\right)$ *and* $\|\cdot\|_{TV}$ *means the total variation norm on bounded signed measures,*

$$\|\pi\|_{TV} = \sup_{B\in\mathcal{B}} \pi(B) - \inf_{B\in\mathcal{B}} \pi(B). \tag{3.17}$$

Remark 10. In addition to the birth-and-death proposal used in Algorithm 1, there is another main proposal studied in the literature, namely the *split-and-merge* strategy. It is worth noting that the convergence of the split-and-merge trans-dimensional strategy can also be established in a similar way. Theoretically, both proposals guarantee that the correspondingly designed algorithms converge to the right target distribution. The two proposals may result in different empirical convergence rates which are problem-dependent. As shown earlier, the adopted birth-and-death proposal yields satisfactory estimation performance in our simulations.

3.5 A Diabetes Data Example

This is a benchmark data set used in [45]. It contains $n = 442$ measurements from diabetes patients. Each measurement has ten baseline predictors: age, sex, body mass index (BMI), average blood pressure (BP), and six blood serum measurements (S1-S6). The response variable is a quantity that measures progression of the diabetes one year after baseline. The response is centered and the predictors are normalized to have zero means and unit variances, before applying any model selection methods. We set the random walk step size of the proposed RJ-MCMC to 7 in order to control the acceptance ratio of proposed models to be around 30%. In our experiment, the acceptance ratio is 30.54%. We empirically observe that a smaller step size would

Table 3.4: Estimated coefficients for the Lasso [115], Lar [45], the Bayesian Lasso based on the Gibbs sampler (GS) [102], BG [59], BG-MCMC, and proposed Bayesian Lasso for diabetes data.

Predictor	Lasso	Lar	GS	BG	BG-MCMC	Proposed BLasso
Age	-0.081	-0.325	-0.337	0	0	0
Sex	-10.920	-11.228	-11.043	-10.646	-10.923	-10.371
BMI	25.013	24.854	24.877	24.905	25.197	24.861
BP	15.074	15.303	15.183	15.380	15.633	14.655
S1	-15.803	-29.293	-20.997	-35.624	-29.872	-7.022
S2	5.196	15.970	9.551	25.314	20.377	0
S3	-4.498	1.232	-2.550	0	-2.778	-7.785
S4	5.839	7.434	6.174	0	0	3.168
S5	27.645	32.648	29.568	37.798	35.675	24.767
S6	3.101	3.174	3.202	0	0	2.227

cause the algorithm to explore the model space more slowly; while a larger one would have a higher rejection rate.

It is clear from Table 3.4 that models under comparison provide results with certain similarities, except for the predictors S1, S2, and perhaps S3. S1 selected by the proposed RJ-MCMC based Bayesian Lasso has a negative coefficient with smaller magnitude than others. For S2, the estimated coefficients are less consistent across different models. For instance, Lasso and Gibbs sampler have positive coefficients with smaller magnitudes than those of the Lar and BG. For the proposed Bayesian Lasso, this predictor is essentially interpreted as insignificant. Therefore, it is not clear whether or not S1-S3 covariates should be selected from a solely computational point of view, and further medical or physical interpretation is needed to justify the choice of different models in a real-world problem.

3.6 A Real fMRI Application

3.6.1 Application Description

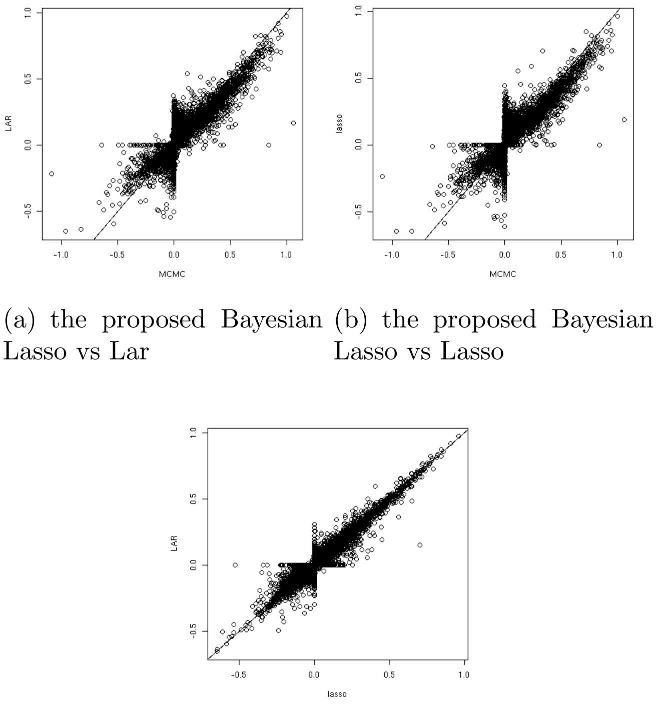
In this section we demonstrate an application to real fMRI data derived from subjects with Parkinson's Disease (PD). The problem of interest here is to employ the sparse linear regression modeling for learning brain functional connectivity using fMRI data.

The fMRI data we have are from ten normal people and eight PD patients. During the fMRI experiment, subjects continually squeezed a bulb in their right hand to control an inflatable ring so that the ring moved through an undulating tunnel without touching the sides. A trial of the task was five minutes. Normal subjects performed only one trial; the PD subjects performed the same task twice, once before medication, the other after the medication. fMRI data were collected with a Philips Achieva 3.0 T scanner with a TR interval of 2s. Hence, we collected 150 data points for each subject. After motion correction, the fMRI time courses of the voxels within each ROI were averaged to represent the summary activity of each ROI. The averaged time courses were then linearly detrended and normalized to unit variance. In this study, based on previous neuroscience knowledge, eighteen brain regions were selected as the ROIs based on each person's individual anatomy.

The model that we assume here is a linear regression one which incorporates both spatial and temporal effects of brain connectivities. To do so, we combine the structural equation modeling (SEM) [92] and multivariate autoregressive model (mAR) [65]. Specially, let $\mathbf{y}(t) = (y_1(t), \cdots, y_p(t))^T$ be a $p \times 1$ dimensional vector, which contains the intensity measurements for the p brain ROIs at time t, for $t = 1, \cdots, T$. For mAR component of the unified model, we consider only up to an m^{th} order process. Hence, by combining the SEM part, the joint SEM and mAR model can be expressed as:

$$\mathbf{y}(t) = \underbrace{\boldsymbol{A}\mathbf{y}(t)}_{\text{SEM}} + \underbrace{\sum_{j=1}^{m} \mathbf{\Phi}(j)\mathbf{y}(t-j)}_{\text{mAR}} + \underbrace{\mathbf{e}(t)}_{\text{noise}} \tag{3.18}$$

where $\boldsymbol{A}$ and $\mathbf{\Phi}(j)$ are $p \times p$ coefficient matrices to be estimated, for $j = 1, \cdots, m$; $\mathbf{e}$ is a noise vector which is assumed to be Gaussian with zero mean and constant variance σ^2. $\boldsymbol{A}$ represents the spatial connection strengths between ROIs, while $\mathbf{\Phi}(j)'$s are the time lag effect strengths. We are aiming at applying the proposed



(a) the proposed Bayesian Lasso vs Lar

(b) the proposed Bayesian Lasso vs Lasso

(c) Lar vs Lasso

Figure 3.2: The correlation between the estimated coefficients when using different algorithms. (a) the proposed Bayesian Lasso vs Lar; (b) the proposed Bayesian Lasso vs Lasso; (c) Lar vs Lasso.

fully Bayesian method to simultaneously select the model order and estimate the functional connectivities between brain regions represented by the coefficients.

3.6.2 Results

We primarily want to check the consistency of the estimates from the aforementioned algorithms. First, we plot and examine the correlations between the estimated coefficients of these algorithms. The result is shown in Figure 3.2. We can see that the proposed Bayesian Lasso, Lar, and Lasso estimates are highly correlated. In particular, the Lar and Lasso estimates have a higher similarity (with the correlation coefficient being 0.967) than being compared with the proposed Bayesian Lasso (with the correlation coefficient being 0.873 and 0.870 for the Lar and Lasso estimates, respectively).

Table 3.5: Correlations between the coefficient estimates on two fMRI data subsets when using MLE, Lar [45], Lasso [115] and the proposed BLasso.

Method	MLE	Lar	Lasso	Proposed BLasso
Correlation	0.293	0.599	0.587	0.702

We further investigate estimation stability. We take one subject's fMRI time-series data and split the 150 time data points into two subsets of size 100, with the middle 50 data points overlapped. We then learn two models from these two subsets and examine the correlation between the estimated coefficients from subsets. The MLE approach for the full (non-sparse) model is also included for comparison. It is noted that the proposed Bayesian Lasso reveals greater estimation stability between models derived from subsets of the fMRI data compared to the other three methods, since it yields the highest similarity between the model coefficients derived from the two subsets (represented by a correlation coefficient being 0.702, see Table 3.5). The Lar and Lasso estimates yield a lower correlation (with the correlation coefficient being 0.599 and 0.587 respectively). The MLE approach provides the lowest consistency (with the correlation coefficient being 0.293), which is not a surprising fact since the variability of the MLE estimate is usually larger than the estimates via sparse regression. Moreover, limiting our attention only to the predictor coefficients which are estimated as non-zero from both subsets, we note that the correlations between the estimates derived from two subsets are even higher for the proposed Bayesian Lasso, Lar, and Lasso approaches.

3.7 Discussion and Conclusion

In this chapter, we proposed a hierarchical, fully Bayesian version of the Lasso model for inferring sparse linear regression from high-dimensional data sets. Since the joint posterior distribution of the parameters involves both discrete and continuous parameters, we developed a reversible jump MCMC algorithm to compute the unbiased minimum variance estimates. Simulations demonstrated that the proposed Bayesian

Lasso estimate yields lower estimation errors when compared with popular Lasso type estimates and a Gibbs sampler based Bayesian estimate. One intuitive explanation of this observation is that model averaging by the fully Bayesian approach provides better stability than selecting only a single *best* model. The simulations further demonsrated that the proposed Bayesian Lasso is robust to correlated predictors, even though the hypothesis of independent priors for predictors is assumed in the model design (3.1). We proved the convergence of the proposed RJ-MCMC algorithm for the Bayesian Lasso. Further, we extended the proposed fully Bayesian framework to the Binomial-Gaussian model, and simulations showed that the proposed stochastic search could substantially improve the performance of the original BG model-based estimate in [59].

One important direction for future work is to improve the sampling strategy. Currently, we use the Gaussian RW proposal with a fixed variance parameter. However, as shown in [66], the adaptive RJ-MCMC sampler usually facilitates mixing speed, and thus a data-driven adaptive sampler is of particular interest. We also observe that different step sizes can lead to different models and affect the empirical convergence.

The proposed RJ-MCMC Bayesian Lasso approach also has limitations. One is its higher computational cost compared with Lasso-type estimates, a limitation common to MCMC-based Bayesian approaches. This limitation prevents using the proposed method from online high-dimensional estimation problems. However for offline/batch estimation problems (e.g. fMRI modeling), the proposed method can usually provide practical accurate estimates with affordable computational complexity. Table 3.6 reports the required CPU times by different methods in the case of $p = 15$ and uncorrelated design. We observed similar empirical complexity results for other settings in our simulations. Basically, among the discussed MCMC-based Bayesian approaches, we note that the Gibbs sampler based Bayesian Lasso requires the highest computational cost, followed by BG-MCMC, while the RJ-MCMC Bayesian Lasso requires the least computational time. In summary, the computational costs of different estimates

Table 3.6: The required CPU time for Lar [45], Lasso [115], the Bayesian Lasso based on the Gibbs sampler (GS) [102], BG [59], BG-MCMC, and proposed Bayesian Lasso, in the case of $p = 15$ and uncorrelated design. CPU times are normalized w.r.t. the Lar running time. The last three simulation based methods run 20,000 iterations.

Method	Lar	Lasso	BG	BG-MCMC	RJ-MCMC	GS
CPU time	1.000	1.296	25.250	346.065	98.749	2848.318

can be ordered as:

$$\begin{aligned}\text{Lar} \prec \text{Lasso} \prec \text{BG} &\prec \text{RJ-MCMC Bayesian Lasso}\\ \prec \text{BG-MCMC} &\prec \text{Gibbs sampler based Bayesian Lasso.}\end{aligned}$$

Chapter 4

Shrinkage-To-Tapering Estimation of Large Covariance Matrices

4.1 Introduction

The main goal of this chapter is to consider the estimation problem of high-dimensional covariance matrix from n iid smaples following a zero-mean p-dimensional multivariate Gaussian distribution $N(\mathbf{0}, \Sigma)$[4]. Importance and motivation of precise estimation of covariance matrices have been discussed in Chapter 1 of this thesis; therefore, we do not repeat them and jump directly to our results. Before proceeding, we remind the audience that the standard and most natural estimator of Σ is the *unstructured sample covariance matrix*

$$\hat{S} = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T,$$

where $\mathbf{x}_i$ means the ith p-dimensional observation sample. This is defined in (1.3) in Chapter 1[5]. Recall that for the classical case where p is fixed and $n \to \infty$, $\hat{S}$ is a consistent estimator of Σ. Unfortunately, the covariance estimation problem becomes fundamentally different and more challenging for high-dimensional settings with a small number of samples where $p \gg n$ meaning the concentration $p/n \to \infty$ (i.e. large-p-small-n). From the eigen-structure perspective, random matrix theory

[4] Without loss of generality (w.l.o.g.), we assume that the diagonal entries of Σ are all normalized to one.

[5] Note that we have temporally changed the notation $\Sigma_n^\star$ for sample covariance matrix to $\hat{S}$ for simplicity in the current chapter.

predicts that the spectrum of $\hat{S}$ is wider than the spectrum of Σ if $p/n \nrightarrow 0$ and $n, p \to \infty$ [6]. For example, the Marchenko-Pastur theorem states that the eigenvalues of $\hat{S}$ have a deterministic semicircular limiting distribution supported on $[(1-\sqrt{y})^2, (1+\sqrt{y})^2]$ where $y = \lim_{n\to\infty} p/n > 0$, while the spectrum of Σ is the Dirac mass at 1. This chapter considers the high-dimensional settings and focuses on the corresponding problem of estimating large covariance matrices.

The rest of the chapter is structured as follows. In Section 4.2, we first introduce the tapering estimator and its minimax risk bounds under Frobenius and spectral norms; we the derive the risk bounds under the same norms for the MMSE shrinkage oracle estimator proposed in [36]. Inconsistency of the shrinkage estimator will be shown by a set of examples. In Section 4.3, we propose a shrinkage-to-tapering oracle (STO) estimator and derive a closed-form expression of the optimal shrinkage weight under MMSE. An approximating algorithm of the STO estimator is further proposed for practical implementation. Section 4.4 compares the numeric performances of the proposed STO estimators, the tapering estimator and other types of shrinkage estimators. The chapter is concluded in Section 4.5.

4.2 Comparison Between Tapering and Shrinkage Estimators

The main contribution of this section is to provide a detailed analysis on risk bounds of tapering and shrinkage estimators for large covariance matrices estimation. Before we formally present our analysis, it is necessary to recall some definitions and results from tapering and shrinkage estimators of covariance matrices.

4.2.1 Tapering Estimator

We consider a class of tapering estimators. Let $\mathcal{S}$ be the set of $p \times p$ symmetric matrix and $A \circ B$ be the Schur product of two matrices A and B: $A \circ B = (a_{ij} b_{ij})$.

Definition 4.2.1. A *covariance matrix taper* (CMT) A is an element in $\mathcal{S}$ such that $\sum_{j=1}^{p} \lambda_j(A \circ B) \leq \sum_{j=1}^{p} \lambda_j(B)$ for all $B \in \mathcal{S}$. In other words, Schur multiplication by any CMT decreases the averaged eigenvalue.

Let W be a CMT, a *tapering estimator* of the covariance matrix is defined as

$$\hat{\Sigma}_{\text{taper}} = W \circ \hat{S}. \tag{4.1}$$

For some $C, C_0 > 0$ and $\alpha > 0$, we consider the following class of covariance matrices

$$\mathcal{G}(\alpha, C, C_0) = \left\{ \Sigma : \max_j \sum_{|i-j|>k} |\sigma_{ij}| \leq Ck^{-\alpha}, \forall k, \text{ and } \lambda_{\max}(\Sigma) \leq C_0 \right\}, \tag{4.2}$$

where α is a smoothing parameter specifying the rate of decay of σ_{ij} from the main diagonal. We state that the matrices in $\mathcal{G}(\alpha, C, C_0)$ *diagonally dominant.* Note that our definition is different from the usual one in the literature and we use this term as a measure of sparsity of covariance matrices when a natural ordering in variables exists, e.g. in time-series models. The following remarkable theorem, proved by Cai, Zhang, and Zhou [20], shows that a covariance tapering estimator based on data generated from i.i.d $N(\mathbf{0}, \Sigma)$ with $\Sigma \in \mathcal{G}(\alpha, C, C_0)$ is minimax.

Theorem 4.2.1. *(Cai, Zhang, and Zhou [20]) Suppose* $\log p = o(n)$ *and* $p \geq n^{\xi}$ *for some* $\xi > 0$*; then we have the following minimax convergence rate*

1. *under the Frobenius risk/normalized MSE:*

$$\frac{1}{p} \inf_{\hat{\Sigma}} \sup_{\Sigma \in \mathcal{G}(\alpha, C, C_0)} E \left\| \hat{\Sigma} - \Sigma \right\|_F^2 \asymp n^{-\frac{2\alpha+1}{2(\alpha+1)}}; \tag{4.3}$$

2. *under the spectral risk:*

$$\inf_{\hat{\Sigma}} \sup_{\Sigma \in \mathcal{G}(\alpha, C, C_0)} E \left\| \hat{\Sigma} - \Sigma \right\|^2 \asymp n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}, \tag{4.4}$$

where the infimum is taken over all possible estimators $\hat{\Sigma} : \mathbb{R}^{n\times p} \to \mathbb{R}^{p\times p}$ *for* Σ *based on the data.*

It is very interesting and important to ask how we can construct CMTs that actually attain the minimax risks. Fortunately, it turns out that there exists such a CMT with different bandwidths (to be defined shortly) that is rate-optimal for each of the two risks in the minimax sense. More specifically, let $W = (w_{ij})$ be defined as

$$w_{ij} = \begin{cases} 1, & \text{for } |i-j| \le k_h \\ 2 - |i-j|/k_h, & \text{for } k_h < |i-j| < k \\ 0, & \text{for } |i-j| \ge k \end{cases}, \tag{4.5}$$

where $k_h = k/2$. First, we can see that such defined W is a valid matrix taper according to Definition 4.2.1 since $\text{diag}(W) = \mathbf{1}$ and therefore

$$\sum_{j=1}^{p} \lambda_j(W \circ \Sigma) = \text{Tr}(W \circ \Sigma) = \text{Tr}(\Sigma) = \sum_{j=1}^{p} \lambda_j(\Sigma).$$

Second, it is clear that such defined W and hence $W \circ \Sigma$ for every Σ vanish off the stripe that is at most $(k-1)$ away from the main diagonal. Therefore, k is defined as the *bandwidth* of the CMT. Third, it should be noted that $W \circ \Sigma$ does not necessarily preserve the positive definiteness of Σ. However, this concern can be mitigated by first diagonalizing $W \circ \Sigma$ and then replacing its negative eigenvalues by zeros. This modification preserves the minimax error bounds (up to a constant of 2) and also the positive definiteness of the resulting estimate. It is noted that the optimal procedures are different under the Frobenius and spectral norms. It has been shown that optimal bandwidths of W under the normalized Frobenius and spectral norms are $n^{1/2(\alpha+1)}$ and $n^{1/(2\alpha+1)}$, respectively.

4.2.2 Shrinkage Estimator

Since the discovery of the Stein's effect on the inadmissibility of the multivariate normal mean vector by the usual sample mean estimator when $p \geq 3$ (see [112] for the original reference), extensive research has been devoted to proposing a broad range of shrinkage estimators such as the James-Stein estimator [71], its truncated version [8], among many others, to improve the performance of the usual estimator in terms of risks induced by a variety of loss functions. Similarly as in the estimation of the mean vector, the sample covariance estimator $\hat{S}$, as we have mentioned, is unsatisfactory for large (high-dimensional) covariance estimation problems. Steinian shrinkage therefore has been an alternatively attractive choice. Estimators of this kind naturally have the form

$$\hat{\Sigma}(\rho) = (1-\rho)\hat{S} + \rho T, \tag{4.6}$$

where $\rho \in [0,1]$ is the shrinkage coefficient and T is the shrinkage target matrix. In general, T is supposed to have the properties of being: (i) well-conditioned; (ii) consistent or even optimal in a subspace of $p \times p$ symmetric matrices. In other words, the shrinkage estimator is a convex combination between the sample covariance matrix and a "good" target matrix. There are several possible and intuitive choices of T. For instance, we consider $T = \hat{F} := p^{-1}\mathrm{Tr}(\hat{S})I$, and the shrinkage estimator has the following form

$$\hat{\Sigma}(\rho) = (1-\rho)\hat{S} + \rho\hat{F}. \tag{4.7}$$

Chen et.al [36] defines an MMSE oracle estimator $\hat{\Sigma}_o := \hat{\Sigma}(\hat{\rho}_o)$ where $\hat{\rho}_o$ is defined as the solution of the optimization problem

$$\hat{\rho}_o = \mathrm{argmin}_{\rho\in[0,1]} \qquad E\left\|\hat{\Sigma}(\rho) - \Sigma\right\|_F^2 \qquad \text{subject to} \qquad \hat{\Sigma}(\rho) = (1-\rho)\hat{S} + \rho\hat{F}.$$

The MMSE oracle estimator seeks the best convex combination between the sample covariance matrix and a scaled identity matrix to approximate the true covariance matrix in terms of the mean-squared errors (MSEs). This estimator is said to be an oracle because the optimal solution depends on Σ which is unknown in practice and is the estimation goal. It is shown in [82] that $\hat{\rho}_o$ can be given by a distribution-free formula

$$\rho_o = \frac{E[\mathrm{Tr}((\Sigma - \hat{S})(\hat{F} - \hat{S}))]}{E\|\hat{S} - \hat{F}\|_F^2}. \tag{4.8}$$

Under additional Gaussian assumption, the closed-form of ρ_o is given in [36]

$$\rho_o = \frac{p - 2 + pt}{p(n+1) - 2 + (p-n)t}, \tag{4.9}$$

where

$$t = \mathrm{Tr}^2(\Sigma)/\mathrm{Tr}(\Sigma^2). \tag{4.10}$$

Here t measures the distribution of the off-diagonal entries of Σ. In particular,

$$\mathrm{Tr}(\Sigma^2) \leq \mathrm{Tr}^2(\Sigma) \leq p\mathrm{Tr}(\Sigma^2),$$

where equalities of the left and right inequalities are attained if and only if $\Sigma = \mathbf{1}\mathbf{1}^T$ and $\Sigma = I$, respectively. So when $t = 1$, the matrix entries have the most spread support (dense); while when $t = p$, the energy of Σ concentrates on the diagonal (sparse) .

A second shrinkage estimator proposed in [53] combines $\hat{S}$ and $T = \mathrm{diag}(\hat{S})$ in the same manner as in (4.7). These two estimators share similar properties since the optimal coefficient can be obtained in a single distribution-free framework. Therefore, in the rest of the paper, we focus on the identity target case in (4.7) which is easier and more expressive for our theoretic analysis.

First, we derive the Frobenius risk of the MMSE oracle estimator (4.7), assuming that the data are from i.i.d. $N(\mathbf{0}, \Sigma)$.

Theorem 4.2.2. *Suppose $\{\mathbf{x}_i\}_{i=1}^n$ are i.i.d. Gaussian $N(\mathbf{0}, \Sigma)$. The Frobenius risk of the MMSE shrinkage oracle estimator (4.7) is given by*

$$E\|\hat{\Sigma}_o - \Sigma\|_F^2 = \left[(1 - \frac{t}{p})\rho_o + \frac{2}{np}\right] \|\Sigma\|_F^2 \,. \tag{4.11}$$

From Theorem 4.2.2, we can see that the Frobenius risk of the shrinkage oracle estimator primarily depends on ρ_o and the property of Σ. The second term in (4.11) contributes negligibly to the total risk when Σ is bounded away from the identity matrix, where $t < p$.

Since it is difficult for us to derive the exact formula, we also derive a lower bound on the risk under the spectral norm.

Theorem 4.2.3. *Suppose $\{\mathbf{x}_i\}_{i=1}^n$ are i.i.d. Gaussian $N(\mathbf{0}, \Sigma)$. The spectral risk of the MMSE shrinkage oracle estimator (4.7) satisfies*

$$E\|\hat{\Sigma}_o - \Sigma\|^2 \geq \rho_o^2(1 - \lambda_{\min}(\Sigma))^2. \tag{4.12}$$

Theorem 4.2.2 and 4.2.3 are important in the sense that, by giving the pointwise explicit risk bounds of Σ in the parameter space $\mathcal{S}$, it is possible for us to analyze the theoretic properties, such as consistency and admissibility, of the MMSE shrinkage oracle estimator (4.7). Indeed, we will shortly see that this shrinkage estimator is inconsistent for some high-dimensional covariance matrices that may often appear in many real-world applications; therefore, it is inadmissible for a subspace of the parameter set $\mathcal{S}$ and this suggests that we shall find alternative solutions. This is the main motivation of the proposed STO estimators which will be introduced shortly in Section 4.3.

4.2.3 Comparison of Risk Bounds Between Tapering and Shrinkage Estimators

We are now ready to compare the risk bounds of the tapering and MMSE shrinkage oracle estimator, thanks to Theorem 4.2.1, Theorem 4.2.2, and Theorem 4.2.3. The comparison is done by studying several specific examples and several interesting conclusions can be drawn. We describe the examples as follows.

Example 4.2.1. Consider, for $0 < \gamma < 1$,

$$\sigma_{ij} = \begin{cases} 1, & \text{for } i = j, \\ \gamma^{|i-j|}, & \text{for } i \neq j. \end{cases} \tag{4.13}$$

In words, the entries of Σ decay exponentially fast when moving away from the main diagonal. This example corresponds to the covariance structure of auto-regression models with order 1, AR(1), and is considered in [11, 36]. We can easily see that, for this Σ, $\mathrm{Tr}(\Sigma) = p$ by using the normalization assumption and $\mathrm{Tr}(\Sigma^2) = \|\Sigma\|_F^2 \asymp p(1+\gamma^2)/(1-\gamma^2)$ by summing the squared ℓ^2 norm of all diagonals of Σ. More specifically,

$$\|\Sigma\|_F^2 = p + 2(p-1)\gamma^2 + 2(p-2)\gamma^4 + \cdots + 2\gamma^{2(p-1)} = 2\sum_{j=0}^{p-1}(p-j)\gamma^{2j} - p$$
$$= \frac{1-\gamma^{2p}}{1-\gamma^2} \times 2p - C_0 - p \to \frac{1+\gamma^2}{1-\gamma^2}p, \qquad \text{for } p \text{ being sufficiently large.}$$

Therefore, it follows that

$$t = \frac{\mathrm{Tr}^2(\Sigma)}{\mathrm{Tr}(\Sigma^2)} \to \frac{1-\gamma^2}{1+\gamma^2}p := Cp, \qquad \text{as } p \to \infty. \tag{4.14}$$

But this then implies that the Frobenius risk $p^{-1}E\|\hat{\Sigma}_o - \Sigma\|_F^2$ in the super-linear high-dimensional situation is asymptotically, as $n \to \infty$, $p \to \infty$, and $n/p \to 0$,

$$C^{-1}\left[(1-C)\frac{Cp^2+p-2}{Cp^2+(1-C)np+p-2}+\frac{2}{np}\right] \to C^{-1}-1 := C(\gamma) > 0$$

where $C(\gamma) = 2\gamma^2/(1-\gamma^2)$, since $C \in (0,1)$. Therefore we can conclude that the Frobenius risk is

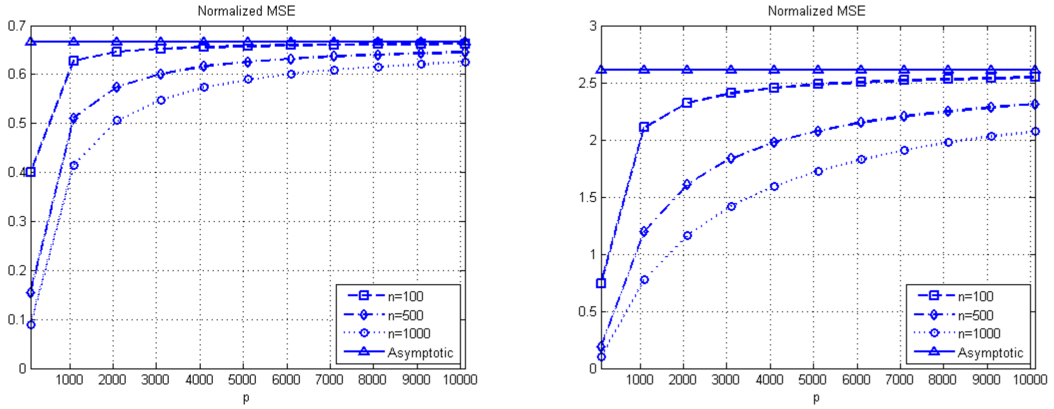
$$\frac{1}{p}E\|\hat{\Sigma}_o - \Sigma\|_F^2 = C(\gamma) + o(1). \tag{4.15}$$

It is clear that the normalized MSE is lower bounded by a positive constant depending on γ and therefore the MMSE shrinkage oracle estimator cannot be a consistency estimator of Σ unless the concentration $p/n \to 0$. Figure 4.1(a) plots the finite sample size behavior of the normalized MSE and its limit (4.15). We can see that the normalized MSE asymptotically approaches to a non-zero value when $n/p \to 0$ with n, p being large enough. On the contrary, note that Σ in this example satisfies any smoothing parameter $\alpha \in (0, \infty)$, we hence deduce that, under the Frobenius risk, the convergence rate, $n^{-(2\alpha+1)/(2\alpha+1)}$ in (4.3), of the tapering estimator $\hat{\Sigma}_{\text{taper}}$ (4.1) with the minimax CMT W in (4.5) can be arbitrarily close to n^{-1}. Comparing these two bounds, it is clear that, for this example, the tapering estimator is uniformly superior than the MMSE shrinkage oracle estimator proposed in [36]. Therefore, this oracle estimator is in fact a *weak oracle* which is overly restrictive in terms of the functional form of the shrinkage estimator. The reason that the tapering estimator outperforms the MMSE shrinkage oracle estimator is that the optimal estimate may not necessarily be decomposed as a simple convex combination between the sample covariance matrix $\hat{S}$ and the scaled identity matrix $\hat{F}$. Example 4.2.1 is a good evidence of this fact.

Furthermore, noticing that (4.9) and (4.14), we have

$$\rho_o \asymp \frac{C(\gamma)p^2+p-2}{C(\gamma)p^2+(1-C(\gamma))np+p-2} \to 1, \qquad \text{when } p \gg n \text{ and } p \to \infty, \tag{4.16}$$

Figure 4.1: Normalized MSE curves of the shrinkage MMSE estimator for the large covariances discussed in Example 4.2.1 and 4.2.2. The MMSE estimator fails to be consistent when $n/p \to 0$, because the normalized Frobenius risks converge to the asymptotic values, calculated from (4.15) and (4.18), that are bounded away from 0.



(a) Example 4.2.1 with $\gamma = 0.5$.

(b) Example 4.2.2 with $\alpha = 0.3$.

we see from the eigen-structure perspective that the spectral risk of the MMSE shrinkage oracle estimator (4.7) obeys

$$E\|\hat{\Sigma}_o - \Sigma\|^2 \geq (1 + o(1))(1 - \lambda_{\min}(\Sigma))^2, \tag{4.17}$$

which implies that $E\|\hat{\Sigma}_o - \Sigma\|^2 \nrightarrow 0$ because $\lambda_{\min}(\Sigma)$ is a monotone decreasing sequence as $p \to \infty$.

As a summary for Example 4.2.1, we conclude that, although the *bona fide* covariance matrix in (4.13) has a diagonal-like structure, shrinkage of $\hat{S}$ to an identity matrix is in consistent and hence it is not a good choice in high-dimensional situations and this procedure shall be improved by taking into account more refined structural information. On the contrary, tapering is minimax in this example. □

We now study a second example that has a slower polynomial decay rate than in Example 4.2.1, as considered in [20].

Example 4.2.2. Consider, for $\alpha > 0$,

$$\sigma_{ij} = \begin{cases} 1, & \text{for } i = j, \\ |i-j|^{-(\alpha+1)}, & \text{for } i \neq j. \end{cases}$$

Based on the definition in (4.2), we can show that $\Sigma \in \mathcal{G}(\alpha^{-1}, 1, C_0)$. It follows from an analogous argument that

$$\|\Sigma\|_F^2 = (1 + C_1)p - C_2,$$

where $C_1 = 2\sum_{j=1}^{p-1} j^{-2(\alpha+1)} > 0$ and $C_2 = 2\sum_{j=1}^{p-1} j^{-2\alpha-1} > 0$, both converging as $p \to \infty$. The rest derivation proceeds as in Example 4.2.1 and consequently we can achieve the argument that there exists a constant $C(\alpha) > 0$ such that

$$\frac{1}{p} E\|\hat{\Sigma}_o - \Sigma\|_F^2 = C(\alpha) + o(1), \tag{4.18}$$

as illustrated in Figure 4.1(b). Similar to the analysis in Example 4.2.1, we can see that the tapering estimator outperforms the MMSE shrinkage oracle estimator under both Frobenius and spectral risks. Again, tapering estimator is minimax and MMSE shrinkage-to-identity is inconsistent in this example when the dimensionality is large. □

Example 4.2.3. In a third example, we consider the covariance structure of a fractional Brownian motion (FBM) with the Hurst parameter $h \in [0.5, 1]$:

$$\sigma_{ij} = 2^{-1}[(|i-j|+1)^{2h} - 2|i-j|^{2h} + (|i-j|-1)^{2h}].$$

The FBM is a model for complex systems that have long-range dependence for h being close to 1, such as modeling the internet traffic [85]. Practical applications usually tune h between 0.5 and 0.9. The covariance matrix in this model does not belong to $\mathcal{G}(\alpha, C, C_0)$ unless $h = 0.5$, which is the case of the Brownian motion with white Gaussian noise process; therefore, tapering estimator does not guarantee to be consistent estimator in this example. To see this, we first observe that $\sigma_{ij} \geq 0$ since x^{2h} is convex in x for $h \in [0.5, 1]$ and then obtain that

$$\begin{aligned}
\|\Sigma\|_1 = \sum_{i=1}^{p}\sum_{j=1}^{p}\sigma_{ij} &= p + \sum_{k=1}^{p-1}(p-k)[(k+1)^{2h} - 2k^{2h} + (k-1)^{2h}] \\
&= p + p\sum_{k=1}^{p-1}\left\{(k+1)^{2h} - k^{2h} - \left[k^{2h} - (k-1)^{2h}\right]\right\} \\
&\quad - \sum_{k=1}^{p-1}\left\{(k+1)^{2h+1} - k^{2h+1} - \left[k^{2h+1} - (k-1)^{2h+1}\right]\right\} \\
&\quad + \sum_{k=1}^{p-1}\left[(k+1)^{2h} - (k-1)^{2h}\right] \\
&= p + p\left[p^{2h} - 1 - (p-1)^{2h}\right] - \left[p^{2h+1} - 1 - (p-1)^{2h+1}\right] + \left[p^{2h} + (p-1)^{2h} - 1\right] \\
&= p^{2h},
\end{aligned}$$

where the second last equality follows from three telescope sums. Consequently, it follows that

$$\max_j \sum_{|i-j|\geq 1} \sigma_{ij} \geq p^{-1} \sum_{1\leq i\neq j\leq p} \sigma_{ij} = p^{2h-1} - 1,$$

where the last term is not summable for $h > 0.5$ as $p \to \infty$. But now, we clearly see from definition (4.2) that $\Sigma \notin \mathcal{G}(\alpha, C, C_0)$.

On the other hand, since the covariance matrix in Example 4.2.3 is Toeplitz, its Frobenius norm is given by

$$\|\Sigma\|_F^2 = p + \frac{1}{2}\sum_{j=1}^{p-1}(p-j)\left[(j+1)^{2h} - 2j^{2h} + (j-1)^{2h}\right]^2.$$

For $x \geq 1$, let us define f as a function of h:

$$f(h) := f_x(h) = (x+1)^{2h} - 2x^{2h} + (x-1)^{2h}.$$

It is clear that f is continuous, $f(1/2) = 0$ and $f(1) = 2$. Consequently we have

$\|\Sigma\|_F^2 = p$ when $h = 0.5$ and $\|\Sigma\|_F^2 = p^2$ when $h = 1$. For $h \in (0.5, 1)$, because the function $(x^{2h} \ln x)$ is asymptotically convex for any $0.5 < h < 1$ as $x \to \infty$, it follows from the Jensen's inequality that $f'(h) \geq 0$ for sufficiently large x. Therefore, we deduce that $f(\cdot)$ will eventually be an increasing function between $[0, 2]$ for $h \in (0.5, 1)$, as x diverges to infinity. So $t \to p$ as $h \to 0.5$, while $t \to 1$ as $h \to 1$.

By Theorem 4.2.2, for the MMSE shrinkage oracle estimator, we now have

$$\begin{aligned} p^{-1}E\|\hat{\Sigma}_o - \Sigma\|_F^2 &= \left[(1 - \frac{t}{p})\rho_o + \frac{2}{np}\right] \frac{\|\Sigma\|_F^2}{p} = (\frac{p}{t} - 1)\rho_o + \frac{2}{nt} \\ &= \frac{p-t}{t} \times \frac{p - 2 + pt}{p(n+1) - 2 + (p-n)t} + o(1). \end{aligned}$$

Therefore, if $t = p$, i.e. $\Sigma = I$, then the Frobenius risk vanishes to zero and the MMSE shrinkage oracle estimator is a consistent estimator; if $t = 1$, i.e. $\Sigma = \mathbf{1}\mathbf{1}^T$, the Frobenius risk is asymptotically $2p/n$, meaning that the Frobenius risk for estimating this large covariance matrix depends on the concentration p/n. □

4.3 A Shrinkage-to-Tapering Estimator

4.3.1 Problem Formulation

Motivated by the above discussions, we propose a Steinian shrinkage type estimator. With the important difference from the shrinkage estimator toward a scaled identity matrix, the proposed estimator shrinks the sample covariance matrix to its tapered version. Basically the proposed estimator subsumes $T = \text{diag}(\hat{S})$ in [53] as one special case where $W = I$. Specifically, the proposed estimator $\hat{\Sigma}^{\text{STO}} := \hat{\Sigma}(\hat{\rho}^{\text{STO}})$ has the form

$$\hat{\Sigma}(\rho^{\text{STO}}) = (1 - \rho^{\text{STO}})\hat{S} + \rho^{\text{STO}}(W \circ \hat{S}), \tag{4.19}$$

where ρ^{STO} is determined by the solution to the optimization problem

$$\rho^{\text{STO}} = \text{argmin}_{\rho \in [0,1]} \qquad E\|\hat{\Sigma}(\rho) - \Sigma\|^2 \qquad \text{subject to} \qquad \hat{\Sigma}(\rho) = (1-\rho)\hat{S} + \rho(W \circ \hat{S}).$$

Here $\|\cdot\|$ can be either Frobenius or spectral norm. By using the tapering estimator as our shrinkage target, we hope this estimator can inherit good properties from both tapering and shrinkage estimators. Throughout the rest of the paper, we shall refer this proposed estimator as the *shrinkage-to-tapering oracle* (STO) estimator.

On one hand, for $\Sigma \in \mathcal{G}(\alpha, C, C_0)$, we can see from Theorem 4.2.1 that the proposed STO estimator reduces to the tapering estimator for large n and p. On the other hand, for $\Sigma \notin \mathcal{G}(\alpha, C, C_0)$, the proposed estimator reduces to an analogy of the MMSE shrinkage oracle estimator. Therefore, we expect that, for an arbitrary large covariance matrix Σ, the proposed estimator could improve upon both tapering and MMSE shrinkage oracle estimators.

The optimal coefficient of the STO estimator can be given in a closed-form.

Theorem 4.3.1. *The coefficient of the proposed STO estimator under the minimum Frobenius risk is*

$$\hat{\rho}^{STO} = \frac{E(\|\hat{S}\|_F^2 - \|V \circ \hat{S}\|_F^2) - (\|\Sigma\|_F^2 - \|V \circ \Sigma\|_F^2)}{E\|\hat{S}\|_F^2 + E\|W \circ \hat{S}\|_F^2 - 2E\|V \circ \hat{S}\|_F^2}, \tag{4.20}$$

where $V = (v_{ij})$ with $v_{ij} = \sqrt{w_{ij}}$. Under further Gaussian assumption, we can write (4.20) in a closed-form given as in (48).

4.3.2 Approximating the Oracle

The proposed STO estimator is nice in theory for developing the closed-form expression of the optimal coefficient $\hat{\rho}^{\text{STO}}$. Nevertheless, in practice, the true Σ is the target of estimation and thus unknown. So the proposed oracle estimator is not feasible in practice. Therefore, we propose a practical algorithm that approximates the oracle estimator. Following the idea of [36], we define a *shrinkage-to-tapering oracle ap-*

proximating (STOA) estimator as an iterative procedure between the following two steps:

1.

$$
\begin{aligned}
\hat{\rho}^{\mathrm{ST}}_{j+1} = & \Big[\mathrm{Tr}(\hat{\Sigma}_j\hat{S}) - \mathrm{Tr}((V\circ\hat{\Sigma}_j)(V\circ\hat{S})) + \mathrm{Tr}^2(\hat{\Sigma}_j) - \mathrm{Tr}(\hat{D}_jV^2\hat{D}_j)\Big] \\
& \Big/ \Big[(n+1)(\mathrm{Tr}(\hat{\Sigma}_j\hat{S}) + \mathrm{Tr}((W\circ\hat{\Sigma}_j)(W\circ\hat{S})) - 2\mathrm{Tr}((V\circ\hat{\Sigma}_j)(V\circ\hat{S}))) \\
& \quad + \mathrm{Tr}^2(\hat{\Sigma}_j) + \mathrm{Tr}(\hat{D}_jW^2\hat{D}_j) - 2\mathrm{Tr}(\hat{D}_jV^2\hat{D}_j)\Big], \qquad (4.21)
\end{aligned}
$$

where $\hat{D}_j$ is a diagonal matrix such that $\hat{D}_j = \mathrm{diag}(\hat{\Sigma}_j)$.

2.

$$
\hat{\Sigma}_{j+1} = (1-\hat{\rho}^{\mathrm{ST}}_{j+1})\hat{S} + \hat{\rho}^{\mathrm{ST}}_{j+1}(W\circ\hat{S}). \qquad (4.22)
$$

With an appropriate initialization, the two steps are operated iteratively until the sequence $\{\hat{\rho}^{\mathrm{ST}}_j\}$ converges. Then the STOA estimator is defined as using its limit

$$
\hat{\rho}^{\mathrm{STOA}} = \lim_{j\to\infty}\hat{\rho}^{\mathrm{ST}}_j.
$$

Currently, for the proposed STOA estimator, we are unable to derive a rigorous theory concerning the convergence as in the oracle approximation shrinkage (OAS) estimator case in [36]. Our empirical experience in the Simulation Section 5.5, however, demonstrates that the STOA algorithm can approximate the STO estimator reasonably well for a broad range of Σ, regardless of its sparsity.

For the proposed estimators, another issue is to determine the bandwidth k of W in the tapering step for calculating the shrinkage target matrix $W\circ\hat{S}$. We adopt a data-driven approach for estimating k. The procedure is as follows: We randomly split the independent data into two subsets and choose k from a set of candidate values. For each k in the chosen set, $\hat{\rho}^{\mathrm{STOA}}$ is estimated based on one data subset which we call the training data set and then we calculate the distance, e.g. induced by the Frobenius or spectral norm, between the estimated $\hat{\Sigma}$ and the sample covariance matrix computed from the other data set, i.e. the testing data set. Finally the optimal k is determined by the index yielding the smallest distance. Due to the extra step of determining k, the proposed estimator is more computationally expensive than the LW, RBLW, OAS, and MMSE shrinkage oracle estimators. Nonetheless, this validation overload in practice is a minor computational issue because the proposed STO and STOA estimators are quite efficient for any pre-specified k and the computational cost of all shrinkage estimators mentioned in this paper are comparable.

We conclude this section by providing the STOA pseudo-code in Algorithm 2.

Algorithm 2: The STOA algorithm.

```
Input: $\hat{S}$, $k_{\max}$
Output: $\hat{\Sigma}^{\text{STOA}}$
begin
    foreach $k = 0 : 2 : k_{\max}$ do
        Construct the CMT $W$ with bandwidth $k$ as in (4.5) ;
        Initialize from $\hat{S}$ and calculate the optimal shrinkage coefficient by
        iterating (4.21) and (4.22) until convergence ;
        Find the best bandwidth of $W$ and corresponding $\hat{\rho}^{\text{STOA}}$ by the
        minimum prediction error on the test data ;
    end
    Return $\hat{\Sigma}^{\text{STOA}} = (1 - \hat{\rho}^{\text{STOA}})\hat{S} + \hat{\rho}^{\text{STOA}}(W \circ \hat{S})$.
end
```

4.4 Simulation

We simulate the three examples discussed earlier in this paper to study the finite sample size numeric performances of the proposed estimators. We fix $p = 100$ for all models and consider different values of n with $n = \{10, 20, 30, 40, 50\}$. The STOA algorithm is initialized at $\hat{\Sigma}_0 = \hat{S}$ and $\rho = 0.5$. The maximum number of iterations in the STOA algorithm is set to be 10. We compare the proposed STO estimator and its variant STOA with the tapering [20] and several shrinkage estimators including the LW [82], Rao-Blackwellized LW (RBLW) [36], MMSE shrinkage oracle (MMSEO) estimator and its variant oracle approximating shrinkage (OAS) [36] estimator.

4.4.1 Model 1: AR(1) Model

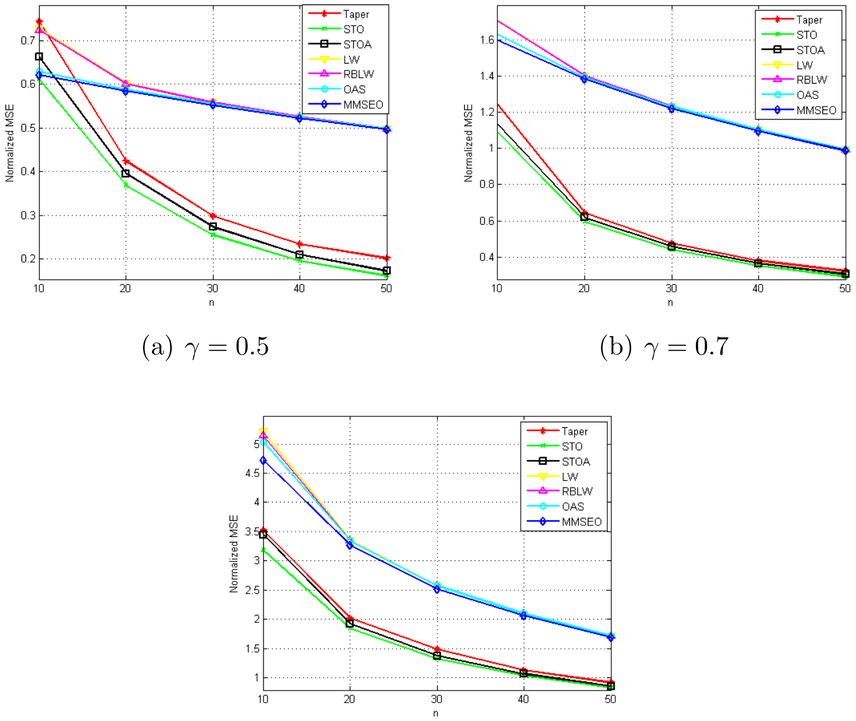
The ith-row and jth-column entry of the covariance matrix Σ is

$$\sigma_{ij} = \gamma^{|i-j|}. \tag{4.23}$$

We chose $\gamma = \{0.5, 0.7, 0.9\}$. A smaller γ essentially makes Σ more like the identity matrix. Note that for any $\gamma \in (0,1)$, $\Sigma \in \mathcal{G}(\alpha, C, C_0)$ for all $\alpha > 0$ and some $C, C_0 > 0$. To specify the tapering bandwidth, we need to determine a proper α by data-driven approaches. Here, our tapering estimator is performed on a training data set over a pre-defined grid and then the optimal α is selected by minimizing the Frobenius loss $\|\hat{\Sigma}_{\text{taper}} - \Sigma\|_F$ on the oracle. (In practical applications, we can use the random splitting scheme discussed above). Estimated normalized MSEs, i.e. the Frobenius risk, and the spectral risk are plotted in Figure 4.2 and Figure 4.3 for various aforementioned estimators.

Several interesting observations can be made from Figure 4.2 and Figure 4.3. First, in terms of estimation risks, the STO, STOA, and tapering estimators uniformly improve upon the previous shrinkage-type estimators including LW, RBLW, OAS, and the MMSEO. This validates our Theorem 4.2.2 and Theorem 4.2.3 on finite sample size data. The improvement is visually appreciable even when n is not so large as considered in the asymptotic setup. Second, the proposed STO and STOA also outperform the tapering estimator, although the improvement is smaller than those from the previous shrinkage-type estimators. The improvement on Frobenius

Figure 4.2: Model 1: The normalized MSE curves as a function of n, averaged over 100 replications. The tapering [20], LW [82], RBLW [36], MMSE shrinkage oracle (MMSEO) [36], and OAS [36] are compared with the proposed STO and STOA estimators.

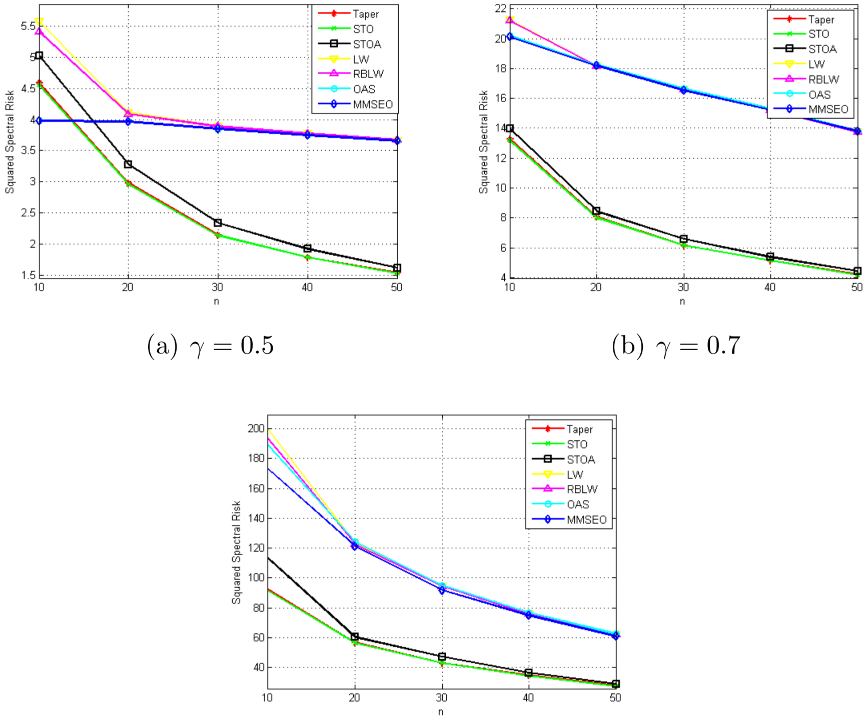


(a) $\gamma = 0.5$

(b) $\gamma = 0.7$

(c) $\gamma = 0.9$

Figure 4.3: Model 1: The spectral risk curves as a function of n, averaged over 100 replications. Here the legends are the same as in Figure 4.2.



(a) $\gamma = 0.5$

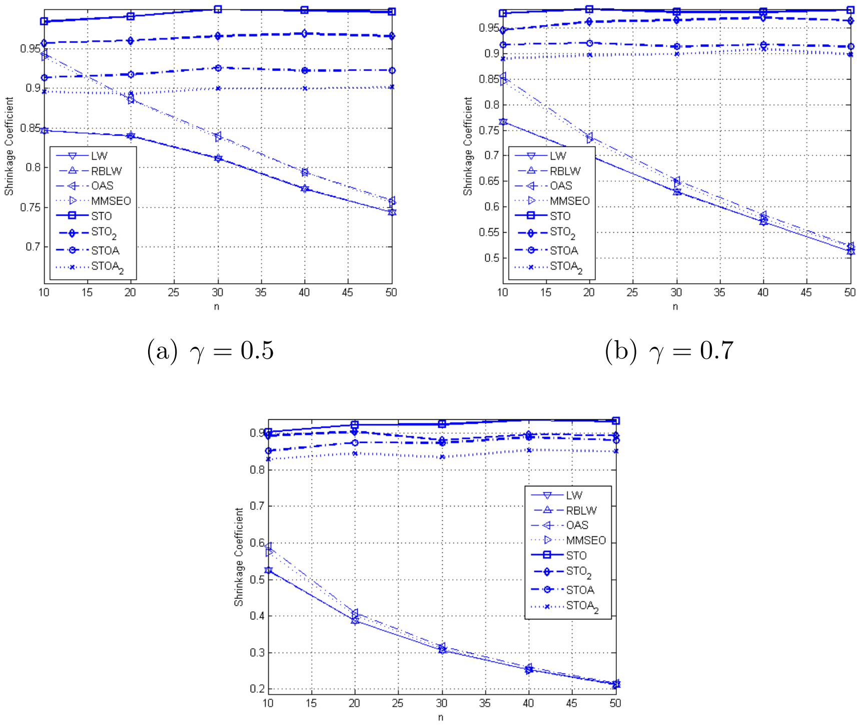
(b) $\gamma = 0.7$

(c) $\gamma = 0.9$

risk is slightly more significant than that on spectral risk. Third, it is clear from these two figures that STOA can well approximate the STO estimator. Finally, despite that the STO estimator minimizes the MSE, the results from the spectral risk are similar to that of the Frobenius risk. It suggests that the STO and STOA are robust against the norm under which the risk is minimized.

The estimated shrinkage coefficients $\hat{\rho}^{\text{STO}}$ and $\hat{\rho}^{\text{STOA}}$ are also plotted in Figure 4.4. It is observed that, in general, the coefficients of STO and STOA are closer to 1 than other shrinkage estimators. This means that STO and STOA essentially use $W \circ \hat{S}$ as the estimator, with a slight adjustment by incorporating information directly from $\hat{S}$. This is confirmative to the theory we have seen since the tapering estimator is minimax. Moreover, shrinkage coefficients of LW, RBLW, MMSEO, and OAS estimators tend to decrease as n increases. This makes sense because the more data collected, the larger amount of information should be used from $\hat{S}$, in which case a

Figure 4.4: Model 1: The estimated shrinkage coefficients for different estimators, averaged over 100 replications. The legends are the same as in Figure 4.2, except that the tapering estimator is excluded and STO/SOTA and STO_2/STOA_2 are the STO/STOA estimates under the Frobenius and spectral risks, respectively.



(a) $\gamma = 0.5$

(b) $\gamma = 0.7$

(c) $\gamma = 0.9$

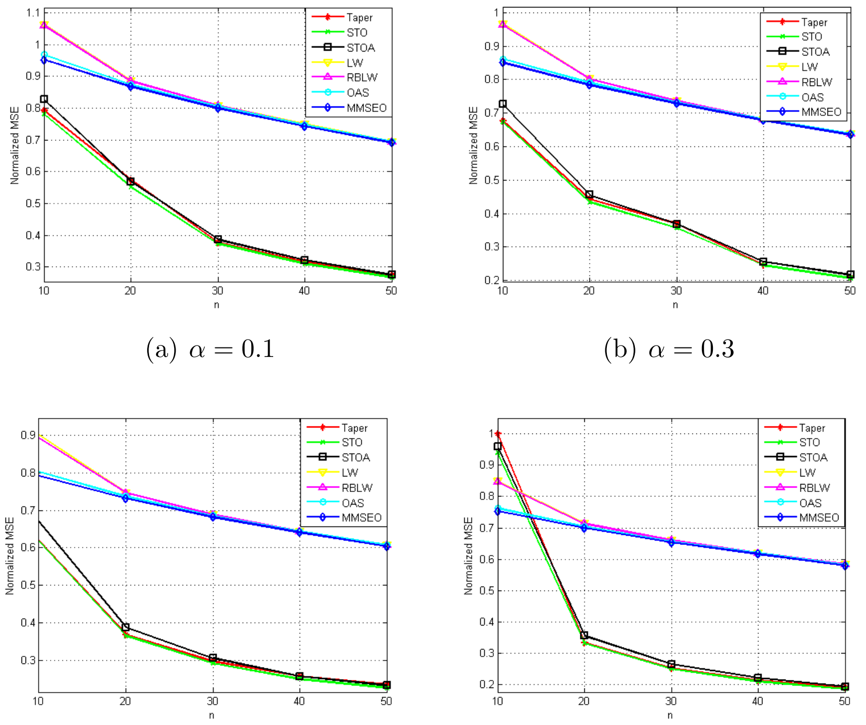
smaller value of ρ_o shall be adaptively chosen. On the contrary, we do not see this observation for the STO and STOA estimators. In fact, their coefficients seem to converge to 1 in this empirical study, as seen in Figure 4.4. This is due to the fact that the shrinkage target, $W \circ \hat{S}$, of these two estimators actually contains the data information. When $W \circ \hat{S}$ is truly optimal, then it is sufficient for the STO and STOA estimators to use only the target component and thus the optimal coefficients converges to 1 in this example.

4.4.2 Model 2: $\Sigma \in \mathcal{G}(\alpha^{-1}, C, C_0)$

We have

$$\sigma_{ij} = \begin{cases} 1 & \text{for } i = j \\ 0.6|i-j|^{-(\alpha+1)} & \text{for } i \neq j \end{cases}, \tag{4.24}$$

Figure 4.5: Model 2: The normalized MSE curves as a function of n, averaged over 100 replications.



(a) $\alpha = 0.1$

(b) $\alpha = 0.3$

(c) $\alpha = 0.5$

(d) $\alpha = 0.7$

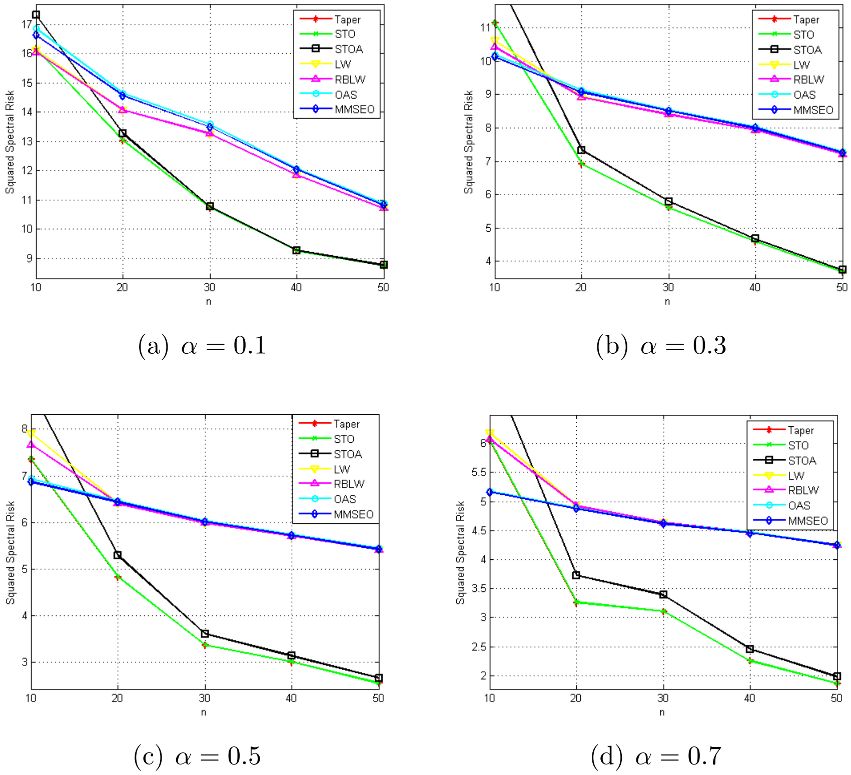
where we choose the smoothing parameter α from $\alpha = \{0.1, 0.3, 0.5, 0.7\}$. We simply set $k = \lfloor n^{1/2(\alpha+1)} \rfloor$ in order to achieve the optimal convergence rate under the Frobineus norm. We remark that the numeric performance can be further improved by cross-validating on a set of bandwidths on the order $n^{1/2(\alpha+1)}$. The risk results of different estimators are shown in Figure 4.5 and Figure 4.6 and the estimated shrinkage coefficients are shown in Figure 4.7. Again, we observe the same pattern on the error curves and essentially same conclusions can be drawn as in Model 1.

4.4.3 Model 3: Fractional Brownian Motion

The numeric performance of the STO and STOA estimators when $\Sigma \notin \mathcal{G}(\alpha, C, C_0)$ is studied in the third model, the FBM model. In our setup, we look at the FBM with the Hurst parameter h selected from $h = \{0.6, 0.7, 0.8, 0.9\}$.

From Figure 4.8, we can see that the normalized MSEs of the MMSE shrinkage

Figure 4.6: Model 2: The spectral risk curves as a function of n, averaged over 100 replications.



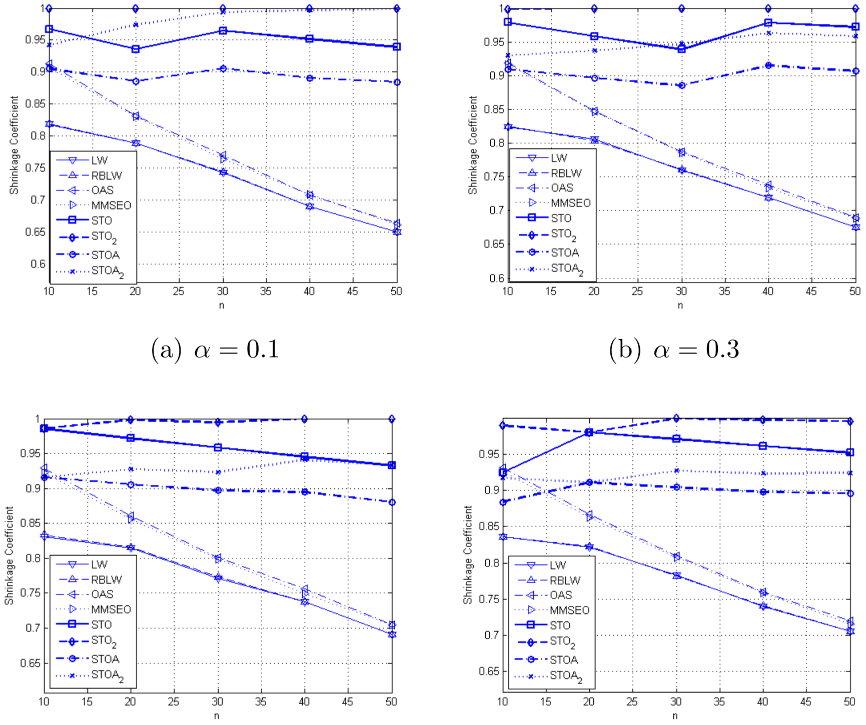
(a) $\alpha = 0.1$

(b) $\alpha = 0.3$

(c) $\alpha = 0.5$

(d) $\alpha = 0.7$

Figure 4.7: Model 2: The estimated shrinkage coefficients for different estimators, averaged over 100 replications.



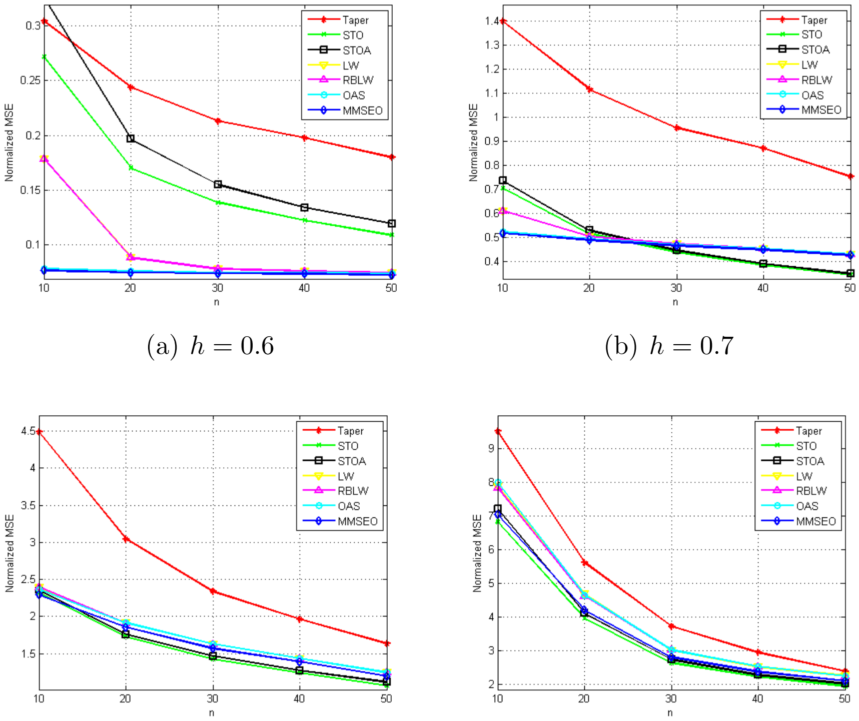
(a) $\alpha = 0.1$

(b) $\alpha = 0.3$

(c) $\alpha = 0.5$

(d) $\alpha = 0.7$

Figure 4.8: Model 3 (FBM): The normalized MSE curves as a function of n, averaged over 100 replications.



(a) $h = 0.6$

(b) $h = 0.7$

(c) $h = 0.8$

(d) $h = 0.9$

estimators are smaller than that of the tapering estimator. This is not surprising because: (i) the assumption $\Sigma \in \mathcal{G}(\alpha, C, C_0)$ is violated and therefore no optimality under the Frobenius risk can be expected in the tapering estimator; (ii) the MMSE estimators are designed to minimize the Frobenius risk. Notwithstanding, when looking at the spectral risk, Figure 4.9, we observe that the risk of the tapering estimator is smaller than those from the MMSE family. Therefore, the tapering estimator is quite robust in the sense that, although being sub-optimal, it still gives better spectral risk performances than the MMSE shrinkage estimators. In contrast, the MMSE shrinkage estimators are sensitive to norms under which the risk performance is measured; in particularly, they are only optimal in the Frobenius norm.

It is observed that the STO and STOA estimators uniformly outperform other shrinkage estimators when $h = 0.8$ and $h = 0.9$. In the case of $h = 0.6$, they are outperformed by LW, RBLW, OAS, and MMSEO estimators but still yield smaller MSEs than the tapering estimator. The case of $h = 0.7$ appears to be non-uniform; however the curve trends shown in Fig 4.8(b) suggest that STO and STOA may eventually yield a smaller Frobenius risk as n gets larger.

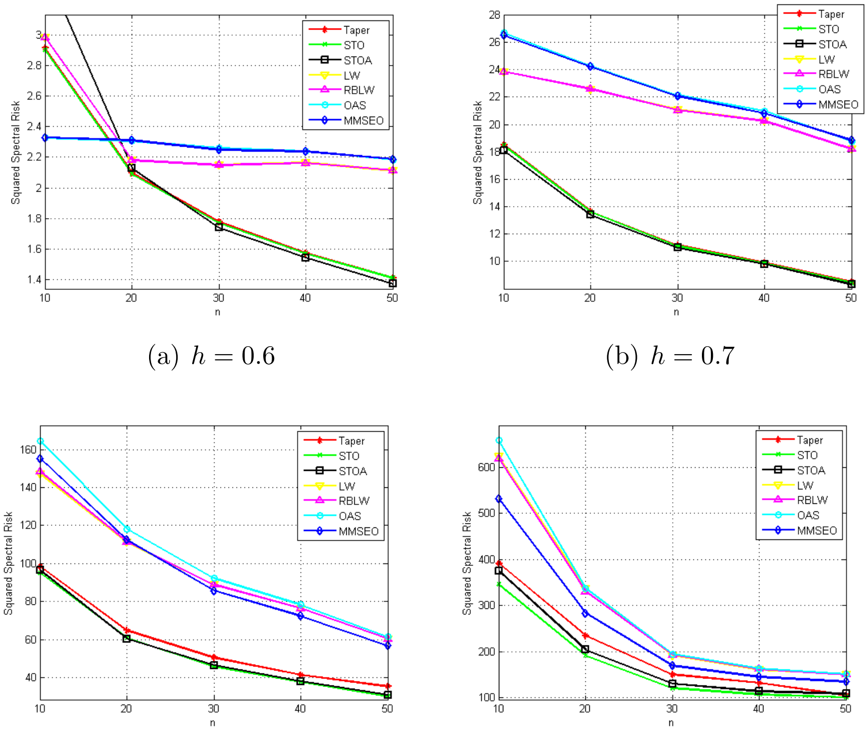
4.5 Conclusion

The main contributions of this chapter are summarized as follows:

1. For high-dimensional covariance estimation problems where $p/n \to \infty$, we showed that the MMSE shrinkage oracle estimator is inconsistent under both Frobenius and spectral risks for some typical covariance matrices in $\mathcal{G}(\alpha, C, C_0)$. Moreover, we showed that the tapering estimator is uniformly superior than the MMSE shrinkage estimator in this case.

2. We proposed a STO estimator that combines the advantages from both the MMSE shrinkage and tapering estimators. In particular, the proposed estimator is suitable for estimating general, high-dimensional covariance matrices. An

Figure 4.9: Model 3 (FBM): The spectral risk curves as a function of n, averaged over 100 replications.



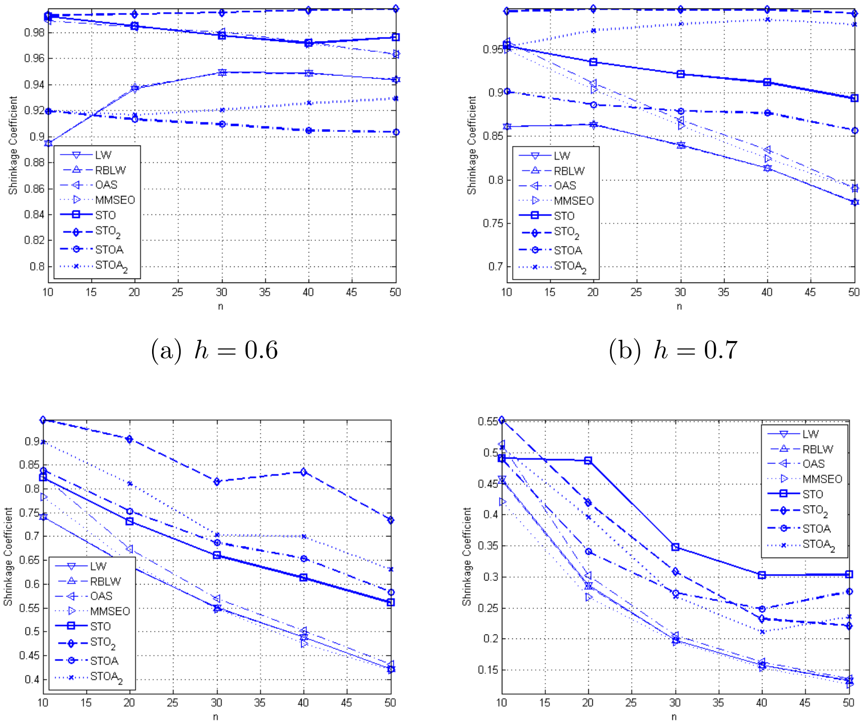
(a) $h = 0.6$

(b) $h = 0.7$

(c) $h = 0.8$

(d) $h = 0.9$

Figure 4.10: Model 3 (FBM): The estimated shrinkage coefficients for different estimators, averaged over 100 replications.



(a) $h = 0.6$

(b) $h = 0.7$

(c) $h = 0.8$

(d) $h = 0.9$

oracle estimator in the closed-form was derived and a practical algorithm to approximate the STO estimator was presented.

Chapter 5

Efficient Minimax Estimation of High-Dimensional Sparse Precision Matrices

5.1 Introduction

In this chapter, we primarily focus on estimating the inverse of the covariance matrix Σ^{-1}, a.k.a. *the precision matrix* Ω, in high-dimensional situations. Estimation of Ω is a more difficulty task than estimating Σ because of the lack of natural and pivotal estimators as $\Sigma_n^\star$ when $p > n$. Nonetheless, accurately estimating Ω has important statistical meanings. For example, in Gaussian graphical models, a zero entry in the precision matrix implies the conditional independence between the corresponding two variables. Further, there are additional concerns in estimating Ω beyond those we have already seen in estimating large covariance matrices; for details see Chapter 1. In light of those challenges in estimating the precision matrix when $p \gg n$, we propose in this paper a new easy-to-implement estimator with attractive theoretic properties and computational efficiency. The proposed estimator is constructed on the idea of the finite Neumann series approximation and constitutes merely matrix multiplication and addition operations. The proposed estimator has a computational complexity of $O(\log(n)p^3)$ for problems with p variables and n observations, representing a significant improvement upon the aforementioned optimization methods. So our estimator is more promising for ultra high-dimensional real-world applications such as gene microarray network modeling.

Remark 11. It is possible to further reduce the computational complexity of the proposed algorithm by employing more sophisticated matrix multiplication algorithms. For instance, the current fastest matrix multiplication algorithm by Coppersmith and Winograd has an asymptotic complexity of $O(p^{2.376})$ [38]. Moreover, by exploring the sparsity structure in the matrices, the complexity can be further reduced to $O(k^{0.7}p^{1.2} + p^{2+o(1)})$ [129], where k is the maximum number of zeros in each of the multipliers. Therefore, we can see that there would be huge computational savings of our algorithm when the covariance matrix is sufficiently sparse.

We now state the assumption regarding the sparse matrices studied in this paper. Our sparse matrix class are built on standard sparse matrices. For $p \gg k$, a p-by-p matrix A with elements a_{ij}'s is said to be *k-sparse* if $A \in S_k$ where

$$S_k = \left\{ A : \sup_j \sum_i \mathbb{I}(a_{ij} \neq 0) \leq k \right\}. \tag{5.1}$$

S_k is a strict class in the sense that there are matrices containing many small entries while they are dense in support and thus excluded from S_k. Therefore, we choose to consider an alternative sparsity measure in terms of the *strong ℓ_q-ball* introduced in [11]. Define

$$\mathcal{G}_q(c_{n,p}) = \left\{ A : \sup_j \sum_i |a_{ij}|^q \leq c_{n,p} \right\}, \tag{5.2}$$

for $0 \leq q < 1$, be the collection of matrices with each column belonging to a strong ℓ_q-ball with size $c_{n,p}$. Note that $\mathcal{G}_q(c_{n,p})$ is closed under the matrix L^1 norm. Think of matrices under consideration are infinite dimensional for a moment. Subset of sparse matrices can be naturally defined as those matrices with finite strong ℓ_q-ball sizes. Therefore, the set $\mathcal{G}_q$ of all possible finite strong ℓ_q-ball volumes is our main target of study

$$\mathcal{G}_q = \bigcup_{c_{n,p} \geq 0} \mathcal{G}_q(c_{n,p}). \tag{5.3}$$

5.1.1 Innovation and Main Results

We summarize the main innovation of this chapter in Theorem 5.1.1, which is an immediate consequence of a series of asymptotic analysis to be reported in Section 5.3. Briefly speaking, we shall describe a computationally efficient algorithm to estimate large precision matrices in a certain approximately inversely closed sparsity class we introduce and show that the resulting estimator is consistent when more and more data are collected. Furthermore, by deriving a lower bound on the estimation error of the precision matrix, the proposed estimator is shown to actually achieve this information-theoretic lower bound and therefore it is rate-optimal.

Theorem 5.1.1. *Assume $p \geq n^{\xi}$ for some $\xi > 0$ and $c_{n,p} \leq C(\log p/n)^{(1-q)/2}$. Then the minimax risk of estimating the precision matrix $\Omega = \Sigma^{-1}$ on $\{\Sigma \in \mathcal{G}_q(c_{n,p}) \cap \mathcal{U}(m)\}$, where $\mathcal{U}(m)$ is defined in (5.8) and the rows of the data $X_{n,p}$ follow i.i.d. sub-Gaussian distribution with covariance Σ, obeys*

$$\inf_{\hat{\Omega}} \sup_{\Sigma \in \mathcal{G}_q(c_{n,p}) \cap \mathcal{U}(m)} E \left\| \hat{\Omega} - \Omega \right\|^2 \asymp c_{n,p}^2 \left(\frac{\log p}{n} \right)^{1-q}, \tag{5.4}$$

where the infimum is taken over all possible estimator $\hat{\Omega} : \mathbb{R}^{n \times p} \to \mathbb{R}^{p \times p}$ for Ω based on the data. Furthermore, the proposed estimator based on the Neumann series representation achieves this minimax risk.

5.1.2 Comparison with Existing Work

It is interesting to observe that the same error bound (5.4) applies to the estimation of the covariance matrix for $\Sigma \in \mathcal{G}_q(c_{n,p})$ [22]. Indeed, a closer examination on our proofs (given in the Appendix) reveals that we actually translate the estimation problem of Ω to the estimation of Σ. Since the latter case is well studied in the literature, there are powerful tools and solid theories to be used for our purpose. Therefore, we will show that our proposed estimator of Ω inherits a large portion of nice theoretic properties from estimation of Σ, such as consistency and rate-optimality [20, 22].

The CLIME estimator of Ω, $\hat{\Omega}_{\text{CLIME}}$ proposed in [18], is the solution of the constrained convex optimization problem

$$\text{minimize } \|\Omega\|_1 \quad \text{subject to } \|\Sigma_n^\star \Omega - I\|_\infty \leq \lambda_n, \tag{5.5}$$

followed by a symmetrization step in order to make $\hat{\Omega}_{\text{CLIME}}$ a self-adjoint matrix. Under a different set of assumptions which are imposed merely on Ω, the CLIME estimator has similar convergence rate as our proposed estimator,

$$E\left\|\hat{\Omega}_{\text{CLIME}} - \Omega\right\|^2 \lesssim c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q}, \tag{5.6}$$

where $c_{n,p}$ is now the size of the precision matrices in their uniform class. Both estimators achieve the optimal convergence rate under the spectral norm; nevertheless, our proposed estimator only consists of thresholding, matrix multiplication and addition operations which require no essential computational overload besides the calculation of $\Sigma_n^\star$. Therefore, the proposed estimator is more computationally efficient than the CLIME estimator and thus can be applied to large-scale precision matrix estimation problems.

Similar spectral/Frobenius norm convergence results in probability were reported for the graphical Lasso and SCAD models in the special case that $q = 0$ [80, 108]. Therefore, our results are more general in the sense that we obtain optimal convergence results for a broader class of sparse matrices in terms of strong ℓ_q balls with small size.

The rest of the chapter is organized as following: Section 5.2 introduces the notion of approximately inverse closeness on the set of sparse matrices and identifies a class of such matrices. Section 5.3 proposes a precision matrix estimator based on the Neumann series representation and proves that it is consistent in probability and in L^2 under the spectral norm. Moreover, the minimax risk of estimating the precision matrix is studied. By comparing the error bound of our proposed estimator with the minimax risk, we show that our estimator is sharp and thus rate-optimal in the sense of minimax risk. In Section 5.4, we discuss the issue of practically determining tuning parameters in the proposed algorithm. Performance comparisons with other optimization based methods using simulations are reported in Section 5.5. A real fMRI application for learning functional brain connectivity of F→STN using the proposed method is presented in Section 5.6. We conclude this paper in Section 5.7 and discuss a few directions for future work.

5.2 Approximately Inversely Closed Sparse Matrices

In order to estimate the precision matrix from multivariate Gaussian observations, it is necessary to assume that Σ is non-singular. We consider the following uniform class

$$\mathcal{U}(m, M) = \{\Sigma \succ 0 : m \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq M\}, \tag{5.7}$$

where $\Sigma \succ 0$ means Σ is strictly positive-definite and $m, M > 0$. Without loss of generality (w.l.o.g.), it is convenient to assume in the sequel that $M = 1/m$ and $0 < m \leq 1$. Therefore, we have the class

$$\mathcal{U}(m) = \{\Sigma \succ 0 : m \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq 1/m\}. \tag{5.8}$$

5.2.1 The Neumann Series Representation of Ω

Let $\Sigma \in \mathcal{U}(m)$. Set $M_\Sigma = \lambda_{\max}(\Sigma)$, $m_\Sigma = \lambda_{\min}(\Sigma)$, and $\eta = 2/(M_\Sigma + m_\Sigma)$. Since Σ is positive-definite and self-adjoint, it is clear that η minimizes $\|I - t\Sigma\|$ over $t > 0$ and

$$\|I - \eta\Sigma\| = \frac{M_\Sigma - m_\Sigma}{M_\Sigma + m_\Sigma} \leq \frac{1 - m^2}{1 + m^2} < 1, \tag{5.9}$$

as shown in [12]. So it follows from the Neumann series expansion that

$$\Omega = \Sigma^{-1} = \eta(\eta\Sigma)^{-1} = \eta[I - (I - \eta\Sigma)]^{-1} = \eta \sum_{j=0}^{\infty}(I - \eta\Sigma)^j \overset{\text{def}}{=} B_r + R_r,$$

where

$$B_r = \eta \sum_{j=0}^{r}(I - \eta\Sigma)^j \tag{5.10}$$

and the residual term is upper bounded by

$$\|R_r\| \leq \eta \sum_{j=r+1}^{\infty} \|I - \eta\Sigma\|^j = \frac{1}{m_\Sigma}\left(\frac{M_\Sigma - m_\Sigma}{M_\Sigma + m_\Sigma}\right)^{r+1} \leq \frac{1}{m}\left(\frac{1-m^2}{1+m^2}\right)^{r+1} \to 0 \tag{5.11}$$

uniformly in $\Sigma \in \mathcal{U}(m)$ as $r \to \infty$.

For our problem of estimating sparse precision matrices, the Neumann representation of Ω motivates us to identify a class of sparse matrices such that the inverse of its any member can be approximated with arbitrary accuracy by elementary linear combinations using only a finite number of members in the class. This idea shall be rigorously formalized in the next section where the notion of *approximately inverse closeness* is introduced.

5.2.2 A Class of Sparse Matrices with Approximately Inverse Closeness

The main contribution of this section is to introduce a class of approximately sparse matrices whose inverses are also approximately sparse.

Definition 5.2.1. The set of all k-sparse matrices that are at most ε-distance from A is defined as:

$$\text{Sparse}_\varepsilon(A, k) = \{B : \|B - A\| \leq \varepsilon \text{ and } B \in S_k\}\,. \tag{5.12}$$

Moreover, we generalize the definition to consider the family of all k-sparse matrices that are within ε-distance from $\mathcal{G}_q(c_{n,p})$

$$\mathcal{S}_\varepsilon(\mathcal{G}_q(c_{n,p}), k) = \bigcup_{A \in \mathcal{G}_q(c_{n,p})} \text{Sparse}_\varepsilon(A, k). \tag{5.13}$$

It is clear from the definition that $B \in \mathcal{S}_\varepsilon(\mathcal{G}_q(c_{n,p}), k)$ if and only if $B \in S_k$ and

$$\text{dist}\,(B, \mathcal{G}_q(c_{n,p})) \leq \varepsilon. \tag{5.14}$$

Finally, let

$$\mathcal{S}_\varepsilon(\mathcal{G}_q(c_{n,p})) = \bigcup_{k \geq 0} \mathcal{S}_\varepsilon(\mathcal{G}_q(c_{n,p}), k) \tag{5.15}$$

be the collection of sparse matrices that can approximate $\mathcal{G}_q(c_{n,p})$ with error ε.

Lemma 5.2.1. *Given $0 \leq q < 1$, $c_{n,p} > 0$, and $\varepsilon > 0$, we let $A \in \mathcal{G}_q(c_{n,p})$ and*

$$k_{\min} = \left[\frac{q}{(1-q)\varepsilon} c_{n,p}^{\frac{1}{q}}\right]^{\frac{q}{1-q}}. \tag{5.16}$$

Then the following statements hold:

1. *There exists $B \in Sparse_\varepsilon(A, k)$ for all $k \geq k_{\min}$, i.e. $Sparse_\varepsilon(A, k)$ is non-empty.*

2. *For each $r \in \mathbb{N}$,*

   $$A^r \in \mathcal{G}_q(C(q)^r c_{n,p}^r).$$

   In particular, for every

   $$k' \geq k_{\min} C(q)^{\frac{r}{1-q}} c_{n,p}^{\frac{r-1}{1-q}}. \tag{5.17}$$

   there exists $B' \in Sparse_\varepsilon(A^r, k')$.

Definition 5.2.2. A collection $\mathcal{S}$ of invertible elements (i.e. S^{-1} exists for all $S \in \mathcal{S}$) is said to be *inversely closed* if $S^{-1} \in \mathcal{S}$ for all $S \in \mathcal{S}$.

It has been shown in [12] that there is a class of band dominated matrices that is inversely closed. In general, for sparse matrices, raising arbitrarily high powers for a sparse matrix can cause the elements mixing so well that the sparsity of its inverse gets violated. Therefore, we define below a relaxed version of the inverse closeness.

Definition 5.2.3. A collection $\mathcal{S}$ of invertible elements is said to be *approximately inversely closed* if for any $\varepsilon > 0$ and $S \in \mathcal{S}$, there is an $S' \in \mathcal{S}$ such that

$$\operatorname{dist}\left(S^{-1}, S'\right) \leq \varepsilon. \tag{5.18}$$

Now, consider the uniform class

$$\mathcal{F}(q, m) = \mathcal{G}_q \cap \mathcal{U}(m). \tag{5.19}$$

It consists of all *bounded* linear functionals $\mathbb{R}^p \to \mathbb{R}^p$ that are: a) within ε-distance of the set $\mathcal{G}_q(c_{n,p})$; b) uniformly bounded away from singularity and thus invertible; c) permutation-invariant. Beyond these facts, there is one more crucial yet less obvious fact that $\mathcal{F}(q, m)$ is actually approximately inversely closed. This is the main theorem of this section:

Theorem 5.2.2. *$\mathcal{F}(q, m)$ is approximately inversely closed. Suppose $\Sigma \in \mathcal{G}_q(c_{n,p}) \cap \mathcal{U}(m)$, then for any $\varepsilon > 0$, there exists*

$$k' = k_{\min} C(q)^{\frac{r}{1-q}} c_{n,p}^{\frac{r-1}{1-q}} \tag{5.20}$$

for sufficiently large r such that $Sparse_{\varepsilon}(\Omega, k')$ is non-empty.

We now summarize here the assumptions assumed in this paper: a) both n and p diverge to ∞ and $n \lesssim p^{\xi}$; b) the observation data $X = \{\mathbf{x}_i\}_{i=1}^{n}$ are sampled from some i.i.d. sub-Gaussian distribution (5.21) with covariance matrix Σ; c) $\Sigma \in \mathcal{G}_q(c_{n,p})$ for some $c_{n,p} > 0$ such that the precision matrix Ω can be approximated within ε distance by some sparse-k' matrix.

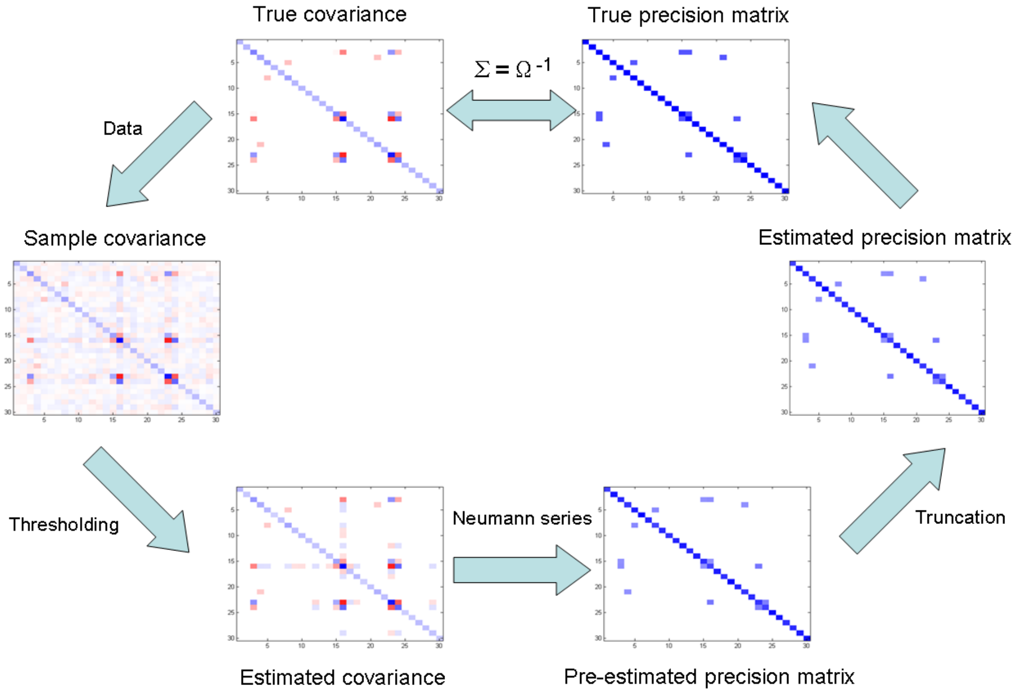


Figure 5.1: The diagram illustrating our proposed Algorithm 3. White spots correspond to zero entries, blue ones are the positive entries, and red ones are negative entries in the matrix. For a sparse Σ with approximately inverse closeness, its inverse Ω can be approximated by the finite Neumann series representation. So consistent procedures such as thresholding applied to $\Sigma_n^\star$ can be used to form a "good" estimator for Ω. Since Ω is not necessarily sparse in the true sense (5.1), an additional truncation is applied to estimate Ω because its strong ℓ_q volume (5.2) can be controlled, e.g. by Theorem 5.2.2.

5.3 Proposed Estimator

5.3.1 Algorithm

The proposed algorithm is based on the Neumann series representation of Ω and the *thresholding operator* T_t defined through $T_t(\Sigma)(i,j) = \sigma_{ij}\mathbb{I}(|\sigma_{ij}| > t)$. A diagram illustrating main ideas of the proposed algorithm is shown in Figure 5.1.

Now, the proposed estimator $\hat{\Omega}$ is described in Algorithm 3.

```
Algorithm 3: The proposed algorithm
Input: Sample covariance matrix $\Sigma_n^\star$, thresholding cutoff $t$, number of
       truncated Neumann series terms $r$, approximation tolerance $\varepsilon$.
Data: $X$.
Output: $\hat{\Omega}$
begin
    Compute $\tilde{\Sigma}_n = T_t(\Sigma_n^\star)$.
    Set $\tilde{\Omega} = \eta \sum_{j=0}^{r} (I - \eta \tilde{\Sigma}_n)^j$.
    Truncate $\tilde{\Omega}$ according to (50) such that $\hat{\Omega} \in \text{Sparse}_\varepsilon(\tilde{\Omega}, k')$, where $k'$ is
    given in (5.20) (i.e. keep the largest $k$ entries of each column of $\tilde{\Omega}$ and set
    others to zeros).
end
```

The output of the above algorithm is an $S_{k'}$ matrix due to the last truncation step. With properly determined ε, we can show that the truncation step does not affect the error bounds when it is removed from the algorithm. The purpose of adding this step is to promote sparsity in the true sense such that the estimator $\hat{\Omega}$ can be clearly interpreted. For instance, a zero ω_{ij} is interpreted as the conditional independence between two variables in the Gaussian graphical model (and hence a missing edge in the graph of $\hat{\Omega}$).

It is still left open to determine the input parameters t and r and estimate η as well (ε can be chosen as a small value). For practical choices of these parameters, we adopt common data-driven approaches such as cross-validation. Regarding this issue of parameter tuning, we will provide the details in Section 5.4.

5.3.2 Consistency

Here following similar ideas as in [11, Theorem 1], we show that the proposed estimator $\hat{\Omega}$ is consistent. In fact, we shall show its consistency for data X generated by some i.i.d. *sub-Gaussian* distribution (a.k.a. rows of X are i.i.d. sub-Gaussian vectors) which is a slightly more general result than $\mathbf{x}_i \sim N(\mathbf{0}, \Sigma)$. More specifically, we assume that there exist absolute constants $C_1, C_2 > 0$ such that

$$P\left(\left|\mathbf{v}^T \mathbf{x}_i\right| \geq t\right) \leq C_1 \exp\left(-\frac{t^2}{2C_2}\right) \tag{5.21}$$

holds for all $t > 0$ and $\|\mathbf{v}\| = 1$. Assume, w.l.o.g., $E\mathbf{x}_i = \mathbf{0}$. With this sub-Gaussian assumption (5.21), standard concentration of measure results (e.g. see [120, Proposition 16 and Corollary 17]) enable us to bound the tail probability of $(\sigma^\star_{ij} - \sigma_{ij})$ in a mixture of exponential and Gaussian-like decaying rate. More precisely, we have

Proposition 5.3.1. *Suppose X is sub-Gaussian obeying (5.21) with covariance $\Sigma \in \mathcal{U}(m)$. Then there exist constants $C_3, C_4, C_5 > 0$, all depending only on C_1 and C_2 in (5.21) and m, such that*

$$P\left(\left|\sigma^\star_{ij} - \sigma_{ij}\right| \geq t\right) \leq C_3 \exp\left[-C_4 \min\left(\frac{t^2}{C_5^2}, \frac{t}{C_5}\right) n\right] \tag{5.22}$$

for all $t \geq 0$.

Focusing on a *neighborhood of zero*, we can see that the large deviation result in [22, Eq.(25)] immediately follows from Proposition 5.3.1.

Corollary 5.3.2. *Under assumptions in Proposition 5.3.1,*

$$P\left(\left|\sigma^\star_{ij} - \sigma_{ij}\right| \geq t\right) \leq C_3 \exp\left(-\frac{8nt^2}{C_4^2}\right) \tag{5.23}$$

for $|t| \leq C_5$.

Then applying [11, Theorem 1] we obtain the following theorem.

Theorem 5.3.3. *Suppose X is sub-Gaussian obeying (5.21) with covariance $\Sigma \in \mathcal{G}_q(c_{n,p}) \cap \mathcal{U}(m)$. Assume $\log p/n \to 0$ and $c_{n,p}(\log p/n)^{(1-q)/2} \to 0$, as $n \to \infty$.*

Choose the threshold parameter $t = \tau\sqrt{\log p/n}$ *for some large* τ*. Then we have, uniformly in* $\Sigma \in \mathcal{G}_q(c_{n,p}) \cap \mathcal{U}(m)$,

$$\left\|\hat{\Omega} - \Omega\right\| \leq C(q, m, \tau)c_{n,p}\left(\frac{\log p}{n}\right)^{(1-q)/2} + \frac{1}{m}\left(\frac{1-m^2}{1+m^2}\right)^{r+1}, \tag{5.24}$$

with probability greater than $1 - C_6 p^{-8\tau^2/C_4^2+2}$ *approaching to 1 whenever* $\tau > C_4/2$*. Here,* r *is the number of terms in the Neumann series pre-estimator* $\tilde{\Omega}$ *of* $\hat{\Omega}$*.*

An immediate consequence of Theorem 5.3.3 is that $\hat{\Omega}$ is a consistent estimator of Ω whenever $c_{n,p}(\log p/n)^{(1-q)/2} = o(1)$ as $n \to \infty$, and the optimal number of terms in the Neumann power series is chosen such that the two terms on the right hand side of (5.24) match on the same magnitude order. Therefore, we have

Corollary 5.3.4. *Let*

$$r = \left\lceil \frac{(1-q)(\log n - \log\log p) - 2\log c_{n,p} - 2\log C}{2[\log(1+m^2) - \log(1-m^2)]} - 1 \right\rceil, \tag{5.25}$$

for some $C > 0$*. Under the same assumptions as in Theorem 5.3.3, then for every* $\Sigma \in \mathcal{G}_q(c_{n,p}) \cap \mathcal{U}(m)$,

$$\left\|\hat{\Omega} - \Omega\right\| \xrightarrow{P} 0, \tag{5.26}$$

as $n \to \infty$*.*

Corollary 5.3.4 follows straightforward after a few algebra based on Theorem 5.3.3; thus we omit its proof. We also show below the entry ∞-norm consistency of $\hat{\Omega}$ which shall be particularly helpful for developing the model selection consistency shortly.

Proposition 5.3.5. *Suppose* X *is sub-Gaussian obeying (5.21) with covariance* Σ*. Assume* $\log p/n \to 0$ *as* $n \to \infty$*. Choose the threshold parameter* $t = \tau\sqrt{\log p/n}$ *for some large* τ*. Then we have, uniformly in* $\Sigma \in \mathcal{G}_q(c_{n,p}) \cap \mathcal{U}(m)$,

$$\left\|\hat{\Omega} - \Omega\right\|_\infty \leq C(q, m, \tau)\left(\sqrt{\frac{\log p}{n}} + \delta^{r+1}\right), \tag{5.27}$$

with probability greater than $(1 - p^{-8\tau^2/C_4^2+2})$ *which asymptotically approaches to 1 whenever* $\tau > C_4/2$*; here* $\delta = (1-m^2)/(1+m^2)$*.*

Remark 12. Proposition 5.3.5 states that the maximal fluctuation of the estimator $\{\hat{\omega}_{ij}\}$ about $\{\omega_{ij}\}$ can be well controlled. Therefore, when the magnitudes of non-zero entries of Ω are uniformly bounded away from zero, we can recover the support of Ω by cutting $\hat{\Omega}$ with a properly determined threshold t' according to (5.27). The cutoff t' can be chosen such that the recovery is successful with probability tending to 1, as more and more data are available.

Based on Theorem 5.3.3, we also provide the consistency of our proposed estimator under the mean squared loss, which shall be useful to establish the upper bound for optimality in the sense of minimax risk. We mention that establishing the upper bound of estimation error is considerably more involved than the weaker claim where the same bound is stated in probability.

Theorem 5.3.6. *Under assumptions in Theorem 5.3.3 and in addition assuming* $p \geq n^{\xi}$ *for some* $\xi > 0$ *and* $r = O(\log n)$*, we have*

$$\sup_{\Sigma \in \mathcal{G}_q(c_{n,p}) \cap \mathcal{U}(m)} E\left\|\hat{\Omega} - \Omega\right\|^2 \lesssim c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q} + \delta^{2(r+1)}. \tag{5.28}$$

One consequence of Theorem 5.3.6 is

Corollary 5.3.7. *With assumptions in Corollary 5.3.4,*

$$\sup_{\Sigma \in \mathcal{G}_q(c_{n,p}) \cap \mathcal{U}(m)} E\left\|\hat{\Omega} - \Omega\right\|^2 \lesssim c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q} \to 0, \tag{5.29}$$

as $n \to \infty$*.*

Remark 13. Since convergence in L^2 implies $\xrightarrow{P}$, it is now clear that Corollary 5.3.4 is an immediate consequence of Corollary 5.3.7.

5.3.3 Sharpness: Optimal Under Minimax Risk

We have so far seen that the proposed estimator is consistent under the spectral norm. In fact, we will also show that it is rate-optimal in the sense of minimaxity among all estimators of Ω. In order to establish the minimaxity of the proposed estimator, it suffices, thanks to Corollary 5.3.7, to find a lower bound on the order of $c_{n,p}^2 \left(\log p/n\right)^{1-q}$ for the mean squared loss. Then this would imply that our proposed estimator is sharp and thus rate-optimal. Now, we tackle this task by assembling ideas in [20–22]. More specifically, we appeal to a general lower bound argument for the minimax risk of estimating the sparse *covariance* matrix, see [22, Lemma 3], and then convert the optimality in terms of *precision* matrix as in [20]. For the completeness of our proof, it is worthy to spend a few paragraphs to describe the construction setup and recap the related lemma, from which our lower bound can follow easily.

The lower bound is established via a carefully constructed finite set of the least favorable multivariate normal distributions that are in the sparse uniform class $\mathcal{G}_q(c_{n,p}) \cap \mathcal{U}(m)$. For $1 \leq J \leq p$, we denote $B \subset \mathbb{R}^p \setminus \{\mathbf{0}\}$ as a non-zero subset in the p-dimensional vector space. Let $\Gamma = \{0,1\}^J$ be the vertex set of the J-dimensional cube and $\Lambda \subset B^J$. Define the product parameter space Θ by

$$\Theta = \Gamma \times \Lambda = \left\{\theta = (\gamma, \lambda) : \gamma \in \Gamma, \lambda \in \Lambda\right\}. \tag{5.30}$$

Then every $J \times p$ matrix D can be identified by a mapping $\Theta \to \mathbb{R}^{J\times p}$ such that the parametrization of D by $\theta = (\gamma, \lambda)$ is interpreted as follows: D is formed by stacking each of the p-dimensional vector $\gamma_1\lambda_1, \cdots, \gamma_J\lambda_J$, where $\lambda = (\lambda_1, \cdots, \lambda_J)$. Moreover, a set of *association operators* $A_j : \mathbb{R}^p \to \mathbb{R}^{p\times p}$ are defined via $M = A_j(\mathbf{b})$ such that the j-th row and column of M are equal to $\mathbf{b}$ and other entries of M are all zeros. Now, combining the product structure of Θ and the association operators, we define a function $\Sigma : \theta \in \Theta \mapsto \Sigma(\theta) \in \mathbb{R}^{p\times p}$ as

$$\Sigma(\theta) = I + \varepsilon \sum_{j=1}^{J} \gamma_j A_j(\lambda_j). \tag{5.31}$$

For $J \leq \lceil p/2 \rceil$ and a set of $\{\lambda_j\}_{1\leq j\leq J}$ where each λ_j has zeros in the first $(p - J)$ positions and the rest entries have either 0 or 1 such that $\|\lambda_j\|_0 = k$ (for some k to be assumed bounded appropriately), it is easy to see that $\Sigma(\theta)$ is an anti-block-diagonal matrix except on the main diagonal where $\sigma_{jj}(\theta)$'s$= 1$. This constitutes the collection of the least favorable multivariate normal distributions that hopefully attain the worst estimation error. Now, we are ready to define a family of matrices that shall be used to obtain the minimax risk lower bound via (5.31):

$$\mathcal{G}_0 = \{\Sigma = \Sigma(\theta) : \theta \in \Theta\}. \tag{5.32}$$

It is not hard to verify that, for a carefully chosen k and ε small and depending on $c_{n,p}$, we have $\mathcal{G}_0 \subset \mathcal{G}_q(c_{n,p})$ and $\mathcal{G}_0 \subset \mathcal{U}(m)$, where the latter is because of diagonal dominance. Finally, to complete the setup, for a given $b \in \{0, 1\}$, a mixture distribution $\bar{P}_{j,b}$, associated with Θ, can be defined as

$$\bar{P}_{j,b}(X; \Theta) = \frac{1}{2^{J-1}|\Lambda|} \sum_{\{\theta\in\Theta:\gamma_j(\theta)=b\}} P(X \mid \Sigma(\theta)), \tag{5.33}$$

where $\gamma_j(\theta)$ projects $\theta = (\gamma, \lambda)$ to the j-th element of γ. Essentially, (5.33) defines a mixture of distributions over all parameter sets with a common projection onto γ_j.

With the above notation, [22, Lemma 3] gives a general lower bound of estimating sparse covariances under the L^2 loss.

Proposition 5.3.8. *(Cai and Zhou [22]) Let Θ be the parameter space of θ constructed in (5.30) and (5.31). For an arbitrary function $\psi(\theta)$, let U be any estimator* *of $\psi(\theta)$ from data X generated from the probability family $\{P_\theta : \theta \in \Theta\}$. Then*

$$\max_{\theta\in\Theta} E_\theta \left\|U - \psi(\theta)\right\|^2 \geq \frac{\alpha}{4}\frac{J}{2} \min_{1\leq j\leq J} \left\|\bar{P}_{j,0} \wedge \bar{P}_{j,1}\right\|, \tag{5.34}$$

where

$$\alpha = \min_{(\theta,\theta'):H(\gamma(\theta),\gamma(\theta'))\geq 1} \frac{\left\|\psi(\theta) - \psi(\theta')\right\|^2}{H(\gamma(\theta),\gamma(\theta'))} \tag{5.35}$$

and $H(\gamma,\gamma')$ is the Hamming distance defined on $\{0,1\}^J$

$$H(\gamma,\gamma') = \sum_{j=1}^{J} |\gamma_j - \gamma'_j|. \tag{5.36}$$

In light of Theorem 5.3.3 and Theorem 5.3.6, we have similar results for estimating the precision matrix as those obtained for estimating covariance matrix [22, Theorem 1].

Theorem 5.3.9. *Suppose $c_{n,p} \leq C(\log p/n)^{(1-q)/2}$. Then the minimax risk of estimating the precision matrix $\Omega = \Sigma^{-1}$ on $\{\Sigma \in \mathcal{G}_q(c_{n,p}) \cap \mathcal{U}(m)\}$, where data $X_{n,p}$ are i.i.d. sub-Gaussian with covariance Σ, obeys*

$$\inf_{\hat{\Omega}} \sup_{\Sigma\in\mathcal{G}_q(c_{n,p})\cap\mathcal{U}(m)} E\left\|\hat{\Omega} - \Sigma^{-1}\right\|^2 \geq Cc_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q}, \tag{5.37}$$

where the infimum is taken over all possible estimator $\hat{\Omega}$ for Ω and here the constant C is independent of n and p.

Now, it is clear that our main Theorem 5.1.1 is a direct consequence of Corollary 5.3.7 and Theorem 5.3.9.

5.3.4 Model Selection Consistency

We have so far shown the matrix L^2 norm consistency of the proposed estimator. Estimation consistency in the spectral norm does not imply the model selection consistency and vice versa. Here by *model selection consistency* we mean that

$$P\left(\{\hat{\omega}_{ij} \neq 0\} = \{\omega_{ij} \neq 0\}\right) \to 1, \tag{5.38}$$

as $n \to \infty$. Therefore, it is interesting to ask whether or not the proposed estimator can accurately recover the *bona fide* statistical structures. This is of particular importance when we need to identify the structure in graphical models. To establish the model selection consistency, it suffices to have the consistency of estimator under the entry ∞-norm. As we have seen that this is indeed the case in Proposition 5.3.5. Hence, it leads to the following theorem.

Theorem 5.3.10. *Let* $\underline{\omega} = \min\{|\omega_{ij}| : \omega_{ij} \neq 0\}$ *and*

$$r = \left\lceil \frac{\log n - \log\log p - 2\log C}{2[\log(1+m^2) - \log(1-m^2)]} - 1 \right\rceil, \tag{5.39}$$

for some $C > 0$*. Suppose* $\underline{\omega} > 2t'$*. Then* $T_{t'}(\hat{\Omega})$ *is model selection consistent.*

In fact, it is clear that Theorem 5.3.10 implies the signs of $\{\omega_{ij}\}$ can also be recovered with high probability.

5.3.5 Extensions

The proposed algorithm framework described in Algorithm 3 can be easily extended to broader situations where additional information regarding the sparsity is available. For instance, when there is an ordering structure between variables such as in autoregression (AR) models, the covariance has a bandable structure. By incorporating this additional information, the proposed estimator can be modified such that even better theoretic properties can be achieved. More precisely, by replacing the thresholding operator in the current Algorithm 3 with the *tapering operator*, the optimal convergence rate

$$\min\left\{n^{-2\alpha/(2\alpha+1)} + \frac{\log p}{n}, \frac{p}{n}\right\} \tag{5.40}$$

can be accomplished by adapting our arguments in Theorem 5.3.3 and 5.3.6 to accommodate the optimality results found for the estimator of Σ in [20]. Here α is a sparsity control parameter specifying the rate of decay of entries moving away from diagonals. We can show, by essentially the same arguments, that this tapering estimator is minimax for the covariance matrices of this form.

The proposed framework can also be extended to the adaptive thresholding case, which has been shown to yield better numeric performances on real data when the homoscedastic assumption appears too restrictive [17]. By allowing location-specific cutoffs which could be estimated in a data-driven means, the adaptive thresholding procedure is shown to be optimal when the heteroscedasticity is indeed present. In this case, the universal (non-adaptive) thresholding operator is suboptimal.

In view of these feasible extensions which can properly handle different real problems, we can see that the proposed algorithm is in fact a quite general and flexible framework. Moreover, the framework can be easily adapted in a way such that various optimality may be achieved while the attractive low computational complexity is maintained.

5.4 Practical Choices of η, r and τ

The construction of the proposed estimator involves determining the parameters η, r and τ.

For the choice of η, we use the estimation $\hat{\eta} = 2/(M_{\tilde{\Sigma}_n} + m_{\tilde{\Sigma}_n})$. Note that the thresholding operator T_t does not necessarily preserve positive-definiteness which means $m_{\tilde{\Sigma}_n} < 0$; nevertheless, T_t with a carefully chosen t can preserve positive-definiteness with high probability [11]. In the exceptional case where $m_{\tilde{\Sigma}_n} < 0$, we can simply project negative eigen-values to 0 and the error bound of approximating Σ remains unchanged (except for a factor of 2).

To determine r and τ, we employ a *random splitting* procedure on a two-dimensional grid $j = (j_1, j_2)$. n data points are randomly partitioned into two sets: one training set of size n_1 and one test set of size $n_2 = (n - n_1)$. The precision matrix is estimated on a collection of tuning parameters r and τ and the optimal tuning parameters $\hat{r}$ and $\hat{t}$ are then determined by maximizing the normal log-likelihood on the test data set, up to an additive constant,

$$\begin{aligned}(\hat{j}_1, \hat{j}_2) &= \arg \max_{j=(j_1,j_2)} \text{log-likelihood}\left(\hat{\Omega}_{n_1, r_{j_1}, t_{j_2}}; \Sigma^{\star}_{n_2}\right) \\ &= \arg \min_{j=(j_1,j_2)} \left\{\text{trace}\left(\Sigma^{\star}_{n_2}\hat{\Omega}_{n_1, r_{j_1}, t_{j_2}}\right) - \log\det(\hat{\Omega}_{n_1, r_{j_1}, t_{j_2}})\right\}, \qquad (5.41)\end{aligned}$$

where $\hat{\Omega}_{n_1, r_{j_1}, t_{j_2}}$ represents the proposed estimate of the precision matrix Ω on the training data with tuning parameters r_{j_1} and t_{j_2}, and $\Sigma^{\star}_{n_2}$ means the sample covariance calculated on the test data.

5.5 Numerical Experiments

In this section, some simulations are conducted to evaluate the performance of the proposed estimator. We consider a Toeplitz model of Ω with entries ω_{ij}'s decaying exponentially fast as they moving away from diagonals. This kind of models arises naturally in time-series data analysis where a natural ordering on variables is present. More specifically, we choose $\omega_{ij} = a^{|i-j|}$ with $a = 0.6$ for our setup. This is tantamount to assume a moving average model by noting that the covariance matrix Σ has a *band* structure

$$\sigma_{ij} = \begin{cases} (1+a^2)/(1-a^2), & \text{for } i = j, \\ -a/(1-a^2), & \text{for } i \neq j. \end{cases} \qquad (5.42)$$

In our case, we therefore have $\sigma_{ii} = 2.1250$ (except for the first and last diagonal elements) and $\sigma_{ij} = -0.9375$ for $i \neq j$. Note that this is also one of the models considered in [18]. We compare the performance of our algorithm with several mainstream optimization-based methods including CLIME, graphical Lasso, and SCAD. Since performances of those optimization methods on this model have been thoroughly studied and become virtual standards, e.g. in accordance to [18] and [108], in

Table 5.1: Estimation error under the spectral norm, specificity, and sensitivity of $\hat{\Omega}, \hat{\Omega}_{\text{taper}}$, CLIME, graphical Lasso (GLasso), and SCAD for $n = 100$.

p	Spectral Norm Loss				
	$\hat{\Omega}$	$\hat{\Omega}_{\text{taper}}$	CLIME	GLasso	SCAD
30	2.28	1.25	2.28	2.48	2.38
60	2.78	1.44	2.79	2.93	2.71
90	2.90	1.60	2.97	3.07	2.76
120	2.97	1.67	3.08	3.14	2.79
200	3.01	1.78	3.17	3.25	2.83

p	Specificity %				
	$\hat{\Omega}$	$\hat{\Omega}_{\text{taper}}$	CLIME	GLasso	SCAD
30	55.28	99.99	78.69	50.65	99.26
60	90.11	100.00	90.37	69.47	99.86
90	99.86	100.00	94.30	77.62	99.88
120	99.88	100.00	96.45	81.46	99.91
200	99.88	100.00	97.41	85.36	99.92

p	Sensitivity %				
	$\hat{\Omega}$	$\hat{\Omega}_{\text{taper}}$	CLIME	GLasso	SCAD
30	56.28	67.20	41.07	60.02	16.93
60	21.54	62.08	25.96	41.72	12.72
90	11.81	59.13	20.32	33.70	11.94
120	11.30	58.94	17.16	29.32	11.57
200	10.87	57.06	15.03	25.34	11.07

the same setup as ours, we only run our algorithm and directly compare with their performances reported under the same setting (see Table 1, 2, and 3 in [18]).

We synthesize 100 training data points and another independent 100 test data points. p is varied over $\{30, 60, 90, 120, 200\}$. We tune the parameters t and r on a two-dimensional grid. We choose $t = \tau\sqrt{\log p/n}$ for τ ranging over $[0, 2]$ evenly spaced with interval size 0.2, and choose r to be integers in the region $[0, 3\lceil \log n \rceil]$. Then the optimal parameter pair $(\hat{t}, \hat{r})$ are determined by maximizing the normal log-likelihood (5.41) on the test data. We further set the *effective zero* level to be 10^{-3}, meaning that estimated entries with absolute values below this level are thought as zeros. All performance results are averaged over 100 simulations and they are shown in Table 5.1 and Figure 5.2, along with the performances reported in [18, Table 1, 2, and 3] for comparison.

First, we examine the estimation performances in terms of the spectral norm loss, the specificity, the sensitivity, and the Mathews correlation coefficients (MCC). From Table 5.1, we see that the proposed estimator $\hat{\Omega}$ improves upon CLIME and graphical Lasso, while being slightly outperformed by the SCAD approach which is most computationally expensive. Secondly, by looking at the specificity (i.e. true negative rate TNR, the proportion of claimed negatives that are true negative) and the sensitivity (i.e. true positive rate TPR, the proportion of detected positives that are true positive), we further observe that the proposed estimator behaves like SCAD for larger p. By comparing these two performance measures separately, we argue that it is difficult to arrive at a safe conclusion concerning the existence of a uniformly superior estimator, at least from a numeric perspective. To jointly examine the specificity and the sensitivity, we also calculate the MCC. The MCC is defined as

$$\frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP}) \times (\mathrm{TP}+\mathrm{FN}) \times (\mathrm{TN}+\mathrm{FP}) \times (\mathrm{TN}+\mathrm{FN})}}. \tag{5.43}$$

MCC values are always between -1 and 1; with 1 being perfect prediction and -1 being the worst performance. From Figure 5.2, it is clear that for large-scale problems, the proposed estimator $\hat{\Omega}$ and SCAD outperform CLIME and graphical Lasso. The MCCs of the proposed estimator and SCAD are nearly identical. We note that it has been empirically observed by existing literature that SCAD often tends to produce sparser estimates. Nevertheless, in terms of computational cost, our proposed algorithm is more preferred than SCAD for ultra large-scale problems.

Further, we show that the performance of the proposed estimator $\hat{\Omega}$ can be improved by taking advantage of the Toeplitz structure of the covariance and precision matrices (as additional information) and replacing thresholding by the tapering procedure proposed in [20]. From Table 5.1 and Figure 5.2, we can see that, incorporation of the bandable structure into the tapering version $\hat{\Omega}_{\text{taper}}$ can significantly and uniformly improve the estimation performances within our framework. Again, it is worthy emphasizing that in this simulation setup, $\hat{\Omega}_{\text{taper}}$ can be proved to remain optimal in the minimax risk sense under the spectral norm.

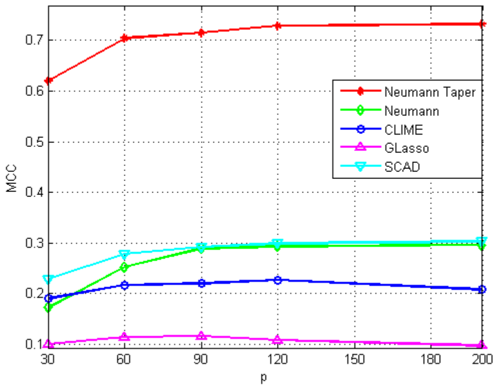
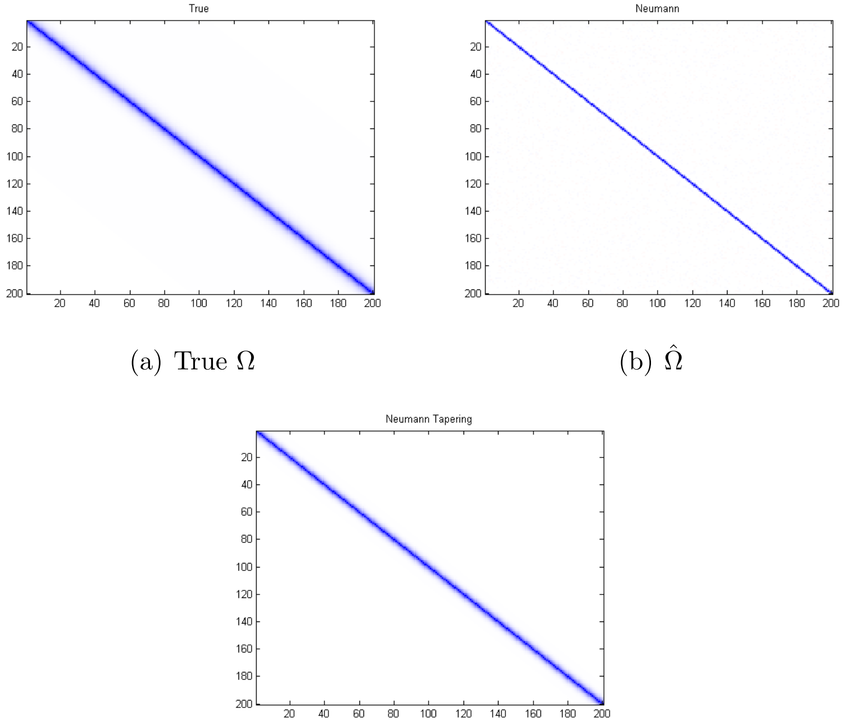


Figure 5.2: The Mathews correlation coefficients (MCC) of estimating Ω when using the proposed algorithm based on the Neumann series representation, its tapering version, CLIME, graphical Lasso (GLasso), and SCAD.



(a) True Ω

(b) $\hat{\Omega}$

(c) $\hat{\Omega}_{\text{taper}}$

Figure 5.3: True sparse precision matrix Ω with $p = 200$. Estimated precision matrix $\hat{\Omega}$ based on the Neumann series and its tapering version $\hat{\Omega}_{\text{taper}}$ with $n = 100$, averaged over 100 replications.

We also plot the entries of Ω in Figure 5.3. A visual inspection shows that the pattern recovered by our proposed algorithm preserves the true structure. On one hand, the high specificity of our proposed algorithm reflects its ability to detect essentially all effective zeros in Ω. On the other hand, the low sensitivity of $\hat{\Omega}$, as well as CLIME, graphical Lasso, and SCAD, comes from the fact that there are many non-zero but small off-diagonal entries in Ω which makes the high accuracy of model selection more difficult to achieve than a truly sparse precision matrix. This observation can be well justified by Theorem 5.3.10, where we require the minimal magnitude of ω_{ij} be greater than a certain threshold to achieve the model selection consistency.

5.6 Application to An Real fMRI data

5.6.1 Modeling F→STN

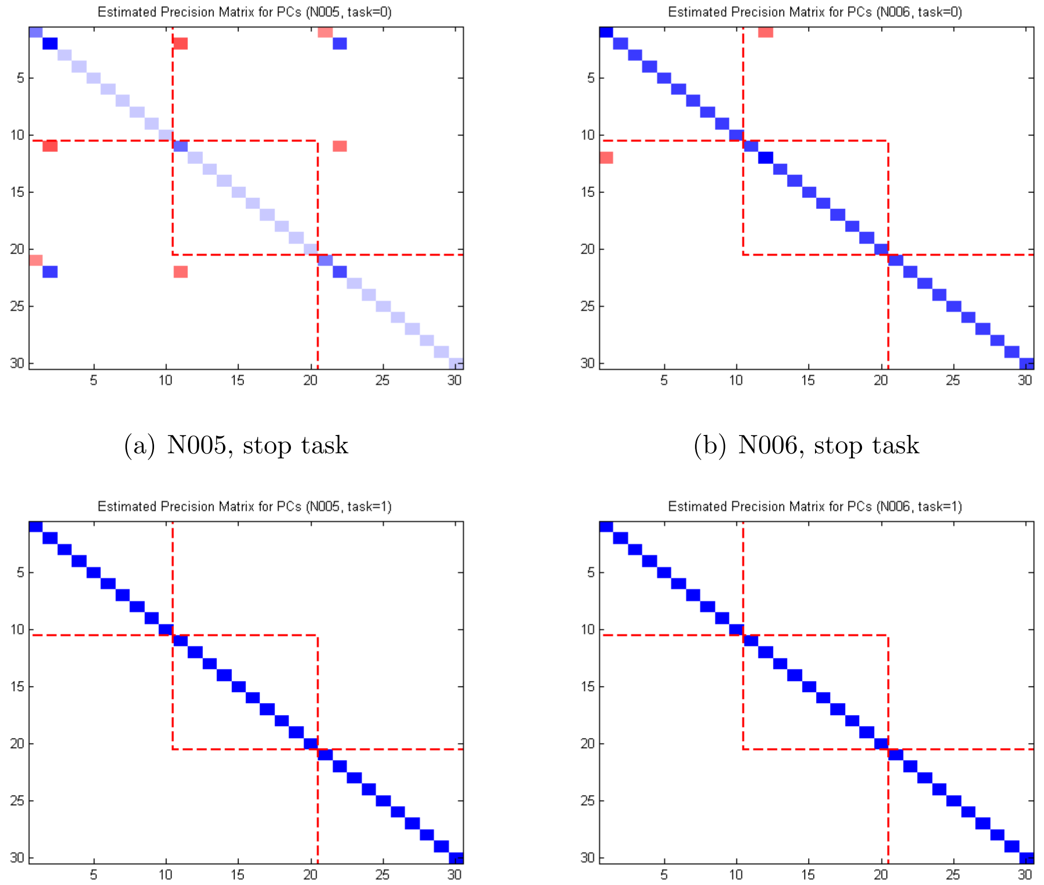
The importance of the "hyperdirect pathway" (from the frontal cortex directly to the subthalamic nucleus (STN), i.e. F→STN) is being increasingly recognized in Parkinson's disease (PD). The hyperdirect pathway has been implicated in impulse control problem. During deep brain stimulation (DBS) surgery, a small electrode is implanted into the STN and an electrical current delivered to disrupt noisy activity in this brain structure. Traditionally, the STN was considered accessible only through DBS. More contemporary work has suggested that there is a direct connection between the frontal cortex in the outer part of the brain and the STN: the hyperdirect pathway. Moreover, despite the small size of the STN, a recent fMRI study in subjects without PD has suggested that it is possible to measure activity in the STN. In the current study, we are interested in test the hypothesis that the strength of connection between frontal cortices and the STN will be significantly different when subjects perform a motor task involving a sudden "stop" command compared to the same motor task where a stop command is not issued. We expect there is some connections in the stop task; but not in the control task. In addition, we shall focus on inferring *direct* connectivity between frontal cortices and the STN, rather than the result of *indirect* influence, e.g. via the thalamus.

Now, we formulate the model. Our goal is to construct connections A → B | C, where A, B, and C are pre-defined brain regions. Here, A = frontal cortex (F), B = STN, and C = thalamus. In words, we would like to learn the brain connectivities *directly* from A to B, by removing the indirect effect of connections of A to B via C; this is exactly tantamount to learn a sparse precision matrix of the three regions. Since the pixels are highly correlated within neighborhoods and their sizes are very large, we first apply PCA to reduce the dimensionality. More specifically, we look at the *eigen-pixels* and learn sparse connections between those eigen-pixels. In particular, in our experiments, we used 10 PCs for each region; then we combine the 10 PCs from the three regions (A, B, C) and run our proposed model to learn a sparse 30×30 precision matrix Ω. As we have seen, a nonzero entry in precision matrix implies an edge in its Gaussian graphical model representation, which in turn implies a connection between the corresponding two eigen-pixels. Therefore, we are interested in the non-zero entries of Ω between rows 1-10 and columns 11-20. We run the proposed method based on the Neumann series representation on this data set. 2-fold CV is used to determine the optimal threshold and number of terms in the Neumann series.

5.6.2 Learned F→STN Connectivities

We first plot the estimated precision matrix Ω for two normal subjects N005 and N006, see Figure 5.4. There are one connections identified for each normal subject in the stop task: PC2 of A↔PC1 of B in N005 and PC1 of A↔PC2 of B in N006. In contrast, there is no connectivity detected by our model in the control task for both subjects. This is in accordance with the biological knowledge that expects connections in the stop task but not in the control task for normal individuals.

Second, the patterns of identified PCs that are connected in the stop tasks for



Estimated Precision Matrix for PCs (N005, task=0)

Estimated Precision Matrix for PCs (N006, task=0)

Estimated Precision Matrix for PCs (N005, task=1)

Estimated Precision Matrix for PCs (N006, task=1)

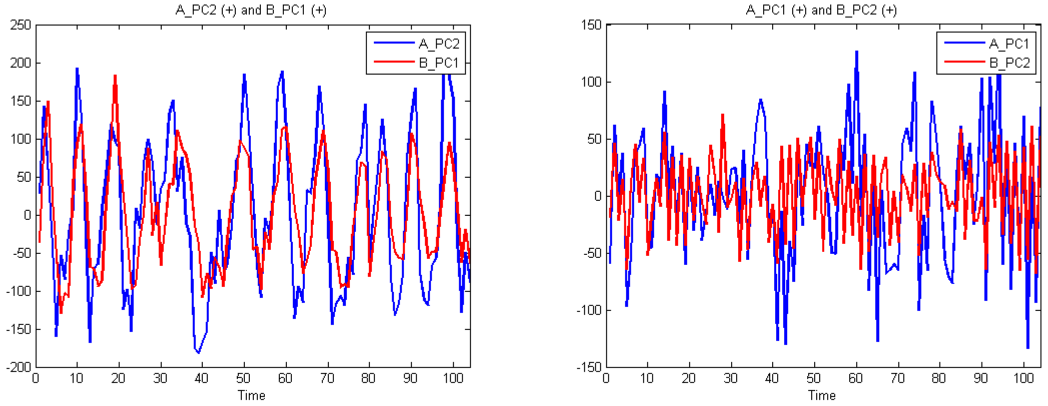
(a) N005, stop task

(b) N006, stop task

(c) N005, control task

(d) N006, control task

Figure 5.4: Estimated precision matrix Ω for two normal subjects N005 and N006. Each subject performs two task: stop and control. From upper left block to bottom right block: A, B, C. We are interested in non-zero entries in the upper middle block: rows 1-10 and columns 11-20.



(a) N005, stop task

(b) N006, stop task

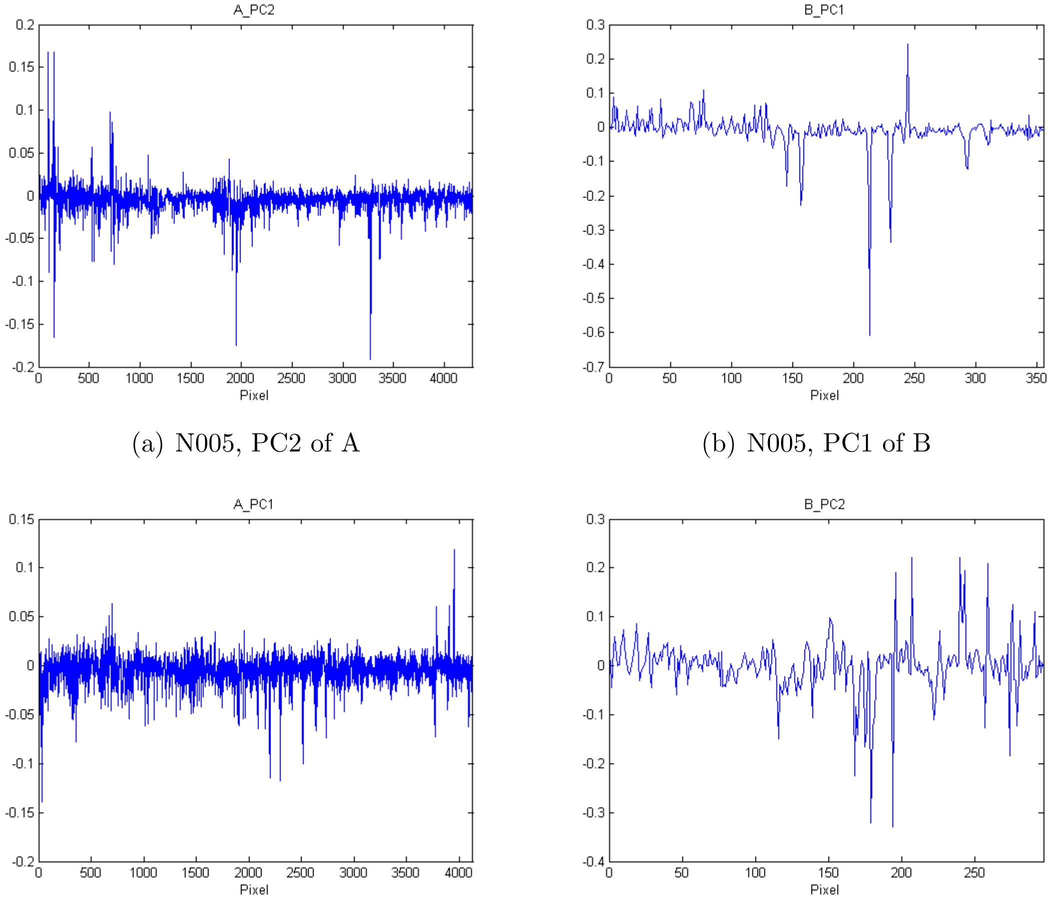
Figure 5.5: The patterns of identified PCs that are connected in the stop task for two normal subjects N005 and N006.

N005 and N006 are also shown in Figure 5.5. It is clear that the associated PCs between A and B are highly correlated with each other.

Finally, we plot the loadings of the identified PCs in Figure 5.6. We can see that there is a clear clustering property for the original pixels associated with the identified PCs, implying that there are a few clusters of spatially close pixels connecting together.

5.7 Conclusion and Discussion

We presented a conceptually simple and computationally efficient algorithm on estimation of large sparse precision matrices. The proposed estimator is motivated by identifying a class of sparse matrices that is approximately inversely closed. Our theoretic analysis showed that the proposed estimator for this class is statistically valid and optimal in the sense of minimax risk under the spectral norm. We further showed that the proposed estimator is model selection consistent which is a direct consequence of the established convergence result on the entry ∞-norm. Then, simulation results demonstrated the encouraging performances of the proposed estimator when compared with state-of-art optimization based methods. Finally, the proposed method was applied to learn direct brain connectivity based on fMRI data and yielded



(a) N005, PC2 of A

(b) N005, PC1 of B

(c) N006, PC1 of A

(d) N006, PC2 of B

Figure 5.6: Loadings of the identifed PCs for two normal subjects N005 and N006.

biologically plausible results.

We would like to point out a few directions for future work. An interesting study following the current work can be on better determining the parameters using a data-driven approach. We will need to analyze its theoretic performance in comparison with an oracle that allows us to know the true covariance/precision matrix in advance of observing any data. We expect that the work in [11] could shed us some light on this future direction. Another future direction is to extend the current work to the complex case. So far we have developed our framework for real data, however, we note that there is nothing preventing extending the obtained results to the complex case, as long as the concentration of measure inequality (5.22) continues to be valid. This requirement is generally true (see the theoretic ground laid in [83]). Extension to the complex case will allow us to apply the proposed framework to a broad range of statistical signal processing problems [1, 87, 114]. Finally, it is also interesting to consider estimating time-varying sparse $\Omega(t)$ in the proposed framework, e.g. for modeling fMRI brain networks.

Chapter 6

fMRI Group Analysis of Brain Connectivity

6.1 Introduction

Studying brain connectivity is crucial in understanding brain functioning and can provide significant insight into the pathophysiology of a number of neurological disorders. Increasingly, inferring brain connectivity using functional Magnetic Resonance Imaging (fMRI) is being explored, and many mathematical formalisms, such as structural equation modeling (SEM) [93], multivariate autoregressive models (mAR) [65], dynamic causal modeling (DCM) and dynamic Bayesian networks (DBNs) [88] have been proposed. Despite significant progress during the last decade, there are still a number of challenges associated with inferring brain connectivity from fMRI. One is the curse of complexity with the above SEM and/or mAR approaches when dealing with practical fMRI data sets where the number of brain regions-of-interest (ROIs) is relatively large and the number of time points is limited. Based on a number of neuroscience studies, the connections between brain regions generally can be considered *a priori* to form a sparse network, suggesting that sparsity should be incorporated into brain connectivity modeling. For instance, sparse mAR models were studied in [118] where the parameters are estimated using penalized regression. Group analysis of effective brain connectivity has long been another challenging topic, since biomedical research is usually conducted at a group level to extract the population features. Efficient group analysis requires appropriate handling of expected inter-subject variability without destroying inter-group differences. To address the above two crucial challenges, this chapter aims at developing a novel, computationally-efficient brain connectivity model that incorporates both sparsity and suitable group analysis.

Several methods for meaningfully extracting group information from fMRI data have been proposed. The *common structure* (CS) model in the DBNs context [88] enforces the same graphical structure for all subjects within a given group, but the connection coefficients are allowed to vary on a subject-by-subject basis. However, CS inference based on DBNs inevitably requires computationally-intensive algorithms such as the Markov Chain Monte Carlo (MCMC). Another proposed method is that of Bayesian group analysis [113] where several possible model structures are considered and the posterior evidence of models for each subject is estimated. Here we present a different group linear regression model to perform group analysis while incorporating the sparsity principle. More specifically, we adopt the modeling concept on the CS level – whereby brain connections generating the fMRI observations are assumed to be structurally identical among subjects within the same group, but individual connection parameters are allowed to vary between subjects – and propose a group robust Lasso framework to perform group analysis. There are several advantages associated with the proposed novel framework:

1. The proposed model is based on the optimization of a convex objective function and thus is computationally more efficient than graphical modeling approaches such as DBNs.

2. The proposed model represents a unified framework whereby group analysis is based on networks learned directly from the time courses in fMRI data.

3. The proposed model is robust against large variance noise and outliers.

Remark 14. We note that the proposed group robust Lasso approach to infer brain connectivity networks is not based on inverse covariance matrix. Nonetheless, this marginal neighborhood selection procedure has been shown to be a consistent variable selection method for constructing a Gaussian graphical model under certain regularity conditions [95].

6.2 Methods

6.2.1 A Group Robust Lasso Model

We propose inferring brain connectivity through a linear regression approach. The Blood Oxygen Level Contrast (BOLD) signal intensity at a target ROI is regarded as a response which is modeled by a linear combination of time courses from ROIs subjected to the corruption by certain noise:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}. \tag{6.1}$$

Here $\mathbf{y}$ is a response vector, $\mathbf{X}$ is a design matrix with columns representing predictors, $\boldsymbol{\beta}$ is a coefficient vector, and $\mathbf{e}$ is a zero-mean random error vector which is assumed to have iid elements with a common finite variance, σ^2. We consider the situation where the number of potential predictors is large while the number of *bona fide* predictors with non-zero coefficients are only a small fraction of the total. Thus, the goal is to determine the correct underlying sub-model. The Lasso [115] is a popular linear model selection tool that continuously shrinks coefficients to zeros. The Lasso is a regularized linear model and minimization of its ℓ_1 penalized squared ℓ_2 loss is known to promote sparsity on the coefficient vector. Nevertheless, the Lasso solution can be unsatisfactory for group analysis because its selection of a predictor is relatively independent of each other and therefore the Lasso estimator in unable to incorporate the potential similarity of structures across subjects. Yet for group analysis, we have certain structural grouping information that is available to us as *a priori*, e.g. the subjects within the same group are assumed to share the same connectivity structure. Therefore, $\boldsymbol{\beta}$ is composed of G groups each of which contains p_g individual coefficients for $g \in \{1, \cdots, G\}$. In matrix notation, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^*, \cdots, \boldsymbol{\beta}_G{}^*)^*$, and $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_G)$ is a block design matrix of dimension $n \times \sum_g p_g$. With this notation, we can refer to *grouping selection* to mean that the sparsity is promoted at the group level, i.e. corresponding subject-specific coefficients within one group are either all non-zeros or all zeros. Thus the Lasso is a special case of the group version when $p_g = 1$ for all g.

To promote sparsity at the group level, we choose to minimize the following objective function:

$$f(\boldsymbol{\beta}) = L\left(\boldsymbol{\beta}; \mathbf{y}, \mathbf{X}\right) + \lambda_n \sum_{g=1}^{G} \left\|\boldsymbol{\beta}_g\right\|_{\ell_2} \tag{6.2}$$

where $L(\cdot)$ can be any cost function. Unlike the ℓ_1 penalty in the Lasso, summation of block Euclidean norms (a.k.a. blocked ℓ_1 norm) in the penalty term encourages grouping selection [124]. The group Lasso is a the special case where L is the standard squared ℓ_2 loss [128]. More generally, we can adopt robust losses that are less sensitive to noise that includes large variability or even outliers. For instance, the convex combination of ℓ_1 and squared ℓ_2 losses [31] or the Huber loss [34] coupled with the block ℓ_1 regularization yields a group robust Lasso. In this paper, we propose a group robust Lasso by using the convex combined loss with a robustness tuning parameter $\delta \in [0, 1]$

$$L\left(\boldsymbol{\beta}; \mathbf{y}, \mathbf{X}\right) = (1 - \delta) \left\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right\|_{\ell_1} + \delta \left\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right\|_{\ell_2}^2 . \tag{6.3}$$

The group Lasso is thus a reduced case when $\delta = 1$. In general, a smaller δ gives more robustness. In the case of robust Lasso ($p_g = 1, \forall g$), the asymptotic behavior of its estimator has been studied in [31] where it is shown therein that the asymptotic variance is stabilized. Furthermore, the robustness tuning parameter can be chosen by the minimal asymptotic variance criterion when the error distribution is known. The proposed group robust Lasso estimator is defined to be any minimizer of (6.2), i.e. $\arg\min_{\boldsymbol{\beta}} \{f(\boldsymbol{\beta})\}$. Note that there is a corresponding model for each λ_n, so determining a proper shrinkage amount is important to make the subsequential inference. The optimal shrinkage parameter λ_n is determined by the BIC which is computed as follows: we first solve for the group robust Lasso for a fixed λ_n. Once the model is determined, we fit the corresponding subset of data to the selected model by unregularized least squares. We then obtain an estimator of $\boldsymbol{\beta}$ with the shrinkage effect removed. An estimator of σ^2 is given by the maximum likelihood

$$\hat{\sigma^2} = \frac{L\left(\hat{\boldsymbol{\beta}}; \mathbf{y}, \mathbf{X}\right)}{(2-\delta)n}.$$

The estimator $(\hat{\boldsymbol{\beta}}, \hat{\sigma^2})$ is called the *Gauss group robust Lasso estimator* which corrects for the bias of underestimating non-zero coefficients and thus is more suitable for accurate estimation. Note that the likelihood under which $\hat{\boldsymbol{\beta}}$ is computed is assumed to be Gaussian (equivalent to the least squares estimator) while the likelihood to estimate the variance, σ^2, is a blend of Gaussian$(0, \sigma^2)$ and Laplace$(2\sigma^2)$ distributions.

Now the BIC can be calculated from the Gauss group robust Lasso estimate

$$\begin{aligned} \text{BIC} &= -2 \times \text{log-likelihood}(\hat{\boldsymbol{\beta}}, \hat{\sigma^2}) + k\log(n) \\ &= \frac{1-\delta}{\hat{\sigma^2}}\left\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right\|_{\ell_1} + \frac{\delta}{\hat{\sigma^2}}\left\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right\|_{\ell_2}^2 \\ &+ 2(1-\delta)n\log(4\hat{\sigma^2}) + \delta n\log(\hat{\sigma^2}) + k\log(n) \end{aligned} \tag{6.4}$$

where k is the number of predictors in the selected model. Finally, the optimal model is chosen by the minimal BIC value among the set of different shrinkages.

To summarize, proposed procedure contains the following steps:

1. Choose a set of shrinkage parameters λ_n. Run the group robust Lasso for each shrinkage.

2. For each shrinkage, identify the non-zero coefficients and use this submodel to compute $(\hat{\boldsymbol{\beta}}, \hat{\sigma^2})$.

3. Compute BICs using estimates from step 2 and choose the model corresponding to the minimal BIC value as the optimal model.

We use the group robust Lasso as a model selection tool and estimate parameters based on the selected model. We refer this (variant) Gauss group robust Lasso as the *grpRLasso* in this paper unless otherwise indicated.

6.2.2 A Group Sparse SEM+mAR(1) Model

The brain connectivity model we assume has a *unified* SEM framework that captures both spatial and temporal brain connections, where we combine the standard SEM model [93] (to represent the relations considered instanteous at the temporal resolution of fMRI) and the 1st-order mAR model [65] (to represent longitudinal temporal relations). Suppose there are p ROIs and the brain is MRI scanned at time $1, \cdots, T$. We also assume that there are S subjects belonging to G groups. Denote by $\mathbf{y}_{s,j}$ the fMRI measurement vector of the j^{th} ROI of subject s, for $j \in \{1, \cdots, p\}$ and $s \in \{1, \cdots, S\}$, as the response variable.

Before introducing the group SEM+mAR(1) model, we introduce a few useful notations first. For the s^{th} subject, let $Y_s^{(0)}$ be the $(T-1) \times p$ matrix with the j^{th} column containing the fMRI measurements of the j^{th} ROI from time 2 to T, and let the $(T-1) \times p$ matrix $Y_s^{(1)}$ be the time-shifted version of $Y_s^{(0)}$ with lag 1. $Y_{s,-j}^{(0)}$ denotes the $(T-1) \times (p-1)$ matrix with the j^{th} column $\mathbf{y}_{s,j}$ being removed. With these notations, for each subject $s \in \{1, \cdots, S\}$, we have the following SEM+mAR(1) model:

$$\mathbf{y}_{s,j} = \underbrace{Y_{s,-j}^{(0)} \boldsymbol{\beta}_{s,j}^{(0)}}_{\text{SEM}} + \underbrace{Y_s^{(1)} \boldsymbol{\beta}_{s,j}^{(1)}}_{\text{mAR(1)}} + \mathbf{e}_s \tag{6.5}$$

where $\mathbf{e}_s$ means the error vector for each subject s, and $\boldsymbol{\beta}_{s,j}^{(0)}$ and $\boldsymbol{\beta}_{s,j}^{(1)}$ represent respectively the SEM and mAR connection strength coefficients to be estimated.

Putting all subjects together and rewriting in matrix form, we can reach the ambient linear regression model (6.6) with a block diagonal design matrix $\mathbf{X}$ and a coefficient vector $\boldsymbol{\beta}$ with a group structure.

$$\underbrace{\begin{pmatrix} \mathbf{y}_{1,j} \\ \mathbf{y}_{2,j} \\ \vdots \\ \mathbf{y}_{S,j} \end{pmatrix} = \operatorname{diag}\Big\{(Y^{(0)}_{s,-j}, Y^{(1)}_{s})\Big\}_{s=1}^{S} \begin{pmatrix} \boldsymbol{\beta}^{(0)}_{1,j} \\ \boldsymbol{\beta}^{(1)}_{1,j} \\ \boldsymbol{\beta}^{(0)}_{2,j} \\ \boldsymbol{\beta}^{(1)}_{2,j} \\ \vdots \\ \boldsymbol{\beta}^{(0)}_{S,j} \\ \boldsymbol{\beta}^{(1)}_{S,j} \end{pmatrix} + \begin{pmatrix} \mathbf{e}_{1} \\ \mathbf{e}_{2} \\ \vdots \\ \mathbf{e}_{S} \end{pmatrix}}_{\overset{\text{def}}{=}\mathbf{X}\boldsymbol{\beta}+\mathbf{e}} \tag{6.6}$$

where

$$\operatorname{diag}\Big\{(Y^{(0)}_{s,-j}, Y^{(1)}_{s})\Big\}_{s=1}^{S} =$$

$$\begin{pmatrix} Y^{(0)}_{1,-j} & Y^{(1)}_{1} & & & & & \\ & & Y^{(0)}_{2,-j} & Y^{(1)}_{2} & & & \\ & & & & \ddots & & \\ & & & & & Y^{(0)}_{S,-j} & Y^{(1)}_{S} \end{pmatrix}.$$

Now for each target ROI j, the proposed grpRLasso can be applied to (6.6) to learn a sparse coefficient vector with grouping structures. Brain connectivity networks are constructed by enumerating all the ROIs and our network analysis is based on the learned grpRLasso coefficient matrices.

We give the asymptotic behavior of the proposed group robust lasso estimator. The obtained theoretic result justifies its usage. Essentially, we shall prove that the proposed group robust lasso can select the correct underlying model with probability approaching to one when a large sample size is available. In other words, the group robust lasso has the oracle property for model selection, namely without knowing the support of the true coefficient vector, the correct model can be identified with probability arbitrarily close to one. Compared with the group lasso, we shall see that the group robust lasso is robust against errors with large variability.

Lasso	RLasso	grpLasso	grpRLasso	Oracle
2.99 (1.59)	2.89 (1.63)	9.62 (5.33)	1.66 (0.52)	1.66 (0.52)

Table 6.1: MSEs for the estimated coefficients with their standard deviations shown in brackets. grpLasso abbreviates for the group Lasso,RLasso for the robust Lasso with the convex combined loss, and grpRLasso for the group robust Lasso. Oracle is the maximum likelihood estimate (for the Gaussian likelihood) obtained from the model with knowing the true non-zero locations.

Theorem 6.2.1. *Suppose that $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n n^{\frac{\gamma-1}{2}} \to \infty$. Assume further that*

1. *$n^{-1}\mathbf{X}^T\mathbf{X} \to \mathbf{C}$, where $\mathbf{C}$ is a positive definite matrix;*
2. *$\{e_i\}$ have a common continuous probability density function in a neighborhood of 0.*

Then the group robust Lasso estimator has the asymptotic normality on non-zero components:

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{nA} - \boldsymbol{\beta}_A\right) \Rightarrow N\left(\mathbf{0}, \frac{(1-\delta)^2 + 4\delta^2\sigma^2 + 4\delta(1-\delta)M_{10}}{4[\delta + (1-\delta)f(0)]^2} C_{11}^{-1}\right). \tag{6.7}$$

Remark 15. Under further adaptive regularization weight on group coefficients with rate $\lambda_g = \left\|\hat{\boldsymbol{\beta}}_{ng}^{LS}\right\|_2^{-\gamma}$, we can show the sign consistency

$$P\left(\operatorname{sgn}\left(\hat{\boldsymbol{\beta}}_{nA}\right) = \operatorname{sgn}\left(\boldsymbol{\beta}_A\right)\right) \to 1, \tag{6.8}$$

as $n \to \infty$.

6.3 A Simulation Example

A synthetic example is used to demonstrate empirical evidence for the improved detection and estimation performance of the proposed grpRLasso over the grpLasso, Lasso, and robust Lasso. A design matrix containing 400 observations and 200 predictors is realized from a Gaussian Ensemble in each synthetic data set. The true sparse coefficient vector $\boldsymbol{\beta}^0$ is set to have a block structure. That is, $\boldsymbol{\beta}^0$ contains three blocks of non-zero coefficients of different magnitudes (1.5, -3, and 2, respectively) distributed in intervals located in $[20, 30]$, $[90, 95]$, and $[180, 188]$, respectively. The shrinkage amount is set on an evenly spaced grid over $[0, 2000]$. BIC (6.4) is used to choose the optimal model. The error $\mathbf{e}$ is simulated from a Student-t distribution with parameters $\nu = 3$ and $\sigma^2 = 9$ so that it has a variance 27.

The mean squared errors (MSEs) computed from 100 simulations are shown in Table 6.1. We note that the grpRLasso achieves the same performance as when the non-zero locations are known in advance (the "oracle property"). Hence, this implies that the proposed grpRLasso is very accurate and robust in terms of both model selection and parameter estimation. In contrast, the grpLasso model has a very large MSE, even worse than the Lasso. This is not surprising after a careful investigation on the nature of a grouping variable selection tool. Suppose that if the group Lasso falsely identifies a non-zero coefficient, then the rest elements in the group are all non-zeros which render a high estimation error. In contrast, since selection procedure of the Lasso is independent among predictors, incorrect selection of one variable has little influence on others.

6.4 fMRI Group Analysis in Parkinson's Disease

In this section, we apply the grpRLasso to a fMRI data collected from subjects with and without Parkinson's disease (PD) and report the group analysis results of the learned brain connectivity networks.

6.4.1 Data Description

fMRI scans for ten normal people and eight subjects with PD were collected in the study. Subjects were asked to continually squeeze a bulb in their right hand to control an inflatable ring so that the ring moved through an undulating tunnel without

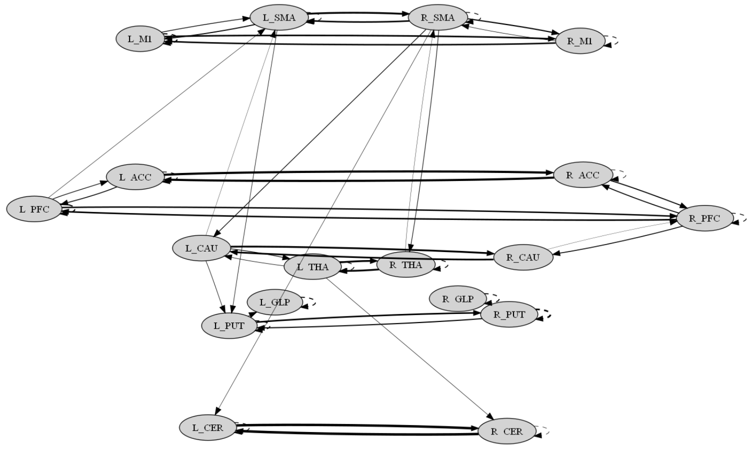


Figure 6.1: Weighted network for the normal group: solid lines are the SEM connections and dashed lines the mAR(1) connections. Node labels are the names of ROIs with the prefixal L indicating the left brain hemisphere and R the right hemisphere. Line width is proportional to the connection strength. The thicker the edge, the larger the magnitude of the estimated coefficient.

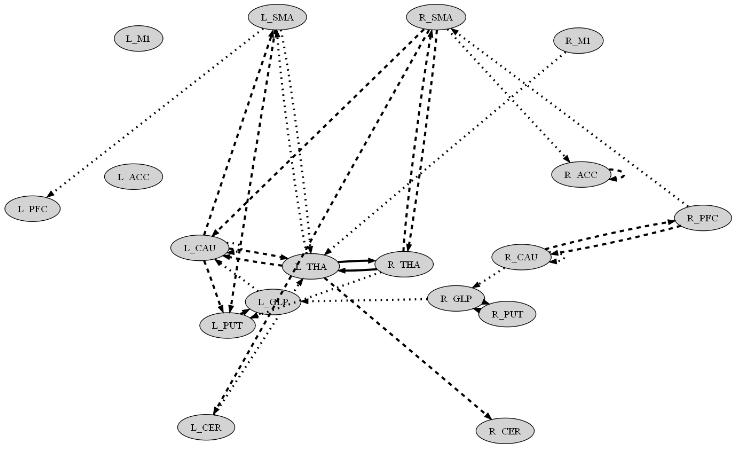


Figure 6.2: Different connections between normal and "off-medication" networks: dashed edges are only present in the normal network while dotted edges only in the "off-medication" network. Solid edges exist in both networks with significantly different means (t-test with size 0.05).

touching the sides. PD subjects performed the same task after been withdrawn from their L-dopa medication for 12hrs. Images were acquired at a sampling rate of 0.5Hz and a trial lasted for five minutes so that 150 data points were obtained for each subject.

6.4.2 Learned Brain Connections

The robustness tuning parameter δ is fixed to 0.5. It is worth noting that this parameter can be further optimized if required [31]. For each target ROI, there are 35 possible directed edges pointing to it, 17 from SEM and 18 from mAR(1). Since we have the normal (10 subjects) and PD off-medication (8 subjects), a linear regression model for each target node has $2 \times 35 = 70$ groups partitioning the total 630 coefficients.

Generally speaking, the (robust) Lasso networks yielded many more connections that are inconsistent among subjects within a group. Hence they are less useful for group studies and not reported further. The network learned from grpRLasso for the normal group is shown in Figure 6.1 and the difference between the normal and PD group networks is shown in Figure 6.2. Edges shown in Figure1 have significant non-zero coefficients with a t-test (against zero mean) of size 0.05. There are four main findings of biological significance. First, as seen in Figure 6.1, there were many reciprocal connections between homologous regions in both groups (e.g. left supplementary motor area (L_SMA)↔right supplementary motor area (R_SMA)). Second, there appeared to be a left↔right shift in the regions active when comparing normal subjects to PD subjects, despite the fact that all subjects were using their right hand during the motor task. For example, while normal subjects recruited there R_SMA and right thalamus (R_THA), PD subjects recruited their L_SMA and L_THA. Presumably the connections between regions homologus (Figure 6.1) provide a mechanism through which PD subjects can recruit regions on the opposite side of the brain. Third, there were connections between the right prefrontal cortex

(R_PFC) and right caudate (R_CAU) in normal subjects that were missing in PD subjects. This likely reflects alterations in the secondary dopaminergic pathway to medial prefrontal regions known to be affected in PD [37]. Fourth, there was enhanced connectivity withing basal ganglia regions (e.g. right putamen (R_PUT)↔right globus pallidus (R_GLP), R_THA→L_PUT, right caudate (R_CAU)→R_GLP). This may reflect that these regions become entrained in oscillations which might enhance the functional connectivity observed with fMRI [49].

Chapter 7

Conclusions and Discussions

7.1 Contribution Summary

In this thesis, we have studied several issues on estimating high-dimensional sparse models from both theoretic and algorithmic perspectives. Now, we summarize the main contributions of this thesis.

In Chapter 2,

- We proposed a convex combined loss of ℓ_1 (LAD) and ℓ_2 (LS), rather than the pure LS cost function, coupled with the ℓ_1 penalty to produce a robust version of the Lasso. Asymptotic normality was established, and we showed that the variance of the asymptotic normal distribution is stabilized. Estimation consistency was proved at different shrinkage rates for $\{\lambda_n\}$ and further proved by a non-asymptotic analysis for the noiseless case.

- Under a simple adaptation procedure, we showed that the proposed robust Lasso is model selection consistent, i.e. the probability of the selected model to be the true model approaches to 1.

- As an extension of the asymptotic analysis of the proposed robust Lasso, we studied an alternative robust version of the Lasso with the Huber loss function, the Huberized Lasso. For the Huberized Lasso, asymptotic normality and model selection consistency were established under much weaker conditions on the error distribution, i.e. no finite moment assumption is required for preserving similar asymptotic results as in the convex combined case. The Huberized

Lasso estimator is well-behaved in the limiting situation when the error follows a Cauchy distribution, which has infinite first and second moments.

- The analysis result obtained for the non-stochastic design was extended to the random design case with additional mild regularity assumptions. These assumptions are typically satisfied for auto-regressive models.

In Chapter 3,

- We proposed a hierarchical, fully Bayesian version of the Lasso model for inferring sparse linear regression from high-dimensional data sets.
- We developed a reversible-jump MCMC algorithm to compute the unbiased minimum variance estimates and proved its convergence.

In Chapter 4,

- For high-dimensional covariance estimation problems where $p/n \to \infty$, we showed that the MMSE shrinkage oracle estimator is inconsistent under both Frobenius and spectral risks for some typical covariance matrices. Moreover, we showed that the tapering estimator is uniformly superior than the MMSE shrinkage estimator in this case.
- We proposed a STO estimator that combines the advantages from both the MMSE shrinkage and tapering estimators. In particular, the proposed estimator is suitable for estimating general, high-dimensional covariance matrices. An oracle estimator in the closed-form was derived and a practical algorithm to approximate the STO estimator was presented.

In Chapter 5,

- We identified a class of sparse matrices that is approximately inversely closed and proposed a conceptually simple and computationally efficient algorithm on estimation of large sparse precision matrices in this class.

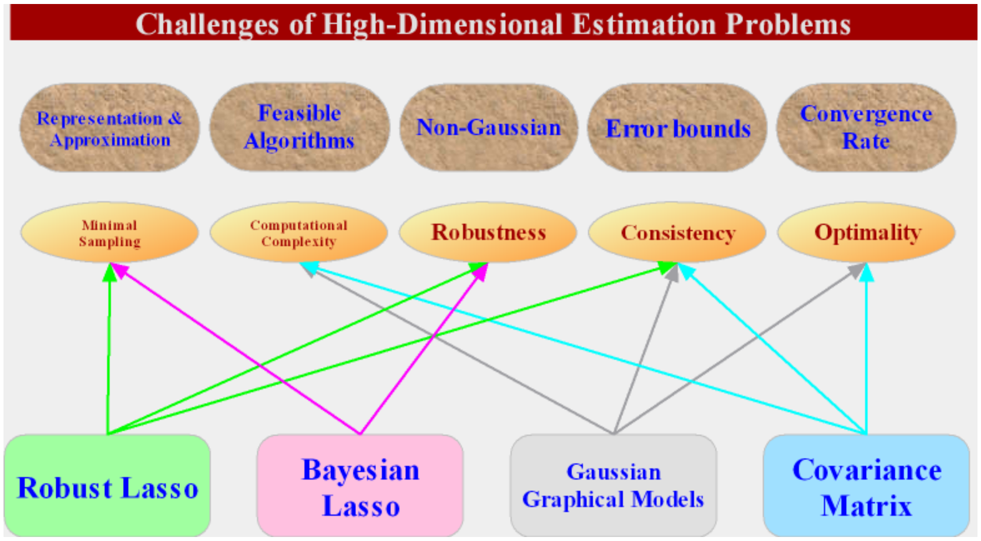


Figure 7.1: This overview summarizes the challenges raised in Chapter 1, the methods proposed in this thesis, and the relationship between proposed methods and the challenges being addressed.

- Our asymptotic analysis showed that the proposed estimator for this class is statistically valid and optimal in the sense of minimax risk under the spectral norm.

- We also established convergence result on the entry ∞-norm and showed that the proposed estimator is model selection consistent.

In Chapter 6,

- We presented a group robust Lasso (grpRLasso) framework that combines SEM and mAR(1) for inferring group-level, sparse brain connectivity networks.

- The grpRLasso was applied to fMRI obtained from subjects with and without PD and significant group differences in biologically plausible regions were found. We suggest that the proposed method provides a computationally efficient means to infer group brain connectivity from fMRI data.

The challenges **C1-C5** raised in Chapter 1 have been individually addressed by the four topics studied in this thesis. The overview picture is shown in Figure 7.1.

7.2 Directions for Future Research

We would like to point out a few directions for the future research based on the accomplished work in this thesis.

7.2.1 Estimation of Conditional Independence for High-Dimensional Non-Stationary Time Series

The independence assumption is of critical importance in all the approaches where the goal is to estimate a (series of) static and sparse precision matrix Ω's. Unfortunately, this assumption is overly restrictive to a broad range of real-world datasets since the data generation mechanisms are hardly static. Equivalently speaking, the underlying Ω for generating $\{\mathbf{x}_i\}$ can change over time and its structures and parameters at a particular time may also depend on the previous ones. For instances, functional brain connectivities are likely to alter when different tasks are performed; genetic regulatory networks are prone to evolving in order to adapt to changing environments. To capture the time-dependent features, the iid hypothesis must be relaxed to accommodate the reality. Nevertheless, there have only been a few research [78, 131] devoted to the estimation of high-dimensional precision matrices of multivariate time-series data.

In the proposed future research, we shall focus on the estimation of high-dimensional sparse precision matrices $\Omega(t)$ for time-series data. We model the time-varying $\Omega(t)$ under the general *physical dependence measure* framework proposed by [125]:

$$\mathbf{x}_i = G(\mathbf{e}_i, \mathbf{e}_{i-1}, \cdots ; i/n), \tag{7.1}$$

where G is a measurable function driving the physical data generation processes and $\{\mathbf{e}_i\}$ are *independent* innovations at $t = i/n$. There are two major advantages over the current state-of-the-art approaches based on the independence hypothesis. First, (7.1) can model non-linear, non-stationary multivariate time-series $\{\mathbf{x}_i\}$. In particular, mild conditions on G allow us to model *locally stationary processes*, a very flexible and general stochastic framework that covers a wide range of many existing time-series models including linear processes (time-varying auto-regression and moving average processes), Volterra series, non-linear transforms of linear processes, and non-linear time series. Second, combining (7.1) with sparsity in $\Omega(t)$ unifies two forms of time-dependence in a single model: changing graph of $\Omega(t)$ and dependence through auto-correlation. Both are new to the current literature.

Compared with the time-varying Gaussian graphical models in [78, 131], the approach (7.1) significantly generalizes their work in several aspects. First, in [131], time-varying $\Omega(t)$ is modeled by *independent* structural changes, rather than the stochastic paradigm we consider here. In fact, their assumption can be seen as a special case in our paradigm in the sense that $\mathbf{x}_i = \Sigma^{1/2}(i/n)\mathbf{e}_i$. Second, our approach has clear regression interpretation on $\Omega(t)$, whereas in [131], estimation of $\Omega(t)$ is done by the 1-norm penalized Gaussian likelihood parameterized by kernel smoothed sample covariance matrix. Third, the smoothness conditions used in [131] are based on the maximal fluctuation of the first and second order derivatives on the entries of $\Sigma(t)$ and $\Omega(t)$. As pointed out in [125], however, using derivatives is not a good way of dealing with time dependence because they may not even exist if G is not well-defined. In contrast, we leverage the *physical dependence measure* based on the coupling idea [125].

Based on the model (7.1), my research plan contains the following stages:

1. We consider the non-parametric estimation problem of the function $G(\cdot)$ (and hence $\Omega(t)$) changing *smoothly* over time. In linear processes, this is tantamount to estimate the time-varying coefficients. We shall derive the asymptotics, including consistency, limiting distributions, and optimality, of the proposed estimator. We will also consider the adaptive procedures on bandwidth selection.

2. We consider the case where there exist abrupt change-points (at unknown positions) in $G(\cdot)$; this corresponds to the structural changes, e.g. adding or deleting

edges, in the graphical model.

3. We shall apply the theoretic results to estimate functional brain connectivity networks from functional Magnetic Resonance Imaging (fMRI) data, a natural and practically important application of the time-dependent model (7.1). The proposed method is novel in the sense that it extends the current linear structural equation modeling (SEM) and multivariate auto-regression (mAR) models to locally stationary, time-varying models.

4. We shall also consider the application of the proposed method to model genetic regulatory networks using gene expression data. It is known that cell cycle is a dynamic system and gene expression levels are adaptive to external varying conditions.

7.2.2 Estimation of Eigen-Structures of High-Dimensional Covariance Matrices

So far, we have seen that the estimation performances for large covariance and precision matrices are measured by the accuracy of eigen-values. A more challenging problem (and largely remains unsolved) is to estimate the eigen-vector of such high-dimensional matrices. Precise estimation of both eigen-values and eign-vectors, for which we call the eigen-structures, is of tremendous importance in many areas of statistics (e.g. PCA and linear discriminant analysis), machine learning (e.g. face recognition and classification), signal processing (e.g. beamforming), and computational biology (e.g. microarray clustering and genome-wide association study). Recently, it has been shown that eigen-vectors of the sample covariance matrix is gradually orthogonal to those of Σ in high-dimensional setups and therefore they essentially contain little information about the eigen-structure of Σ [73]. Without special structures in Σ, current approaches and theory can deal with the problem size comparable with the sample size, see [103] for the spiked covariance model as an example; nonetheless, I am very interested in estimating the eigen-structure in regularized subclasses of Σ in the situations where $p \gg n$; for instance, p grows at sub-exponential rate of n. Tools from random matrix theory and optimization are extremely useful for this purpose.

7.2.3 Multi-Task Lasso for Group Analysis

The group analysis presented in [35] requires *strictly grouping* structures. This homogeneity assumption can be overly restrictively in practice since for example there may be sub-types of Parkinson's disease within the patient group. Therefore, it is desirable to allow inter-subject variability within groups. To this end, multi-task Lasso models with structured sparsity can be adopted [29, 76, 84]. Moreover, by considering multi-task Lasso with structured sparsity, we can easily integrate prior domain knowledge from neurology experts and thus improve the performance of group analysis.

Bibliography

[1] Richard Abrahamsson, Yngve Selen, and Petre Stoica. Enhanced Covariance Matrix Estimators in Adaptive Beamforming. *2007 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 969–972, 2007.

[2] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[3] T.W. Anderson. *An Introduction to Multivariate Statistical Analysis.* Wiley, Hoboken, NJ, third edition, 2003.

[4] Christophe. Andrieu and Arnaud. Doucet. Joint bayesian model selection and estimation of noisy sinusoids via reversible jump mcmc. *IEEE Transactions on Signal Processing*, 47(10):2667–2676, 1999.

[5] Z.D. Bai and Y.Q. Yin. Limit of the smallest eigenvalue of a large-dimensional sample covariance matrix. *The Annals of Probability*, 21:1275–1294, 1993.

[6] Zhidong Bai and Jack W. Silverstein. No Eigenvalues Outside the Support of the Limiting Spectral Distribution of Large-Dimensional Sample Covariance Matrices. *Annals of Probability*, 26(1):316–345, 1998.

[7] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data. *Journal of Machine Learning Research*, 9:485–516, 2008.

[8] A. J. Baranchick. A Family of Minimax Estimators of the Mean of a Multivariate Normal Distribution. *Annals of Mathematical Statistics*, 41:642–645, 1970.

[9] Asbjorn Berge, Are C. Jensen, and Anne H. Schistad Solberg. Sparse Inverse Covariance Estimates for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 45(5):1399–1407, 2007.

[10] Peter J. Bickel and Elizaveta Levina. Covariance Regularization by Thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

[11] Peter J. Bickel and Elizaveta Levina. Regularized Estimation of Large Covariance Matrices. *The Annals of Statistics*, 36(1):199–227, 2008.

[12] Peter J. Bickel and Marko Lindner. Approximating the Inverse of Banded Matrices by Banded Matrices with Applications to Probability and Statistics. *Preprint, available at arXiv:1002.4545v2*, 2010.

[13] Peter J Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[14] Peter Bloomfield and William L. Steiger. *Least Absolute Deviations: Theory, Applications and Algorithms*. Birkhäuser, 1983.

[15] S.P. Brooks, P. Giudici, and G.O. Roberts. Efficient construction of of reversible jump markov chain monte carlo proposal distributions. *J. Royal. Statist. Soc B.*, 65:3–55, 2003.

[16] Florentina Bunea, Alexandre Tsybakov, and Marten Wegkamp. Sparsity orcale inequalities for the LASSO. *Electronic Journal of Statistics*, 1:169–194, 2007.

[17] Tony Cai and Weidong Liu. Adaptive Thresholding for Sparse Covariance Matrix Estimation. *To appear in Journal of American Statistical Association*, 2011.

[18] Tony Cai, Weidong Liu, and Xi Luo. A Constrained ℓ_1 Minimization Approach to Sparse Precision Matrix Estimation. *To appear in Journal of American Statistical Association*, 2011.

[19] Tony Cai, Guangwu Xu, and Jun Zhang. On recovery of sparse signals via ℓ_1 minimization. *IEEE Transactions on Information Theory*, 57(7):3388–3397, 2009.

[20] Tony Cai, Cun-Hui Zhang, and Harrison Zhou. Optimal Rates of Convergence for Covariance Matrix Estimation. *The Annals of Statistics*, 38(4):2118–2144, 2010.

[21] Tony Cai and Harrison Zhou. Minimax Estimation of Large Covariance Matrices under ℓ_1-norm. *To appear in Statistica Sinica*, 2011.

[22] Tony Cai and Harrison Zhou. Optimal Rates of Convergence for Sparse Covariance Matrix Estimation. *Preprint*, 2011.

[23] Emmanuel Candès and Yaniv Plan. Near-ideal model selection by ℓ_1 minimization. *The Annals of Statistics*, 37(5):2145–2177, 2009.

[24] Emmanuel Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52:489–509, 2004.

[25] Emmanuel Candès, Justin Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. on Pure and Applied Math.*, 59(8):1207–1223, 2006.

[26] Emmanuel Candès and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[27] Emmanuel Candès and Terence Tao. The Dantzig selector: Statistical estima-

tion when p is much larger than n. *The Annals of Statistics*, 35:2313–2351, 2007.

[28] Scott Shaobing Chen, David L. Donoho, and Michael Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

[29] Xi Chen, Seunghak Kim, Qihang Lin, Jaime G. Carbonell, and Eric P. Xing. Graph-Structured Multi-task Regression and an Efficient Optimization Method for General Fused Lasso. *Preprints, arXiv*, 2010.

[30] Xiaohui Chen, Young-Heon Kim, and Z. Jane Wang. Efficient Minimax Estimation of A Class of High-Dimensional Sparse Precision Matrices. *IEEE Transactions on Signal Processing, to appear*, 2012.

[31] Xiaohui Chen, Z. Jane Wang, and Martin J. McKeown. Asymptotic analysis of robust LASSOs in the presence of noise with large variance. *IEEE Transactions on Information Theory*, 56(10):5131–5149, 2010.

[32] Xiaohui Chen, Z. Jane Wang, and Martin J. McKeown. A Bayesian Lasso via reversible-jump MCMC. *Signal Processing*, 91(8):1920–1932, 2011.

[33] Xiaohui Chen, Z. Jane Wang, and Martin J. McKeown. Shrinkage-to-tapering estimation of large covariance matrices. *IEEE Transactions on Signal Processing, revision submitted*, 2011.

[34] Xiaohui Chen, Z.Jane Wang, and Martin. J McKeown. Asymptotic analysis of the Huberized LASSO estimator. *The 35th International Conference on Acoustics, Speech, and Signal Processing*, pages 1898–1901, 2010.

[35] Xiaohui Chen, Z.Jane Wang, and Martin J. McKeown. fMRI group studies of brain connectivity via a group robust LASSO. *International Conference on Image Processing*, pages 1–4, 2010.

[36] Yilun Chen, Ami Wiesel, Yonina C. Eldar, and Alfred O. Hero. Shrinkage Algorithms for MMSE Covariance Estimation. *IEEE Transactions on Signal Processing*, 58(10):5016–5029, 2010.

[37] Roshan Cools, Elka Stefanova, Roger A. Barker, Trevor W. Robbins, and Adrian M. Owen. Dopaminergic modulation of high-level cognition in parkinson's disease: the role of the prefrontal cortex revealed by pet. *Brain*, 125(4):584–594, 2002.

[38] Don Coppersmith and Shmuel Winograd. Matrix Multiplication via Arithmetic Progressions. *Journal of Symbolic Computation*, 9(3):251–280, 1990.

[39] Hidde de Jong. Modeling and Simulation of Genetic Regulatory Systems: A Literature Review. *Journal of Computational Biology*, 9(1):67–103, 2002.

[40] Forster J. Dellaportas, P. and I. Ntzoufras. On bayesian model and variable selection using mcmc. *Statistics and Computing*, 12:27–36, 2002.

[41] David L. Donoho. For most large underdetermined systems of equations, the minimal ℓ_1-norm near-solution approximates the sparsest near-solution. *Communications on Pure and Applied Mathematics*, 59(6):907–934, 2006.

[42] David L. Donoho. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, 2006.

[43] David L. Donoho, Michael Elad, and Vladimir Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52:6–18, 2006.

[44] Rick Durrett. *Probability: Theory and Examples.* Duxbury Advanced Series, third edition, 2005.

[45] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least Angle Regression (with discussion). *The Annals of Statistics*, 32(2):407–499, 2004.

[46] Noureddine El Karoui. Tracy-Widom Limit for the Largest Eigenvalue of a Large Class of Complex Sample Covariance Matrices. *The Annals of Probability*, 35(2):663–714, 2007.

[47] Noureddine El Karoui. Operator Norm Consistent Estimation of Large-dimensional Sparse Covariance Matrices. *The Annals of Statistics*, 36(6):2717–2756, 2008.

[48] Yonina C. Eldar and Moshe Mishali. Robust recovery of signals from a structured union of subspaces. *IEEE Transactions on Information Theory*, 11:5302–5316, 2009.

[49] Alexandre Eusebio, Alek Pogosyan, Shouyan Wang, and et al. Resonance in subthalamo-cortical circuits in parkinson's disease. *Brain*, 132(8):2139–2150, 2009.

[50] Jianqing Fan, Yang Feng, and Yichao Wu. Network Exploration via the Adaptive Lasso and SCAD penalties. *The Annals of Applied Statistics*, 3(2):521–541, 2009.

[51] JianQing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association*, 96(4456):1348–1360, 2001.

[52] M. Figueiredo. Adaptive sparseness for supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:1150–1159, 2003.

[53] Thomas J. Fisher and Xiaoqian Sun. Improved Stein-type Shrinkage Estimators for the High-Dimensional Multivariate Normal Covariance Matrix. *Computational Statistics and Data Analysis*, 55:1909–1918, 2011.

[54] Dean P. Foster and Edward I. George. The risk inflation criterion for multiple regression. *The Annals of Statistics*, 22:1947–1975, 1994.

[55] Ildiko E. Frank and Jerome H. Friedman. A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, 35:109–148, 1993.

[56] Jerome Friedman and Tibshirani Robert Hastie, Travor. Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics*, 9(3):432–441, 2008.

[57] Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

[58] E. George and R. McCullogh. Approaches for bayesian variable selection. *Statistica Sinica*, 7:339–373, 1997.

[59] Edward I. George and Dean P. Foster. Calibration and empirical bayes variable selection. *Biometrika*, 87(4):731–747, 2000.

[60] J. Charles Geyer. On the Asymptotics of Convex Stochastic Optimization. *Unpublished manuscripts*, pages 1–17, 1996.

[61] W.. Gilks, S. Richardson, and David Spiegelhalter. *Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics*. Chapman & Hall/CRC, 1 edition, 1995.

[62] Michael Grant and Boyd Stephen. CVX: Matlab software for disciplined convex programming (web page and software), 2009. `http://stanford.edu/~boyd/cvx`.

[63] Peter Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82:711–732, 1995.

[64] Joseph R. Guerci. Theory and Application of Covariance Matrix Tapers for Robust Adaptive Beamforming. *IEEE Transactions on Signal Processing*, 47(4):977–985, 1999.

[65] L. Harrison, W. Penny, and K. Friston. Multivariate autoregressive modeling of fMRI time series. *Neuroimage*, 19:1273–1302, 2003.

[66] David Hastie. *Towards Automatic Reversible Jump Markov Chain Monte Carlo*. PhD thesis, University of Bristol, 2004.

[67] W.K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[68] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(3):55–67, 1970.

[69] Patrik O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.

[70] David Hunter and Runze Li. Variable selection using mm algorithms. *Annals of Statistics*, 33:1617–1642, 2005.

[71] W James and Charles Stein. Estimation with Quadratic Loss. *In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pages 361–379, 1961.

[72] Iain Johnstone. On the Distribution of the Largest Eigenvalue in Principle Components Analysis. *The Annals of Statistics*, 29(2):295–327, 2001.

[73] Iain M. Johnstone and Arthur Yu Lu. Sparse Principal Components Analysis. *ArXiv:0901.4392v1*, 2004.

[74] Joseph Kadane and Nicole Lazar. Methods and criteria for model selection. *Journal of the American Statistical Association*, 99:279–290, 2004.

[75] Jafar A. Khan, Stefan Van Aelst, and Ruben H. Zamare. Robust Linear Model Selection Based on Least Angle Regression. *Journal of American Statistical Association*, 102(480):1289–1299, 2007.

[76] Seunghak Kim and Eric P. Xing. Tree-Guided Group Lasso for Multi-Task Regression with Structured Sparsity. *ICML*, 2010.

[77] Keith Knight and Wenjiang Fu. Asymptotics for LASSO-type estimators. *The Annals of Statistics*, 28(5):1356–1378, 2000.

[78] Mladen Kolar and Eric Xing. On time varying undirected graphs. *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS) 2011*, (JMLR) 15:407–415, 2011.

[79] L. Kuo and B. Mallick. Variable selection for regression models. *Sankhya*, 60:65–81, 1998.

[80] Clifford Lam and Jianqing Fan. Sparsistency and Rates of Convergence in Large Covariance Matrix Estimation. *The Annals of Statistics*, 37(6):4254–4278, 2009.

[81] K. Lange, D. Hunter, and H. Yang. Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, 9(1):1–20, 2000.

[82] Olivier Ledoit and Michael Wolf. A Well-Conditioned Estimator for Large Dimensional Covariance Matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

[83] Michel Ledoux. *The Concentration of Measure Phenomenon.* Mathematical Surveys and Monographs Volume 89, American Mathematical Society, 2003.

[84] Seunghak Lee, Jun Zhu, and Eric P. Xing. Adaptive Multi-Task Lasso: with Application to eQTL Detection. *NIPS*, 2010.

[85] Will E. Leland, Murad S. Taqqu, Walter Willinger, and Daniel V. Wilson. On the Self-Similar Nature of Ethernet Traffic (Extended Version). *IEEE/ACM Transactions on Networking*, 2(1):1–15, 1994.

[86] Chenlei Leng, Yi Lin, and Grace Wahba. A note on the LASSO and related procedures in model selection. *Statistica Sinica*, 16:1273–1284, 2006.

[87] Jian Li, Petre Stocia, and Zhisong Wang. On Robust Capon Beamforming and Diagonal Loading. *IEEE Transactions on Signal Processing*, 51(7):1702–1715, 2003.

[88] Junning Li, Z.Jane Wang, and Martin. J McKeown. Dynamic Bayesian Network Modelling of fMRI: A Comparison of Group Analysis Methods. *NeuroImage*, 41(2):398–407, 2008.

[89] Dennis Lindley. A statistical paradox. *Biometrika*, 44(4):187–192, 1957.

[90] Karim Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics*, 2:90–102, 2008.

[91] Guillaume Marrelec, Pierre Bellec, and Habib Benali. Exploring Large-Scale Brain Networks in Functional MRI . *Journal of Physiology-Paris*, 100(4):171–181, 2006.

[92] A. McIntosh, C. Grady, L. Ungerleider, J. Haxby, and et al. Network analysis of cortical visual pathways mapped with pet. *Journal of Neuroscience*, 14(2):655–666, 1994.

[93] A. McIntosh, C. Grady, L. Ungerleider, J. Haxby, and et al. Network analysis of cortical visual pathways mapped with pet. *Journal of Neuroscience*, 14(2):655–666, 1994.

[94] Nicolai Meinshausen. Relaxed Lasso. *Computational Statistics and Data Analysis*, 52(1):374–393, 2007.

[95] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

[96] Nicolai Meinshausen and Bin Yu. LASSO-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246–270, 2009.

[97] Sean P. Meyn and Richard Tweedie. *Markov Chains and Stochastic Stability.* Springer-Verlag, London, 1999.

[98] T. Mitchell and J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83:1023–1036, 1988.

[99] Crispin M. Mutshinda and Mikko J. Sillanpää. Extended bayesian lasso for multiple quantitative trait loci mapping and unobserved phenotype prediction. *Genetics*, 186(3):1067–1075, 2010.

[100] Michael R. Osborne, Brett Presnell, and Berwin A. Turlach. On the LASSO and Its Dual. *Journal of Computational and Graphical Statistics*, 9:319–337, 1999.

[101] Michael R. Osborne, Brett Presnell, and Berwin A. Turlach. A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20(3):389–403, 2000.

[102] T. Park and G. Casella. The bayesian lasso. *Journal of the American Statistical Association*, Vol. 103(482):681–686, 2008.

[103] Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica*, 17:1617–1642, 2007.

[104] P. Pikkuhookana and Mikko J. Sillanpää. Correction of relatedness in bayesian models for genomic data association analysis. *Heredity*, 103:223–237, 2009.

[105] David Pollard. Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7(2):186–199, 1991.

[106] O'Hara RB and Mikko J. Sillanpää. A review of bayesian variable selection methods: what, how and which. *Bayesian Analysis*, 4:85–118, 2009.

[107] Saharon Rosset and Ji Zhu. Piecewise linear regularized solution paths. *The Annals of Statistics*, 35(3):1012–1030, 2007.

[108] Adam J. Rothman, Peter J. Bickel, Elizaveta Levina, and Ji Zhu. Sparse Permutation Invariant Covariance Estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

[109] Gideon E. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[110] Jun Shao. *Mathematical Statistics*. Springer, second edition, 2003.

[111] Mikko J. Sillanpää and Elja Arjas. Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics*, 148:1373–1388, 1998.

[112] Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *In Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 197–206, 1956.

[113] Klaas Enno Stephan, Will D. Penny, Daunizeau Jean, Rosalyn J. Moran, and Karl J. Friston. Bayesian model selection for group studies. *NeuroImage*, 46(4):1004–17, 2009.

[114] Lennart Svensson and Lundberg Nordenvaad. The Reference Prior for Complex Covariance Matrices with Efficient Implementation Strategies. *IEEE Transactions on Signal Processing*, 58(1):53–66, 2010.

[115] Robert Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

[116] Luke Tierney. Markov chains for exploring posterior distributions (with discussion). *The Annals of Statistics*, 22(4):1701–1762, 1994.

[117] Joel.A. Tropp. Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52:1030–1051, 2006.

[118] P.A. Valdes-Sosa, J.M. Sanchez-Bornot, and et al. Estimating brain functional connectivity with sparse multivariate autoregression. *Phil. Trans. R. Soc. B*, 360:969981, 2005.

[119] Sanchez-Bornot Jose Valdés-Sosa, Pedro, Agustín Lage-Castellanos, Mayrim Vega-Hernández, Jorge Bosch-Bayard, Lester Melie-García, and Erick Canales-Rodríguez. Estimating brain functional connectivity with sparse multivariate autoregression. *Phil. Trans. R. Soc. B*, 360:969–81, 2005.

[120] Roman Vershynin. Introduction to the Non-asymptotic Analysis of Random Matrices. *Preprint*, 2010.

[121] Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy recovery of sparsity. *Technical Report 708, Department of Statistics, UC Berkeley*, 2006.

[122] Hansheng Wang, Guodong Li, and Guohua Jiang. Robust regression shrinkage and consistent variable selection via the LAD-LASSO. *Journal of Business and Economic Statistics*, 11:1–6, 2006.

[123] Li Wang, Michael D. Gordon, and Ji Zhu. Regularized Least Absolute Deviations Regression and an Efficient Algorithm for Parameter Tuning. *Sixth International Conference on Data Mining*, pages 690–700, 2006.

[124] Tongtong Wu and Kenneth Lange. Coordinate descent algorithms for LASSO penalized regression. *The Annals of Applied Statistics*, 2(1):224–244, 2008.

[125] Wei Biao Wu. Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences*, 102(40):14150–14154, 2005.

[126] Ming Yuan. High Dimensional Inverse Covariance Matrix Estimation via Linear Programming. *Journal of Machine Learning Research*, 11:2261–2286, 2010.

[127] Ming Yuan and Yi Lin. Efficient Empirical Bayes Variable Selection and Estimation in Linear Models. *Journal of the American Statistical Association*, 100:1215–1225, 2005.

[128] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2005.

[129] Raphael Yuster and Uri Zwick. Fast Sparse Matrix Multiplication. *ACM Transactions on Algorithms*, 1(1):2–13, 2005.

[130] Peng Zhao and Bin Yu. On model selection consistency of LASSO. *Journal of Machine Learning Research*, 7:2451–2563, 2006.

[131] Shuheng Zhou, John Lafferty, and Larry Wasserman. Time varying undirected graphs. *Machine Learning*, 80(2-3):295–319, 2010.

[132] Hui Zou. The Adaptive Lasso and Its Oracle Properties. *Journal of American Statistical Association*, 101(476):1418–1429, 2006.

Appendix

A.1 Notations

We fix the notation system that is used throughout the thesis. Additional specialized symbols are listed individually in each chapters. We will denote the number of variables/parameters and the number of samples by p and n, respectively. In particular, p will be the number of the coefficients in the linear regression models and the number of variables in the covariance/precision matrix estimation problems.

Vector and Matrix

Bold lower letters, e.g. $\mathbf{a}, \mathbf{x}, \cdots$, are used to denote vectors; capital letters, e.g. $A, B, \Sigma, \Omega, \cdots$, for matrices; and curly capital letters, e.g. $\mathcal{G}, \mathcal{S}, \cdots$, for a collection of elements such as matrices. For a generic vector $\mathbf{a}$, standard norm notations are used such as $\|\mathbf{a}\|_{\ell_1} := \|\mathbf{a}\|_1 = \sum_j |a_j|$, $\|\mathbf{a}\|_{\ell_2} := \|\mathbf{a}\|_2 = \sqrt{\sum_j |a_j|^2}$, $\|\mathbf{a}\|_{\ell_\infty} := \|\mathbf{a}\|_\infty = \sup_j |a_j|$, and $\|\mathbf{a}\|_{\ell_0} := \|\mathbf{a}\|_0 = \sum_j \mathbb{I}(a_j \neq 0)$. For a generic matrix A and $1 \leq r \leq \infty$, the matrix L^r norm is defined as

$$\|A\|_{L^r} = \sup_{\|\mathbf{x}\|_r = 1} \|A\mathbf{x}\|_r \,. \tag{2}$$

For a square matrix A, the matrix L^1, L^2, L^∞, and Frobenius norms are defined as:

$$\begin{aligned} \|A\|_{L^1} &= \sup_j \sum_i |a_{ij}| && (3) \\ \|A\|_{L^2} &= \sup_{\|\mathbf{x}\|_2=1} \sqrt{\mathbf{x}^T A^T A \mathbf{x}} && (4) \\ \|A\|_{L^\infty} &= \sup_i \sum_j |a_{ij}| && (5) \\ \|A\|_F &= \sqrt{\sum_i \sum_j a_{ij}^2}, && (6) \end{aligned}$$

respectively. One verifies that $\|A\|_{L^2} \leq \|A\|_{L^1}$. Note that for a symmetric square matrix A, $\|A\|_{L^1} = \|A\|_{L^\infty}$ and $\|A\|_{L^2} = \sup_j \{|\lambda_j(A)|\}$ where $\lambda_j(A)$ is the j-th eigenvalue of A, i.e. $\|A\|_{L^2}$ is the spectral norm of A (a.k.a. the operator norm of A as a linear functional from $\ell_2 \to \ell_2$). Note that, in the rest of the paper, it is assumed $A^T = A$ unless otherwise indicated since we focus on considering covariance and precision matrices. Moreover, for simplicity, we shall skip the subscript in $\|\cdot\|_2$ for the Euclidean norm of a vector and the spectral norm of a matrix. We will also use the entry-wise norm on matrices, which is tantamount to regard matrices as vectors. For instances, $\|A\|_\infty$ stands for the maximum magnitude of the entries of A and $\|A\|_1 = \sum_i \sum_j |a_{ij}|$. The "dist" function defined on matrices for the distance between a point and a set, which coincides with the usual definition $\text{dist}(A, \mathcal{S}) = \inf\{\|A - B\| : B \in \mathcal{S}\}$. If $\mathcal{S}$ is a closed subset, the infimum is attained.

Let $\mathbf{x}_i$ be the i^{th}-row of X and $\mathbf{x}^j$ be the j^{th}-column of a matrix X. Let $\text{Tr}(M)$ denote the trace of a squared matrix M. M_{11}, M_{12}, M_{21}, and M_{22} are submatrices of M partitioned according to

$$M = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}.$$

For a generic vector $\mathbf{u}$, we interchangeably use u_j and $u[j]$ to denote the j^{th}-element of $\mathbf{u}$. Define $\text{supp}(\mathbf{u}) = \{j \in \{1, \cdots, p\} | u_j \neq 0\}$ and define $|\mathbf{u}|$ to be the cardinality of $\mathbf{u}$.

Probability

$Z_n \xrightarrow{P} Z$ refers to convergence in probability and $Z_n \Rightarrow Z$ convergence in distribution. Note that we shall also use capital letters to mean random variables; this should not cause confusion with the matrix notation when the context is clear. We use the phase *with high probability* to mean that the referred probability approaches to 1 when $n \to \infty$ (thus, $p \to \infty$ as well). $1(A)$ is the indicator function of some measurable set A. For a random variable Z, $E(Z; A) = \int_A Z \, dP$ is the expectation of Z taken on A. For two probability measures P and Q with a common dominating measure μ, let p and q be the density of P and Q, respectively. Then the total variation affinity between P and Q is defined as $\|P \wedge Q\| = \int \min(p, q) \, d\mu$. We use $|A|$ to denote the size or cardinality of a set A. $\hat{\boldsymbol{\beta}}_n$ is said to be a $\sqrt{n}$-consistent estimator for $\boldsymbol{\beta}$ if $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ has an unbiased limiting distribution with respect to (w.r.t.) $\mathbf{0}$.

A.2 Proofs for Chapter 2

A.2.1 Proof of Theorem 2.2.1

Let $Z_n(\mathbf{u}) = n^{-1} \sum_{i=1}^n L(\mathbf{u}; y_i, \mathbf{x}_i) + n^{-1}\lambda_n \|\mathbf{u}\|_{\ell_1}$ and $\hat{\boldsymbol{\beta}}_n$ minimize Z_n. Then it follows that $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ minimizes V_n, where

$$\begin{aligned} V_n(\mathbf{u}) = & \sum_{i=1}^n \left[\delta \left(\left(e_i - \frac{\mathbf{u}^* \mathbf{x}_i}{\sqrt{n}} \right)^2 - e_i^2 \right) + (1-\delta) \left(\left| e_i - \frac{\mathbf{u}^* \mathbf{x}_i}{\sqrt{n}} \right| - |e_i| \right) \right] \\ & + \lambda_n \sum_{j=1}^p \left[\left| \beta_j + \frac{u_j}{\sqrt{n}} \right| - |\beta_j| \right]. \end{aligned} \tag{7}$$

Without loss of generality (WLOG), by symmetry we can assume $\frac{\mathbf{u}^* \mathbf{x}_i}{\sqrt{n}} \geq 0$. Put

$$Y_i = (1-\delta)(1(e_i < 0) - 1(e_i \geq 0)) - 2\delta e_i$$

and

$$Z_i = 2 \times \left(\frac{\mathbf{u}^* \mathbf{x}_i}{\sqrt{n}} - e_i \right) 1 \left(0 \le e_i < \frac{\mathbf{u}^* \mathbf{x}_i}{\sqrt{n}} \right).$$

Then

$$\begin{aligned} V_n(\mathbf{u}) &= \delta \sum_{i=1}^{n} \left(\mathbf{u}^* \frac{\mathbf{x}_i \mathbf{x}_i^*}{n} \mathbf{u} \right) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{u}^* \mathbf{x}_i Y_i + (1-\delta) \sum_{i=1}^{n} Z_i \\ &\quad + \lambda_n \sum_{j=1}^{p} \left[\left| \beta_j + \frac{u_j}{\sqrt{n}} \right| - |\beta_j| \right]. \end{aligned} \tag{8}$$

With the error assumption, we have

$$EY_i = (1-\delta)(P(e_i < 0) - P(e_i \ge 0)) - 2\delta E e_i = 0$$

and

$$\begin{aligned} \mathrm{Var}\,(Y_i) &= (1-\delta)^2 E[1(e_i < 0) - 1(e_i \ge 0)]^2 + 4\delta^2 E e_i^2 \\ &- 4\delta(1-\delta)\left[E(e_i; e_i < 0) - E(e_i; e_i \ge 0)\right] \\ &= (1-\delta)^2 + 4\delta^2\sigma^2 + 4\delta(1-\delta)E|e_i|, \end{aligned}$$

since e_i has a symmetric distribution. Note that $\sigma^2 < \infty$ is used here.

Now by assumption 1, we have

$$\sum_{i=1}^{n} \left(\mathbf{u}^* \frac{\mathbf{x}_i \mathbf{x}_i^*}{n} \mathbf{u} \right) \to \mathbf{u}^* C \mathbf{u},$$

by Lemma A.2.5

$$\sum_{i=1}^{n} Z_i \xrightarrow{P} f(0) \mathbf{u}^* C \mathbf{u}$$

and

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{u}^* \mathbf{x}_i Y_i \Rightarrow \mathbf{u}^* \mathbf{W}$$

where $\mathbf{W} \sim N(\mathbf{0}, ((1-\delta)^2 + 4\delta^2\sigma^2 + 4\delta(1-\delta)M_{10})\, C)$, and

$$\lambda_n \sum_{j=1}^{p} \left[\left| \beta_j + \frac{u_j}{\sqrt{n}} \right| - |\beta_j| \right] \to \lambda_0 \sum_{j=1}^{p} [u_j \mathrm{sgn}\,(\beta_j)\, 1(\beta_j \neq 0) + |u_j| 1(\beta_j = 0)].$$

Combining all terms together and applying Slutsky's lemma (c.f. p60, [110]), we deduce that $V_n(\mathbf{u}) \Rightarrow V(\mathbf{u})$, where

$$\begin{aligned} V(\mathbf{u}) &= \delta \mathbf{u}^* C \mathbf{u} + (1-\delta) f(0) \mathbf{u}^* C \mathbf{u} + \mathbf{u}^* \mathbf{W} \\ &+ \lambda_0 \sum_{j=1}^{p} [u_j \mathrm{sgn}\,(\beta_j)\, 1(\beta_j \neq 0) + |u_j| 1(\beta_j = 0)]. \end{aligned} \tag{9}$$

It is obvious that the finite-dimensional convergence holds. Finally, since V_n is convex and V has a unique minimum, the epi-convergence result from [60] implies that $\arg\min(V_n) = \sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \Rightarrow \arg\min(V)$. □

A.2.2 Proof of Corollary 2.2.2

$\lambda_0 = 0$ implies that

$$V(\mathbf{u}) = (\delta + (1-\delta) f(0))\, \mathbf{u}^* C \mathbf{u} + \mathbf{u}^* \mathbf{W},$$

which is minimized at

$$\arg\min(V) = -\frac{C^{-1}\mathbf{W}}{2(\delta + (1-\delta) f(0))} \sim N\left(\mathbf{0}, \frac{(1-\delta)^2 + 4\delta^2\sigma^2 + 4\delta(1-\delta)M_{10}}{4(\delta + (1-\delta)f(0))^2} C^{-1} \right).$$

□

A.2.3 Proof of Corollary 2.2.3

By the law of large numbers and the fact that $\hat{\boldsymbol{\beta}}_{LS}$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}$ under assumptions 1 and 2(c),

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n-p}\sum_{i=1}^{n}\left(e_i - \left(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}\right)^* \mathbf{x}_i\right)^2 \\
&= \frac{n}{n-p}\left[\frac{1}{n}\sum_{i=1}^{n} e_i^2 - 2\left(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}\right)^* \frac{1}{n}\sum_{i=1}^{n} e_i \mathbf{x}_i \right. \\
&\quad \left. + \left(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}\right)^* \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^* \left(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}\right)\right] \\
&\xrightarrow{P} \sigma^2.
\end{aligned} \tag{10}$$

The corollary then follows from Theorem 2.2.1. □

A.2.4 Proof of Theorem 2.2.4

Define $Z_n(\mathbf{u}) = n^{-1}\sum_{i=1}^{n} L(\mathbf{u}; y_i, \mathbf{x}_i) + n^{-1}\lambda_n \|\mathbf{u}\|_{\ell_1}$, where $L(u; y_i, \mathbf{x}_i) = \delta\left(y_i - \mathbf{u}^*\mathbf{x}_i\right)^2 + (1-\delta)\left|y_i - \mathbf{u}^*\mathbf{x}_i\right|$. It suffices to show that

1. For any compact subset $K \subseteq \mathbb{R}^p$,

$$\sup_{\mathbf{u}\in K} |Z_n(\mathbf{u}) - Z(\mathbf{u})| \xrightarrow{P} 0. \tag{11}$$

2. $\hat{\boldsymbol{\beta}}_n = O_p(1)$.

Put $r_i = y_i - \mathbf{u}^*\mathbf{x}_i$. For part 1, consider

$$\begin{aligned}
Z_n(\mathbf{u}) = \delta(\mathbf{u}-\boldsymbol{\beta})^*\left(\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^*}{n}\right)(\mathbf{u}-\boldsymbol{\beta}) + \frac{\delta}{n}\sum_{i=1}^{n} e_i^2 \\
- \frac{2\delta}{n}\sum_{i=1}^{n}(\mathbf{u}-\boldsymbol{\beta})^*\mathbf{x}_i e_i + (1-\delta)\frac{1}{n}\sum_{i=1}^{n}|r_i| + \frac{\lambda_n}{n}\|\mathbf{u}\|_{\ell_1}.
\end{aligned} \tag{12}$$

The first three terms are easily seen to converge in probability to $\delta[(\mathbf{u}-\boldsymbol{\beta})^*C(\mathbf{u}-\boldsymbol{\beta})+\sigma^2]$. For the fourth term, note that,

$$\frac{\sum_{i=1}^n Er_i^2}{n^2} \to 0.$$

By a weak law of large numbers (c.f. Theorem 1.14 in [110]), we have

$$\frac{1}{n}\sum_{i=1}^n (|r_i| - E|r_i|) \xrightarrow{P} 0.$$

So it follows that

$$\frac{1}{n}\sum_{i=1}^n |r_i| \xrightarrow{P} \frac{1}{n}\sum_{i=1}^n E|r_i|, \tag{13}$$

provided that the limit of $\frac{1}{n}\sum_{i=1}^n E|r_i|$ exists. Now, we show that the sequence $\{\frac{1}{n}\sum_{i=1}^n E|r_i|\}_{n\in\mathbb{N}}$ is Cauchy. Consider

$$\begin{aligned}
& \left|\frac{1}{n+1}\sum_{i=1}^{n+1} E|r_i| - \frac{1}{n}\sum_{i=1}^{n} E|r_i|\right| \\
= & \left|\frac{1}{n+1}\sum_{i=1}^{n+1} E|r_i| - \frac{1}{n}\sum_{i=1}^{n+1} E|r_i| + \frac{1}{n}\sum_{i=1}^{n+1} E|r_i| - \frac{1}{n}\sum_{i=1}^{n} E|r_i|\right| \\
\leq & \frac{1}{n(n+1)}\sum_{i=1}^{n+1} E|r_i| + \frac{1}{n}E|r_{n+1}| \\
\leq & \frac{1}{n^2}\sum_{i=1}^{n+1}\left[M_{10} + |(\boldsymbol{\beta}-\mathbf{u})^*\mathbf{x}_i|\right] + \frac{1}{n}E\left[|e_{n+1}| + |(\boldsymbol{\beta}-\mathbf{u})^*\mathbf{x}_{n+1}|\right].
\end{aligned}$$

Then it follows that, with probability 1, the last quantity converges to 0 by Lemma A.2.1 and A.2.2, as $\max_{1\leq i\leq n}|\mathbf{x}_i| = o(\sqrt{n})$. Thus, the limit of $n^{-1}\sum_{i=1}^n E|r_i|$ exists, which is denote by r. Therefore we can conclude that Z_n converges in probability to a function Z, where

$$Z(\mathbf{u}) = \delta(\mathbf{u}-\boldsymbol{\beta})^*C(\mathbf{u}-\boldsymbol{\beta}) + \delta\sigma^2 + (1-\delta)r + \lambda_0\left\|\mathbf{u}\right\|_{\ell_1}.$$

Since $\{Z_n\}_{n\in\mathbb{N}}$ are convex, it follows from the convexity lemma [105] that Z is necessarily convex and the pointwise convergence in probability can be strengthened to the uniform convergence on compact sets. Part 1 is thus proved.

For part 2, since

$$Z_n(\mathbf{u}) \geq \frac{\delta}{n}\sum_{i=1}^{n}(y_i - \mathbf{u}^*\mathbf{x}_i)^2,$$

where the minimum of the RHS is bounded in probability, it follows that $\hat{\boldsymbol{\beta}}_n = O_p(1)$.

As a particular case that $\lambda_n = o(n)$, we can see that

$$\liminf_{n\to\infty} Z_n(\mathbf{u}) \geq \delta(\mathbf{u}-\boldsymbol{\beta})^*C(\mathbf{u}-\boldsymbol{\beta}) + \delta\sigma^2 + (1-\delta)M_{10} \tag{14}$$

and

$$\begin{aligned}\limsup_{n\to\infty} Z_n(\mathbf{u}) \leq{} & \delta(\mathbf{u}-\boldsymbol{\beta})^*C(\mathbf{u}-\boldsymbol{\beta}) + \delta\sigma^2 + (1-\delta)M_{10} \\ & + 2(1-\delta)\limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}|(\mathbf{u}-\boldsymbol{\beta})^*\mathbf{x}_i|.\end{aligned} \tag{15}$$

It is then clear that $\limsup_{n\to\infty} Z_n(\boldsymbol{\beta}) \leq \limsup_{n\to\infty} Z_n(\mathbf{u})$. Since we have shown $Z_n \xrightarrow{P} Z$, it follows that

$$Z(\boldsymbol{\beta}) \leq Z(\mathbf{u}).$$

I.e. $\boldsymbol{\beta}$ minimizes $Z(\mathbf{u})$. The proof is complete. □

A.2.5 Proof of Proposition 2.2.5

Define

$$Z_n(\mathbf{u}) = \sum_{i=1}^{n}\left[\delta(\mathbf{u}-\boldsymbol{\beta})^*\mathbf{x}_i\mathbf{x}_i^*(\mathbf{u}-\boldsymbol{\beta}) + (1-\delta)\left|(\mathbf{u}-\boldsymbol{\beta})^*\mathbf{x}_i\right|\right] + \lambda_n\left\|\boldsymbol{\beta}\right\|_{\ell_1}$$

and

$$\begin{aligned} V_n(\mathbf{u}) &= \sum_{i=1}^{n} \left[\delta \mathbf{u}^* \mathbf{x}_i \mathbf{x}_i^* \mathbf{u} + (1-\delta) \left|\mathbf{u}^* \mathbf{x}_i\right|\right] + \lambda_n \sum_{j=1}^{p} \left[|\beta_j + u_j| - |\beta_j|\right] \\ &= \delta n \mathbf{u}^* C_n \mathbf{u} + (1-\delta) \sum_{i=1}^{n} |\mathbf{u}^* \mathbf{x}_i| + \lambda_n \sum_{j=1}^{p} \left[|\beta_j + u_j| - |\beta_j|\right]. \end{aligned}$$

Let $\hat{\boldsymbol{\beta}}_n^0$ be a minimizer of $Z_n(\mathbf{u})$ over $\mathbf{u}$. Then $\mathbf{h} = \hat{\boldsymbol{\beta}}_n^0 - \boldsymbol{\beta}$ minimizes $V_n(\mathbf{u})$. Note that $V_n(\mathbf{0}) = 0$, therefore $V_n(\mathbf{h}) \leq 0$. Since

$$\delta n \mathbf{h}^* C_n \mathbf{h} + (1-\delta) \sum_{i=1}^{n} |\mathbf{h}^* \mathbf{x}_i| \geq 0,$$

we have

$$\sum_{j \notin A} |h_j| + \sum_{j \in A} (|\beta_j + h_j| - |\beta_j|) \leq 0$$

such that

$$\|\mathbf{h}\|_{\ell_1(A^c)} \leq \left| \sum_{j \in A} (|\beta_j + h_j| - |\beta_j|) \right| \leq \sum_{j \in A} \left||\beta_j + h_j| - |\beta_j|\right| \leq \sum_{j \in A} |h_j| = \|\mathbf{h}\|_{\ell_1(A)}.$$

Since $\|\mathbf{h}\|_{\ell_0(A)} \leq S$, it follows that

$$\|\mathbf{h}\|_{\ell_1(A)} \leq \sqrt{S} \, \|\mathbf{h}\|_{\ell_2(A)} \leq \sqrt{S} \, \|\mathbf{h}\|_{\ell_2}, \tag{16}$$

whereas

$$\|\mathbf{h}\|_{\ell_1} = \|\mathbf{h}\|_{\ell_1(A)} + \|\mathbf{h}\|_{\ell_1(A^c)} \leq 2 \, \|\mathbf{h}\|_{\ell_1(A)} \leq 2\sqrt{S} \, \|\mathbf{h}\|_{\ell_2}. \tag{17}$$

Now we can bound $V_n(\mathbf{u})$ from below as

$$V_n(\mathbf{u}) \geq \delta n \mathbf{u}^* C_n \mathbf{u} + (1-\delta)\sqrt{n} \left(\mathbf{u}^* C_n \mathbf{u}\right)^{1/2} + \lambda_n \sum_{j=1}^{p} \left[|\beta_j + u_j| - |\beta_j|\right], \tag{18}$$

where we used

$$\|X\mathbf{u}\|_{\ell_1} \geq \|X\mathbf{u}\|_{\ell_2} = \sqrt{n} \left(\mathbf{u}^* C_n \mathbf{u}\right)^{1/2}.$$

Substituting $\mathbf{h}$ into (18), we get

$$0 \geq V_n(\mathbf{h}) = \delta n \mathbf{h}^* C_n \mathbf{h} + (1-\delta)\sqrt{n}\,(\mathbf{h}^* C_n \mathbf{h})^{1/2} + \lambda_n \sum_{j=1}^{p} \left[|\beta_j + h_j| - |\beta_j|\right]. \qquad (19)$$

So by (16) and (18), we obtain

$$\begin{aligned} \lambda_n \sqrt{S}\, \|\mathbf{h}\|_{\ell_2} &\geq \lambda_n \|\mathbf{h}\|_{\ell_1(A)} \\ &\geq \delta n \mathbf{h}^* C_n \mathbf{h} + (1-\delta)\sqrt{n}\,(\mathbf{h}^* C_n \mathbf{h})^{1/2}. \end{aligned} \qquad (20)$$

Now, we bound $\mathbf{h}^* C_n \mathbf{h}$ from below. WLOG, we can assume that $\mathbf{h}$ is in decreasing order of magnitudes. Let T_1 be the S_0-largest positions of $\mathbf{h}$. Decompose $\mathbf{h} = \mathbf{h}(T_1) + \mathbf{h}(T_1^c)$, where $\mathbf{h}(T_1)$ is the $p \times 1$ vector that is a restricted version of $\mathbf{h}$ to the set T_1 and 0 elsewhere. We note that

$$\begin{aligned} \|\mathbf{h}(T_1^c)\|_{\ell_2}^2 &\leq \sum_{j=S_0+1}^{p} \frac{\|\mathbf{h}\|_{\ell_1}^2}{j^2} \leq \|\mathbf{h}\|_{\ell_1}^2 \sum_{j=S_0+1}^{p} \left(\frac{1}{j-1} - \frac{1}{j}\right) \\ &\leq \frac{\|\mathbf{h}\|_{\ell_1}^2}{S_0} \leq \frac{4S\,\|\mathbf{h}\|_{\ell_2}^2}{S_0}, \end{aligned}$$

where the third inequality follows from the telescope sum. So it follows that

$$\begin{aligned} \mathbf{h}(T_1)^* C_n \mathbf{h}(T_1) &\geq \phi_{\min}(S_0)\, \|\mathbf{h}(T_1)\|_{\ell_2}^2 \\ &= \phi_{\min}(S_0) \left(\|\mathbf{h}\|_{\ell_2}^2 - \|\mathbf{h}(T_1^c)\|_{\ell_2}^2\right) \\ &\geq \phi_{\min}(S_0)\, \|\mathbf{h}\|_{\ell_2}^2 \left(1 - \frac{4S}{S_0}\right). \end{aligned}$$

Also, since $\|\mathbf{h}\|_{\ell_0(T_1^c)} \leq p - S_0$, we have

$$\begin{aligned} \mathbf{h}(T_1^c)^* C_n \mathbf{h}(T_1^c) &\leq \phi_{\max}(p - S_0)\, \|\mathbf{h}\|_{\ell_0(T_1^c)}^2 \\ &\leq \phi_{\max}(p - S_0) \frac{4S}{S_0} \|\mathbf{h}\|_{\ell_2}^2. \end{aligned}$$

So applying Minkowski's inequality, we conclude that

$$\begin{aligned}\mathbf{h}^*C_n\mathbf{h} &\geq (\mathbf{h}(T_1)^*C_n\mathbf{h}(T_1) - \mathbf{h}(T_1^c)^*C_n\mathbf{h}(T_1^c))^2 \\ &\geq \|\mathbf{h}\|_{\ell_2}^2\left(1 - 4\sqrt{\frac{S\phi_{\max}(p-S_0)}{S_0\phi_{\min}(S_0)}}\right), \end{aligned} \tag{21}$$

where we used $\phi_{\max}(p - S_0) \geq \phi_{\min}(S_0)$. Set $D_0 = 1 - 4\sqrt{\frac{S\phi_{\max}(p-S_0)}{S_0\phi_{\min}(S_0)}} > 0$ where the positivity of D_0 is a consequence of the inherent design. By inserting this estimate of $\mathbf{h}^*C_n\mathbf{h}$ into (20), we get

$$\begin{aligned}\lambda_n\sqrt{S}\,\|\mathbf{h}\|_{\ell_2} &\geq \delta n\mathbf{h}^*C_n\mathbf{h} + (1-\delta)\sqrt{n}\,(\mathbf{h}^*C_n\mathbf{h})^{1/2} \\ &\geq \delta nD_0\,\|\mathbf{h}\|_{\ell_2}^2 + (1-\delta)\sqrt{nD_0}\,\|\mathbf{h}\|_{\ell_2}. \end{aligned} \tag{22}$$

Canceling $\|\mathbf{h}\|_{\ell_2}$ on both sides yields

$$\|\mathbf{h}\|_{\ell_2} \leq \frac{\lambda_n\sqrt{S} - (1-\delta)\sqrt{nD_0}}{\delta nD_0}. \tag{23}$$

This completes the proof. □

A.2.6 Proof of Theorem 2.3.1

As in the proof of Theorem 2.2.1, we let $Z_n(\mathbf{u}) = n^{-1}\sum_{i=1}^n L(\mathbf{u}; y_i, \mathbf{x}_i) + n^{-1}\lambda_n\sum_{j=1}^p \hat{w}_j\,|u_j|$ and $\hat{\boldsymbol{\beta}}_n$ minimize Z_n. Define

$$\begin{aligned}V_n(\mathbf{u}) = \sum_{i=1}^n &\left[\delta\left(\left(e_i - \frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}}\right)^2 - e_i^2\right) + (1-\delta)\left(\left|e_i - \frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}}\right| - |e_i|\right)\right] \\ &+ \lambda_n\sum_{j=1}^p \hat{w}_j\left[\left|\beta_j + \frac{u_j}{\sqrt{n}}\right| - |\beta_j|\right]. \end{aligned} \tag{24}$$

Then $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ minimizes V_n. Rewrite V_n as

$$\begin{aligned} V_n(\mathbf{u}) &= \delta \sum_{i=1}^{n} \left(\mathbf{u}^* \frac{\mathbf{x}_i \mathbf{x}_i^*}{n} \mathbf{u} \right) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{u}^* \mathbf{x}_i Y_i \\ &+ (1-\delta) \sum_{i=1}^{n} Z_i + \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^{p} \hat{w}_j \sqrt{n} \left[\left| \beta_j + \frac{u_j}{\sqrt{n}} \right| - |\beta_j| \right], \end{aligned} \tag{25}$$

where

$$Y_i = (1-\delta)(1(e_i < 0) - 1(e_i \geq 0)) - 2\delta e_i$$

and

$$Z_i = 2 \times \left(\frac{\mathbf{u}^* \mathbf{x}_i}{\sqrt{n}} - e_i \right) 1 \left(0 \leq e_i < \frac{\mathbf{u}^* \mathbf{x}_i}{\sqrt{n}} \right).$$

We have already seen that the first three terms together converge in distribution to

$$(\delta + (1-\delta) f(0)) \mathbf{u}^* C \mathbf{u} + \mathbf{u}^* \mathbf{W},$$

where $\mathbf{W} \sim N(\mathbf{0}, ((1-\delta)^2 + 4\delta^2\sigma^2 + 4\delta(1-\delta)M_{10})\, C)$. For the last term, we divide into two cases. If $\beta_j \neq 0$, then $\hat{w}_j \xrightarrow{P} |\beta_j|^{-\gamma}$ by the continuous mapping theorem (CMT). So it follows that

$$\underbrace{\frac{\lambda_n}{\sqrt{n}}}_{\to 0} \times \underbrace{\hat{w}_j}_{\xrightarrow{P} |\beta_j|^{-\gamma}} \times \underbrace{\sqrt{n} \left(\left| \beta_j + \frac{u_j}{\sqrt{n}} \right| - |\beta_j| \right)}_{\to u_j \mathrm{sgn}(\beta_j)} \xrightarrow{P} 0.$$

If $\beta_j = 0$, then $\sqrt{n} \left(\left| \beta_j + \frac{u_j}{\sqrt{n}} \right| - |\beta_j| \right) = |u_j|$ and

$$\frac{\lambda_n}{\sqrt{n}} \hat{w}_j = \underbrace{\lambda_n n^{\frac{\gamma-1}{2}}}_{\to \infty} \underbrace{\left| \sqrt{n} \hat{\beta}_{LS}[j] \right|^{-\gamma}}_{= O_p(1)} \xrightarrow{P} \infty.$$

Applying the Slutsky's lemma, we deduce that $V_n \Rightarrow V$ pointwise, where

$$V(\mathbf{u}) = \begin{cases} (\delta + (1-\delta) f(0)) \mathbf{u}^* C \mathbf{u} + \mathbf{u}^* \mathbf{W} & \text{if } u_j = 0 \forall j \notin A, \\ \infty & \text{otherwise.} \end{cases} \tag{26}$$

Since V_n is convex and V has unique minimum, it follows from the standard epi-convergence results ([60] and [77]) that $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) = \arg\min(V_n) \Rightarrow \arg\min(V)$. That is,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})[A] \Rightarrow C_{11}^{-1}\mathbf{W}_A \text{ and } \sqrt{n}\hat{\boldsymbol{\beta}}_n[A^c] \Rightarrow \mathbf{0},$$

where

$$\mathbf{W}_A = N\left(\mathbf{0}, \frac{(1-\delta)^2 + 4\delta^2\sigma^2 + 4\delta(1-\delta)M_{10}}{4[\delta + (1-\delta)f(0)]^2}C_{11}\right).$$

Then part 1 of the theorem follows and it is ready to see that $\arg\min(V_n)$ converges in probability to $\mathbf{0}$ on A^c.

For part 2, it suffices to show the following two cases:

1. For $j \in A$, the asymptotic normality proved in part 1) yeilds $\hat{\beta}_n[j] \xrightarrow{P} \beta_j$. Since $|\beta_j| > 0$, it follows that $P(j \in A_n) \to 1$.

2. For $j \in A^c$, we want to show that $P(j \in A_n) \to 0$. Observe that on the event $\{j \in A_n\}$, the first-order sub-differential optimality condition implies that

$$\begin{aligned} &\left| 2\delta \frac{\mathbf{x}^{j*}\left(\mathbf{y} - X\hat{\boldsymbol{\beta}}_n\right)}{\sqrt{n}} + (1-\delta)\sum_{i\in B}\frac{x_{ij}\mathrm{sgn}\left(y_i - \hat{\boldsymbol{\beta}}_n^*\mathbf{x}_i\right)}{\sqrt{n}} \right. \\ &\left. - \frac{\lambda_n \hat{w}_j \mathrm{sgn}\left(\hat{\beta}_n[j]\right)}{\sqrt{n}} \right| \leq (1-\delta)\sum_{i\notin B}\frac{|x_{ij}|}{\sqrt{n}}, \end{aligned} \tag{27}$$

   where $B = \{i \in \{1, \cdots, n\} : y_i - \hat{\boldsymbol{\beta}}_n^*\mathbf{x}_i \neq 0\}$. Since the RHS of (27) is bounded in probability while the LHS diverges in probability, it follows that the probability with which (27) holds vanishes as $n \to \infty$. Hence, it follows that $P(j \in A_n) \to 0$.

□

A.2.7 Proof of Theorem 2.4.1

Let $Z_n(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^n L(\mathbf{u}; y_i, \mathbf{x}_i) + \frac{\lambda_n}{n}\|\mathbf{u}\|_{\ell_1}$ and $\hat{\boldsymbol{\beta}}_n^H$ minimize Z_n. Anticipating $1/\sqrt{n}$ convergence rate so that we define V_n as

$$\begin{aligned} V_n(\mathbf{u}) = \sum_{i=1}^n \Big[&\left(e_i - \frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}}\right)^2 1\left(\left|e_i - \frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}}\right| \le \delta\right) \\ &- e_i^2 1\left(|e_i| \le \delta\right)\Big] \\ &+ \sum_{i=1}^n \Big[\left(2\delta\left|e_i - \frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}}\right| - \delta^2\right) 1\left(\left|e_i - \frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}}\right| > \delta\right) \\ &- (2\delta|e_i| - \delta^2)1(|e_i| > \delta)\Big] \\ &+ \lambda_n \sum_{j=1}^p \left[\left|\beta_j + \frac{u_j}{\sqrt{n}}\right| - |\beta_j|\right]. \end{aligned} \tag{28}$$

Then $\sqrt{n}(\hat{\boldsymbol{\beta}}_n^H - \boldsymbol{\beta})$ minimizes V_n. WLOG, by symmetry we can assume $\frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}} \ge 0$. We can decompose V_n as

$$V_n(\mathbf{u}) = \sum_{i=1}^n (S_i + T_i + Y_i + Z_i) + \lambda_n \sum_{j=1}^p \left[\left|\beta_j + \frac{u_j}{\sqrt{n}}\right| - |\beta_j|\right], \tag{29}$$

where

$$S_i = \frac{\mathbf{u}^*\mathbf{x}_i\mathbf{x}_i^*\mathbf{u}}{n} 1\left(\left|e_i - \frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}}\right| \le \delta\right),$$

$$T_i = (|e_i| - \delta)^2 \left(1\left(\left|e_i - \frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}}\right| \le \delta\right) - 1\left(|e_i| \le \delta\right)\right),$$

$$\begin{aligned} Y_i = \frac{2\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}} &\left[\delta(1(e_i < 0) - 1(e_i \ge 0))1\left(\left|e_i - \frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}}\right| > \delta\right)\right. \\ &\left. - e_i 1\left(\left|e_i - \frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}}\right| \le \delta\right)\right], \end{aligned} \tag{30}$$

and

$$\begin{aligned} Z_i &= 4\delta\left(\frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}} - e_i\right) 1\left(0 \le e_i < \frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}}\right) \\ &\times\ 1\left(\left|e_i - \frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}}\right| > \delta\right). \end{aligned}$$

First, we observe that

$$ES_i = \frac{\mathbf{u}^*\mathbf{x}_i\mathbf{x}_i^*\mathbf{u}}{n} P\left(\left|e_i - \frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}}\right| \le \delta\right).$$

Since f is continuous, the continuity of the probability measure P implies that

$$P\left(\left|e_i - \frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}}\right| \le \delta\right) \to K_{0\delta}.$$

So it follows from Lemma A.2.2 that

$$E\left(\sum_{i=1}^n S_i\right) \to K_{0\delta}\mathbf{u}^*C\mathbf{u}.$$

Since S_i is a Bernoulli r.v., we have

$$\begin{aligned} \mathrm{Var}\left(S_i\right) &= \left(\frac{\mathbf{u}^*\mathbf{x}_i\mathbf{x}_i^*\mathbf{u}}{n}\right)^2 P\left(\left|e_i - \frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}}\right| \le \delta\right) \\ &\times\ P\left(\left|e_i - \frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}}\right| > \delta\right). \end{aligned}$$

So we deduce that

$$\sum_{i=1}^n \mathrm{Var}\left(S_i\right) \le \sum_{i=1}^n \frac{\mathbf{u}^*\mathbf{x}_i\mathbf{x}_i^*\mathbf{u}}{n}\left(\frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}}\right)^2 \to 0,$$

whereas Chebyshev's inequality implies that

$$\sum_{i=1}^n S_i \xrightarrow{P} K_{0\delta}\mathbf{u}^*C\mathbf{u}. \tag{31}$$

Next, we show that the second term T_i stochastically vanishes, i.e.

$$\sum_{i=1}^{n} T_i \xrightarrow{P} 0. \tag{32}$$

To see this, it is straightforward to check that $ET_i = o\left(n^{-\frac{3}{2}}\right)$ so $E\sum_{i=1}^{n} T_i = o\left(n^{-\frac{1}{2}}\right)$. Furthermore, by the symmetry of e_i we can express ET_i^2 as

$$\begin{aligned} ET_i^2 &= E\left((e_i-\delta)^4; \delta - \frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}} \le e_i \le \delta + \frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}}\right) \\ &= \frac{2}{5} f(\delta)\left(\frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}}\right)^5 + o\left(n^{-\frac{5}{2}}\right) \end{aligned}$$

such that

$$\operatorname{Var}(T_i) = \frac{2}{5} f(\delta)\left(\frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}}\right)^5 + o(n^{-\frac{5}{2}}),$$

whereby Lemma A.2.1 implies that $\sum_{i=1}^{n} \operatorname{Var}(T_i) \to 0$. So it follows that $\sum_{i=1}^{n} T_i = o_p(1)$. The last two terms converge as in the proof of Theorem 2.2.1. Specifically, a routine calculation shows that

$$\begin{aligned} EY_i = {} & \frac{2\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}}\left[\delta P\left(\delta - \frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}} < e_i < \delta + \frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}}\right)\right. \\ & \left. - E\left(e_i; \delta - \frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}} < e_i < \delta + \frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}}\right)\right] \\ = {} & \frac{2\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}}\left[2\delta f(\delta)\frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}} + o\left(n^{-1/2}\right) - 2\delta f(\delta)\frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}}\right] \\ = {} & o\left(n^{-1}\right), \end{aligned} \tag{33}$$

which implies that $E\sum_{i=1}^n Y_i = o(1)$. Note that the two components of Y_i are orthogonal and it can be easily shown that

$$\begin{aligned}\operatorname{Var}\left(\sum_{i=1}^n Y_i\right) &= \sum_{i=1}^n \Big[\frac{4\mathbf{u}^*\mathbf{x}_i\mathbf{x}_i^*\mathbf{u}}{n}\Big(\delta^2 M_{0\delta} + K_{2\delta} \\ &\quad + 2\delta f(\delta)\left(\frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}}\right)^2 + o(n^{-1/2})\Big)\Big] \\ &\to \left(\delta^2 M_{0\delta} + K_{2\delta}\right) C.\end{aligned} \tag{34}$$

So the CLT implies that

$$\sum_{i=1}^n Y_i \Rightarrow 2\mathbf{u}^*\mathbf{W}, \tag{35}$$

where $\mathbf{W} \sim N\left(\mathbf{0}, (\delta^2 M_{0\delta} + K_{2\delta})C\right)$. Finally, note that $Z_i = 0$ for large n, so we can deduce that

$$\sum_{i=1}^n Z_i \xrightarrow{P} 0. \tag{36}$$

Combining (31), (32), (35), and (36) together and applying Slutsky's lemma, we deduce that $V_n(\mathbf{u}) \Rightarrow V(\mathbf{u})$ where

$$\begin{aligned} V(\mathbf{u}) &= K_{0\delta}\mathbf{u}^*C\mathbf{u} + 2\mathbf{u}^*\mathbf{W} \\ &+ \lambda_0 \sum_{j=1}^p [u_j \operatorname{sgn}(\beta_j)\, 1(\beta_j \neq 0) + |u_j| 1(\beta_j = 0)]. \end{aligned}$$

Since V_n is convex and V has a unique minimum, it follows from [60] that $\arg\min(V_n) = \sqrt{n}(\hat{\boldsymbol{\beta}}_n^H - \boldsymbol{\beta}) \Rightarrow \arg\min(V)$. □

A.2.8 Proof of Theorem 2.5.1

The proof is essentially similar to that of Theorem 2.2.1, with additional complexity from the extra randomness of X. We use the same notation as in the proof of Theorem 2.2.1 unless otherwise indicated. Let V_n be defined as in (8). Recall that

$$Y_i = (1-\delta)(1(e_i < 0) - 1(e_i \geq 0)) - 2\delta e_i,$$

$$Z_i = 2 \times \left(\frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}} - e_i\right) 1\left(0 \leq e_i < \frac{\mathbf{u}^*\mathbf{x}_i}{\sqrt{n}}\right),$$

and

$$\begin{aligned} V_n(\mathbf{u}) &= \delta \sum_{i=1}^{n} \left(\mathbf{u}^* \frac{\mathbf{x}_i\mathbf{x}_i^*}{n}\mathbf{u}\right) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{u}^*\mathbf{x}_i Y_i \\ &+ (1-\delta) \sum_{i=1}^{n} Z_i + \lambda_n \sum_{j=1}^{p} \left[\left|\beta_j + \frac{u_j}{\sqrt{n}}\right| - |\beta_j|\right]. \end{aligned}$$

First, we observe that $\{\mathbf{u}^*\mathbf{x}_i Y_i\}_{i\in\mathbb{N}}$ forms a martingale difference sequence since

$$E\left(\mathbf{u}^*\mathbf{x}_i Y_i \mid \mathcal{F}_{i-1}\right) = \mathbf{u}^*\mathbf{x}_i E\left(Y_i\right) = 0,$$

where the second equality follows from the hypotheses that $\mathbf{x}_i \in \mathcal{F}_{i-1}$ and $\sigma(e_i)$ is independent of $\mathcal{F}_{i-1}$. Put $T_n = \sum_{i=1}^{n} E\left((\mathbf{u}^*\mathbf{x}_i Y_i)^2 \mid \mathcal{F}_{i-1}\right)$. A simple calculation shows that

$$\begin{aligned} \frac{T_n}{n} &= ((1-\delta)^2 + 4\delta^2\sigma^2 + 4\delta(1-\delta)M_{10}) \\ &\times \mathbf{u}^* \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^*\right) \mathbf{u} \\ &\xrightarrow{P} ((1-\delta)^2 + 4\delta^2\sigma^2 + 4\delta(1-\delta)M_{10})\mathbf{u}^*C\mathbf{u}. \end{aligned}$$

Moreover, consider

$$\begin{aligned} & n^{-1}\sum_{i=1}^{n} E\left(\mathbf{u}^*\mathbf{x}_i\mathbf{x}_i^*\mathbf{u}Y_i^2\right) \\ =\ & E\left[\mathbf{u}^* \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^*\right) \mathbf{u}E\left(Y_i^2 | \mathcal{F}_{i-1}\right)\right] \\ =\ & E\left(Y_i^2\right) E\left[\mathbf{u}^* \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^*\right) \mathbf{u}\right] \\ \to\ & ((1-\delta)^2 + 4\delta^2\sigma^2 + 4\delta(1-\delta)M_{10})\mathbf{u}^*C\mathbf{u}, \end{aligned}$$

where the last conclusion follows from the dominated convergence theorem. So it follows that $\sum_{i=1}^n n^{-1}\left(\mathbf{u}^*\mathbf{x}_iY_i\right)^2 \in L^1$. By Lemma A.2.3, we have

$$1\left(|\mathbf{u}^*\mathbf{x}_iY_i| > \epsilon\sqrt{n}\right) = 1\left(\left|\frac{\mathbf{u}^*\mathbf{x}_iY_i}{\sqrt{n}}\right| > \epsilon\right) \downarrow 0, \quad P\text{-a.s.}$$

Again by the dominated convergence and Lemma A.2.1, we deduce that

$$\frac{1}{n}\sum_{i=1}^n E\left((\mathbf{u}^*\mathbf{x}_iY_i)^2; |\mathbf{u}^*\mathbf{x}_iY_i| > \sqrt{n}\epsilon\right) \to 0$$

as $n \to \infty$. Now the martingale central limit theorem (p414, [44]) implies that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{u}^*\mathbf{x}_iY_i \Rightarrow N(\mathbf{0}, ((1-\delta)^2 + 4\delta^2\sigma^2 + 4\delta(1-\delta)M_{10})\mathbf{u}^*C\mathbf{u}).$$

Secondly, the term involving Z_i is similar to the non-stochastic design case. That is, we need to show that $\sum_{i=1}^n Z_i \xrightarrow{P} f(0)\mathbf{u}^*C\mathbf{u}$, then the rest of the proof proceeds as in Theorem 2.2.1. Observe that by Lemma A.2.5 and the measureability assumption on $\mathbf{x}_i$, $\sum_{i=1}^n E\left(Z_i|\mathcal{F}_{i-1}\right) = f(0)\sum_{i=1}^n \mathbf{u}^*\frac{\mathbf{x}_i\mathbf{x}_i^*}{n}\mathbf{u} \xrightarrow{P} f(0)\mathbf{u}^*C\mathbf{u}$. So it follows from the Skorokhod representation (p83-84, [44]) that

$$\sum_{i=1}^n EZ_i = E\left(\sum_{i=1}^n E\left(Z_i|\mathcal{F}_{i-1}\right)\right) \to f(0)\mathbf{u}^*C\mathbf{u}.$$

Similarly, we can show that $\sum_{i=1}^n \mathrm{Var}\left(Z_i\right) \to 0$. This completes the proof. □

A.2.9 Proof of Auxiliary Lemmas

The following two useful lemmas (Lemma A.2.1 and A.2.2) concerning the convergence of two real sequences will be used repeatedly in the proofs of the theorems.

Lemma A.2.1. *Let $0 \le c_1 < c_2 < \cdots$ and $0 \le d_1 < d_2 < \cdots$ with $c_n \to \infty$ and $d_n \to \infty$. Let $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ be two real sequences such that $a_n \ge 0$, $\frac{1}{c_n}\sum_{i=1}^n a_i \to a \ge 0$, and $\frac{1}{d_n}\max_{1\le i\le n}|b_i| \to 0$. Then $\frac{1}{c_nd_n}\sum_{i=1}^n a_ib_i \to 0$.*

Proof. Fix an $\epsilon > 0$. Since $\frac{1}{d_n}\max_{1\le i\le n}|b_i| \to 0$, there is an $N = N(\epsilon)$ such that

$n \geq N$ implies that $\frac{|b_i|}{d_n} \leq \epsilon$ for all $i \in \{1, \cdots, n\}$. But then

$$\begin{aligned} \left| \frac{1}{c_n d_n} \sum_{i=1}^{n} a_i b_i \right| &\leq \frac{1}{c_n d_n} \sum_{i=1}^{N-1} |a_i b_i| + \sum_{i=N}^{n} \left| \frac{a_i}{c_n} \right| \left| \frac{b_i}{d_n} \right| \\ &\leq \epsilon + \left(\sum_{i=1}^{n} \frac{a_i}{c_n} \right) \epsilon \\ &\leq \epsilon + (a + \epsilon)\epsilon \\ &= (a+1)\epsilon + \epsilon^2 \end{aligned}$$

for sufficiently large n. Since ϵ is arbitrary, it follows that $\frac{1}{c_n d_n} \sum_{i=1}^{n} a_i b_i \to 0$ as $n \to \infty$. □

Lemma A.2.2. *Let $0 \leq c_1 < c_2 < \cdots$ with $c_n \to \infty$. Let $\{a_n\}_{n \in \mathbb{N}}$ and $\{b_n\}_{n \in \mathbb{N}}$ be two real sequences such that $a_n \geq 0$, $\frac{1}{c_n} \sum_{i=1}^{n} a_i \to a \geq 0$ and $b_n \to b$. Then $\frac{1}{c_n} \sum_{i=1}^{n} a_i b_i \to ab$.*

Proof. Fix an $\epsilon > 0$. Since $b_n \to b$, there is an $N = N(\epsilon)$ such that $n \geq N$ implies that $|b_n - b| \leq \epsilon$. Consider

$$\begin{aligned} &\left| \frac{1}{c_n} \sum_{i=1}^{n} a_i b_i - ab \right| \\ = &\left| \sum_{i=1}^{n} \frac{a_i}{c_n} b_i - \sum_{i=1}^{n} \frac{a_i}{c_n} b + \sum_{i=1}^{n} \frac{a_i}{c_n} b - ab \right| \\ \leq &\sum_{i=1}^{n} \left| \frac{a_i}{c_n}(b_i - b) \right| + \left| \sum_{i=1}^{n} \frac{a_i}{c_n} - a \right| b \\ \leq &\sum_{i=1}^{N-1} \left| \frac{a_i}{c_n}(b_i - b) \right| + \sum_{i=N}^{n} \frac{a_i}{c_n} |b_i - b| + b\epsilon \\ \leq &\epsilon + \left(\sum_{i=1}^{n} \frac{a_i}{c_n} \right) \epsilon + b\epsilon \\ \leq &\epsilon + (a + \epsilon)\epsilon + b\epsilon \\ = &(a + b + 1)\epsilon + \epsilon^2 \end{aligned}$$

for sufficiently large n. Since ϵ is arbitrary, the lemma follows. □

Lemma A.2.3. *Under assumption 1, we have*

$$\frac{1}{n} \max_{1 \leq i \leq n} \mathbf{x}_i^* \mathbf{x}_i \to 0$$

and

$$\max_{1 \leq i \leq n} \mathbf{x}_i^* \left(\sum_{i=1}^{n} \mathbf{x}_i^* \mathbf{x}_i \right)^{-1} \mathbf{x}_i \to 0$$

as $n \to \infty$.

Proof. Let $\epsilon > 0$ and $T = \operatorname{tr}(C)$. Since $C_n \to C$ and C is positive definite, we have $\operatorname{tr}(C_n) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^* \mathbf{x}_i \to T > 0$. So there is an $N = N(\epsilon)$ such that

$$\left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^* \mathbf{x}_i - T \right| \leq \epsilon$$

for all $n \geq N$. But then

$$\begin{aligned}
0 &\leq \frac{1}{n} \max_{N \leq i \leq n} \mathbf{x}_i^* \mathbf{x}_i \\
&= \frac{1}{n} \max_{N \leq i \leq n} \left(\sum_{j=1}^{i} \mathbf{x}_j^* \mathbf{x}_j - \sum_{j=1}^{i-1} \mathbf{x}_j^* \mathbf{x}_j \right) \\
&\leq \frac{1}{n} \max_{N \leq i \leq n} \left((T + \epsilon) i - (T - \epsilon)(i - 1) \right) \\
&\leq \max_{N \leq i \leq n} \left(\frac{T + 2i\epsilon}{n} \right) \leq 3\epsilon
\end{aligned}$$

for sufficiently large n. Letting $n \to \infty$, the first part of the lemma follows since

$$0 \leq \limsup_{n \to \infty} \frac{1}{n} \max_{1 \leq i \leq n} \mathbf{x}_i^* \mathbf{x}_i \leq 3\epsilon,$$

whereas $\epsilon > 0$ is arbitrary. The second claim is an easy consequence of the first part by the finite-dimensionality of $\mathbf{x}_i$'s. □

Lemma A.2.4. *Suppose $X_1, X_2, \cdots$ is a sequence of i.i.d. r.v.'s with a continuous, positive probability density function f around 0 (w.r.t. the Lebesgue measure). Let F be the common cumulative distribution function with $F(0) = \frac{1}{2}$. Let $g_n(u) =$*

$\sum_{i=1}^n \left(\left| X_i - \frac{u}{\sqrt{n}} \right| - |X_i| \right)$. *Then* $g_n(u) \Rightarrow g(u)$ *where* $g(u) = uZ + f(0)u^2$ *with* $Z \sim N(0,1)$.

Proof. First assume $u \geq 0$ and rewrite g_n as following

$$g_n(u) = \frac{u}{\sqrt{n}} \sum_{i=1}^n Y_i + \sum_{i=1}^n Z_i, \tag{37}$$

where

$$Y_i = 1(X_i < 0) - 1(X_i \geq 0)$$

and

$$Z_i = 2 \times \left(\frac{u}{\sqrt{n}} - X_i \right) 1 \left(0 \leq X_i < \frac{u}{\sqrt{n}} \right).$$

Since $F(0) = 1/2$, i.e. X_i has median 0, we have $EY_i = 0$ and $\mathrm{Var}\,(Y_i) = EY_i^2 = 1$. Hence the central limit theorem (CLT) implies that

$$\frac{\sum_{i=1}^n Y_i}{\sqrt{n}} \Rightarrow N(0,1).$$

Now consider the second term concerning Z_i. First observe that for $|u'|$ small, by the continuity assumption of f about 0, we have

$$\begin{aligned}
P(0 \leq X_i < u') &= \int_0^{u'} f(x_i)\, dx_i \\
&= \int_0^{u'} (f(0) + o(1))\, dx_i \\
&= f(0)u' + o(u'), \\
E(X_i; 0 \leq X_i < u') &= \int_0^{u'} x_i(f(0) + o(1))\, dx_i \\
&= \frac{1}{2} f(0)u'^2 + o(u'^2), \\
E(X_i^2; 0 \leq X_i < u') &= \int_0^{u'} x_i^2(f(0) + o(1))\, dx_i \\
&= \frac{1}{3} f(0)u'^3 + o(u'^3).
\end{aligned}$$

Applying $u' = u/\sqrt{n}$, we deduce that

$$\begin{aligned} EZ_i &= E\left(2\left(\frac{u}{\sqrt{n}} - X_i\right); 0 \leq X_i < \frac{u}{\sqrt{n}}\right) \\ &= \frac{2u}{\sqrt{n}} P\left(0 \leq X_i < \frac{u}{\sqrt{n}}\right) - 2E\left(X_i; 0 \leq X_i < \frac{u}{\sqrt{n}}\right) \\ &= \frac{2u}{\sqrt{n}}\left(f(0)\frac{u}{\sqrt{n}} + o\left(\frac{u}{\sqrt{n}}\right)\right) - f(0)\frac{u^2}{n} + o\left(\left(\frac{u}{\sqrt{n}}\right)^2\right) \\ &= \frac{f(0)}{n}u^2 + o\left(\frac{u^2}{n}\right). \end{aligned} \tag{38}$$

So $\sum_{i=1}^n EZ_i = f(0)u^2 + o(1)$. It can be similarly shown that

$$\mathrm{Var}\,(Z_i) = \frac{4f(0)}{3}\left(\frac{u}{\sqrt{n}}\right)^3 + o\left(\left(\frac{u}{\sqrt{n}}\right)^3\right)$$

so that $\sum_{i=1}^n \mathrm{Var}\,(Z_i) = O(n^{-1/2})$. So it follows from Chebyshev's inequality that

$$P\left(\left|\sum_{i=1}^n Z_i - \sum_{i=1}^n EZ_i\right| > \epsilon\right) \leq \frac{\sum_{i=1}^n \mathrm{Var}\,(Z_i)}{\epsilon^2} \to 0,$$

i.e.

$$\sum_{i=1}^n Z_i \xrightarrow{P} f(0)u^2.$$

Invoking Slutsky's lemma, we obtain $g_n \Rightarrow g$ pointwise. The case for $u < 0$ follows the similar lines. □

Lemma A.2.5. *Let $X_1, X_2, \cdots$ be an i.i.d. sequence of r.v.'s with a common continuous, positive p.d.f. f around 0 and the median of X_i equal 0 for all i. Let $\mathbf{u}, \mathbf{v}_i \in \mathbb{R}^p$ and $g_n(\mathbf{u}) = \sum_{i=1}^n \left(\left|X_i - \frac{\mathbf{u}^*\mathbf{v}_i}{\sqrt{n}}\right| - |X_i|\right)$. Then $g_n(\mathbf{u}) \Rightarrow g(\mathbf{u})$ where $g(\mathbf{u}) = \mathbf{u}^*\mathbf{W} + f(0)\mathbf{u}^*C\mathbf{u}$ with $\mathbf{W} \sim N(\mathbf{0}, C)$ and $C = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mathbf{v}_i\mathbf{v}_i^*$.*

Proof. The proof is an easy consequence of Lemma A.2.4 and we use the same notation unless explicitly indicated. Without loss of generality, we can assume that $\mathbf{u}^*\mathbf{v}_i \geq 0$ for all i by symmetry. As in the proof of Lemma A.2.4, we have $EZ_i = \frac{f(0)}{n}(\mathbf{u}^*\mathbf{v}_i\mathbf{v}_i^*\mathbf{u}) +$

$o\left(\frac{\mathbf{u}^*\mathbf{v}_i\mathbf{v}_i^*\mathbf{u}}{n}\right)$. So

$$\sum_{i=1}^{n} EZ_i = f(0)\mathbf{u}^* \frac{\sum_{i=1}^{n} \mathbf{v}_i\mathbf{v}_i^*}{n}\mathbf{u} + o(1) \to f(0)\mathbf{u}^* C\mathbf{u}.$$

Moreover, observe that

$$\sum_{i=1}^{n} \mathrm{Var}\,(Z_i) = \frac{4f(0)}{3}\left(\sum_{i=1}^{n} \frac{\mathbf{u}^*\mathbf{v}_i\mathbf{v}_i^*\mathbf{u}}{n}\frac{\mathbf{u}^*\mathbf{v}_i}{\sqrt{n}}\right) + o(\frac{1}{\sqrt{n}}).$$

Applying Lemma A.2.1 with $a_i = \mathbf{u}^*\mathbf{v}_i\mathbf{v}_i^*\mathbf{u}$, $b_i = \mathbf{u}^*\mathbf{v}_i$, $c_n = n$, and $d_n = \sqrt{n}$, it then follows from Lemma A.2.3 that $\sum_{i=1}^{n} \mathrm{Var}\,(Z_i) \to 0$ as $n \to \infty$. Then Chebyshev's inequality implies that $\sum_{i=1}^{n} Z_i \xrightarrow{P} f(0)\mathbf{u}^* C\mathbf{u}$. Let $\tilde{Y}_i = (\mathbf{u}^*\mathbf{v}_i)Y_i$. The CLT and Slutsky's lemma together imply that

$$\frac{\sum_{i=1}^{n} \tilde{Y}_i}{\sqrt{n}} \Rightarrow N(\mathbf{0}, \mathbf{u}^* C\mathbf{u}) = \mathbf{u}^*\mathbf{W},$$

where $\mathbf{W} \sim N(\mathbf{0}, C)$. The conclusion follows by invoking Slutsky's lemma once again. □

A.3 Appendix for Chapter 3

A.3.1 Derivation of the Joint Posterior Distribution in (3.3)

Denote the likelihood of the model by $L(\mathbf{y}|X,\cdot)$. We can factorize the joint posterior distribution according to the conditional independence relationships encoded in the hierarchical model as:

$$p(\boldsymbol{\beta}, \tau, \sigma^2, \boldsymbol{\gamma}, \lambda|\mathbf{y}, X)$$

$$\begin{aligned}
&\propto \; p(\boldsymbol{\beta}, \tau, \sigma^2, \boldsymbol{\gamma}, \lambda) \times L(\mathbf{y}|X, \boldsymbol{\beta}, \tau, \sigma^2, \boldsymbol{\gamma}, \lambda) \\
&\propto \; \frac{1}{\binom{p}{k}} p(\lambda) p(\tau, \sigma^2) p(k|\lambda) \prod_{j \in \boldsymbol{\gamma}} p(\beta_j|\boldsymbol{\gamma}, \tau) \times (\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{||\mathbf{y} - X\boldsymbol{\beta}||_2^2}{2\sigma^2}\right) \\
&\propto \; \frac{1}{\binom{p}{k}} (\lambda\tau\sigma)^{-1} \frac{e^{-\lambda}\lambda^k}{k!} \tau^{-k} \exp\left(-\frac{\sum_{j\in\boldsymbol{\gamma}} |\beta_j|}{\tau}\right) \times \sigma^{-n} \exp\left(-\frac{||\mathbf{y} - X\boldsymbol{\beta}||_2^2}{2\sigma^2}\right),
\end{aligned}$$

which gives (3.3).

A.3.2 MCMC Algorithm for the Binomial-Gaussian Model

We apply the proposed fully Bayesian framework to the Binomial-Gaussian model proposed in [59]. More specifically, let $\boldsymbol{\gamma}$ be a p-length binary vector representing an active set. The active set is assumed to be binomially distributed

$$p(\boldsymbol{\gamma}|w) = w^{q_{\boldsymbol{\gamma}}} (1-w)^{p-q_{\boldsymbol{\gamma}}} \tag{39}$$

where w is the probability of taking value one and $q_{\boldsymbol{\gamma}}$ means the number of ones in $\boldsymbol{\gamma}$. Conditioning on $\boldsymbol{\gamma}$, Zellner's g-prior is used for the coefficient:

$$p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}, g) = N\left(\mathbf{0}, g\sigma^2 (X_{\boldsymbol{\gamma}}^T X_{\boldsymbol{\gamma}})^{-1}\right). \tag{40}$$

Then, assigning a conjugate beta prior on w and Jeffrey's non-informative priors on σ^2 and g, we have

$$\begin{aligned}
p(w) &= \text{Beta}(a, b), \\
p(\sigma^2, g) &\propto (\sigma^2 g)^{-1}.
\end{aligned}$$

Therefore we can analytically integrate out $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and obtain

$$\begin{aligned} p(\mathbf{y}, g, \sigma^2, w, \boldsymbol{\gamma}) &\propto p(w)p(g)p(\sigma^2)p(\boldsymbol{\gamma}|w)p(\mathbf{y}|g, \sigma^2, \boldsymbol{\gamma}) \\ &\propto w^{a-1}(1-w)^{b-1}(\sigma^2)^{-1}g^{-1}w^{q_{\boldsymbol{\gamma}}}(1-w)^{p-q_{\boldsymbol{\gamma}}} \\ &\times (2\pi\sigma^2)^{-n/2}(1+g)^{-q_{\boldsymbol{\gamma}}/2}\exp\left(-\frac{\mathbf{y}^T\mathbf{y}}{2\sigma^2} + \frac{g}{2\sigma^2(1+g)}R_{\boldsymbol{\gamma}}\right), \end{aligned} \tag{41}$$

where $R_{\boldsymbol{\gamma}} = \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}^T X_{\boldsymbol{\gamma}}^T X_{\boldsymbol{\gamma}} \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ is least squares estimate of model $\boldsymbol{\gamma}$. By further integrating out σ^2 and w, we obtain

$$p(\mathbf{y}, g, \boldsymbol{\gamma}) \propto B(a+q_{\boldsymbol{\gamma}}, b+p-q_{\boldsymbol{\gamma}})g^{-1}(1+g)^{-q_{\boldsymbol{\gamma}}/2}\left(-\frac{\mathbf{y}^T\mathbf{y}}{2} + \frac{g}{2(1+g)}R_{\boldsymbol{\gamma}}\right)^{-n/2} \tag{42}$$

where $B(a+q_{\boldsymbol{\gamma}}, b+p-q_{\boldsymbol{\gamma}}) = \frac{\Gamma(a+q_{\boldsymbol{\gamma}})\Gamma(b+p-q_{\boldsymbol{\gamma}})}{\Gamma(a+b-p)}$.

Now, it follows that the full conditional probabilities

$$p(\boldsymbol{\gamma}|\mathbf{y}, g) \propto B(a+q_{\boldsymbol{\gamma}}, b+p-q_{\boldsymbol{\gamma}})(1+g)^{-q_{\boldsymbol{\gamma}}/2}\left(-\frac{\mathbf{y}^T\mathbf{y}}{2} + \frac{g}{2(1+g)}R_{\boldsymbol{\gamma}}\right)^{-n/2}, \tag{43}$$

$$p(g|\mathbf{y}, \boldsymbol{\gamma}) \propto g^{-1}(1+g)^{-q_{\boldsymbol{\gamma}}/2}\left(-\frac{\mathbf{y}^T\mathbf{y}}{2} + \frac{g}{2(1+g)}R_{\boldsymbol{\gamma}}\right)^{-n/2}\mathbb{I}(g \geq 0). \tag{44}$$

Thus the updating probability from $\boldsymbol{\gamma} \to \boldsymbol{\gamma}'$ is the MH ratio

$$\min\left\{\frac{p(\boldsymbol{\gamma}'|\mathbf{y}, g)}{p(\boldsymbol{\gamma}|\mathbf{y}, g)} \times \frac{p(\boldsymbol{\gamma}' \to \boldsymbol{\gamma})}{p(\boldsymbol{\gamma} \to \boldsymbol{\gamma}')}, 1\right\}. \tag{45}$$

We note that the above BG-MCMC algorithm can be interpreted as an RJ-MCMC approach with the birth-and-death proposal. Here since the Jacobian for the birth-and-death proposal is 1 for model jumping, there is no explicit Jacobian term reflected in the generalized MH ratio and thus it is coincides with the ordinary MH ratio with $\boldsymbol{\gamma}$ being considered as a regular parameter.

Regarding the parameter g, we employ a symmetric random walk sampler for updating. More specifically, we let $g' = g + u$ where $u \sim N(0, \epsilon^2)$ and accept with probability $\min\left\{\frac{p(g'|\mathbf{y},\boldsymbol{\gamma})}{p(g|\mathbf{y},\boldsymbol{\gamma})}, 1\right\}$.

The MCMC algorithm for the BG model is summarized in Algorithm 4.

Algorithm 4: BG-MCMC algorithm

```
Input: The number of iterations $T$. Random walk step size $\epsilon$.
Data: $X$ and $\mathbf{y}$.
Output: $\left\{\left\{\boldsymbol{\gamma}^{(t)}, g^{(t)}\right\} \middle| t \in \{0, \cdots, T\}\right\}$.
begin
    Initialization: choose $\left\{\boldsymbol{\gamma}^{(0)}, g^{(0)}\right\}$ and $t = 1$.
    repeat
        if $k^{(t-1)} = 1$ then
            $k^{(t)} \leftarrow k^{(t-1)} + 1$.
        else if $k^{(t-1)} = p$ then
            $k^{(t)} \leftarrow k^{(t-1)} - 1$.
        else
            $k^{(t)} \leftarrow k^{(t-1)} + U(\{-1, 1\})$.
        end
        if $k^{(t)} = k^{(t-1)} + 1$ then
            Propose a $\boldsymbol{\gamma}'$ such that $|\boldsymbol{\gamma}'| = |\boldsymbol{\gamma}| + 1$ (birth move);
            Accept $\boldsymbol{\gamma}'$ with the probability given by MH ratio.
        else
            Propose a $\boldsymbol{\gamma}'$ such that $|\boldsymbol{\gamma}'| = |\boldsymbol{\gamma}| - 1$ (death move);
            Accept $\boldsymbol{\gamma}'$ with the probability given by MH ratio.
        end
        Sample $g$ with the Gaussian random walk with step size $\epsilon$. $t \leftarrow t + 1$.
    until $t = T$.
end
```

A.4 Proofs for Chapter 4

A.4.1 Proof of Theorem 4.2.2

By definition, LHS of (4.11) can be written as

$$\begin{aligned} E\|\hat{\Sigma}_o - \Sigma\|_F^2 &= E\{\mathrm{Tr}[(\hat{\Sigma}_o - \Sigma)^2]\} \\ &= \mathrm{Tr}(E\hat{\Sigma}_o^2) - \mathrm{Tr}[(E\hat{\Sigma}_o)\Sigma] - \mathrm{Tr}[\Sigma(E\hat{\Sigma}_o)] + \mathrm{Tr}(\Sigma^2) \\ &= \mathrm{Tr}(E\hat{\Sigma}_o^2) - 2\left[(1-\rho_o)\mathrm{Tr}(\Sigma^2) + \rho_o \frac{\mathrm{Tr}^2(\Sigma)}{p}\right] + \mathrm{Tr}(\Sigma^2). \end{aligned}$$

We analyze the first term in the last equation and find that it equals

$$I = (1-\rho_o)^2 E[\mathrm{Tr}(\hat{S}^2)] + p^{-1}\rho_o^2 E[\mathrm{Tr}^2(\hat{S})] + 2p^{-1}(1-\rho_o)\rho_o E[\mathrm{Tr}^2(\hat{S})].$$

Let

$$A = E\left[\mathrm{Tr}(\hat{S}^2) - \frac{\mathrm{Tr}(\hat{S})}{p}\right] \qquad \text{and} \qquad B = \mathrm{Tr}(\Sigma^2) - \frac{\mathrm{Tr}^2(\Sigma)}{p}.$$

Note that for any square matrix M, we have $\mathrm{Tr}(M^2) \geq \mathrm{Tr}^2(M)/p$; therefore $A \geq 0$ and $B \geq 0$. Expanding further by substituting $\hat{\rho}_o$ in (4.8) into I, we can obtain that

$$I = B(1 - \frac{B}{A}) + E\left[\frac{\mathrm{Tr}^2(\hat{S})}{p} - \frac{\mathrm{Tr}^2(\Sigma)}{p}\right].$$

Now, by the Gaussian assumption, we have

$$\begin{aligned} E\mathrm{Tr}(\hat{S}^2) &= \frac{n+1}{n}\mathrm{Tr}(\Sigma^2) + \frac{1}{n}\mathrm{Tr}^2(\Sigma) \\ E\mathrm{Tr}^2(\hat{S}) &= \mathrm{Tr}^2(\Sigma) + \frac{2}{n}\mathrm{Tr}(\Sigma^2). \end{aligned}$$

Plugging this expression into I, we see the theorem follows. □

A.4.2 Proof of Theorem 4.2.3

By the definition of the matrix spectral norm and noting that $E\hat{S} = \Sigma$ and $E\mathrm{Tr}(\hat{S}) = \mathrm{Tr}(\Sigma) = p$, we can obtain the following chain inequalities

$$\begin{aligned}
&E\|\hat{\Sigma}_o - \Sigma\|^2 = E\left(\sup_{\|\mathbf{x}\|_2=1} \mathbf{x}^T(\hat{\Sigma}_o - \Sigma)\mathbf{x}\right)^2 \geq \left(E\sup_{\|\mathbf{x}\|_2=1} \mathbf{x}^T(\hat{\Sigma}_o - \Sigma)\mathbf{x}\right)^2 \\
&\geq \sup_{\|\mathbf{x}\|_2=1} (E\mathbf{x}^T(\hat{\Sigma}_o - \Sigma)\mathbf{x})^2 = \sup_{\|\mathbf{x}\|_2=1} \left(\mathbf{x}^T(E\hat{\Sigma}_o - \Sigma)\mathbf{x}\right)^2 \\
&= \sup_{\|\mathbf{x}\|_2=1} \left(\mathbf{x}^T((1-\rho_o)E\hat{S} + \rho_o E\hat{F} - \Sigma)\mathbf{x}\right)^2 = \sup_{\|\mathbf{x}\|_2=1} \left(\mathbf{x}^T(p^{-1}\rho_o \mathrm{Tr}(\Sigma)I - \rho_o\Sigma)\mathbf{x}\right)^2 \\
&= \left(\rho_o - \rho_o \inf_{\|\mathbf{x}\|_2=1} \mathbf{x}^T\Sigma\mathbf{x}\right)^2 = \rho_o^2(1 - \lambda_{\min}(\Sigma))^2.
\end{aligned}$$

Here we used Jensen's inequality twice at the second and third steps. □

A.4.3 Proof of Theorem 4.3.1

Let $\rho \in [0,1]$ and note that

$$E\mathrm{Tr}(\hat{\Sigma}\Sigma) = (1-\rho)\mathrm{Tr}(\Sigma^2) + \rho\mathrm{Tr}((W \circ \Sigma)\Sigma) = (1-\rho)\|\Sigma\|_F^2 + \rho\|V \circ \Sigma\|_F^2, \tag{46}$$

where $V = (v_{ij})$ with $v_{ij} = \sqrt{w_{ij}}$. Moreover, since

$$\mathrm{Tr}(\hat{\Sigma}^2) = (1-\rho)^2\mathrm{Tr}(\hat{S}^2) + 2\rho(1-\rho)\mathrm{Tr}((W \circ \hat{S})\hat{S}) + \rho^2\mathrm{Tr}((W \circ \hat{S})^2),$$

taking expectation on both sides, we can obtain that

$$E\mathrm{Tr}(\hat{\Sigma}^2) = (1-\rho)^2 E\|\hat{S}\|_F^2 + 2\rho(1-\rho)E\|V \circ \hat{S}\|_F^2 + \rho^2 E\|W \circ \hat{S}\|_F^2. \tag{47}$$

To calculate $\hat{\rho}^{\mathrm{STO}}$, we need to find the minimizer of $E[\mathrm{Tr}(\hat{\Sigma}^2)] - 2\mathrm{Tr}((E\hat{\Sigma})\Sigma)$. Expanding this using (46) and (47), we get

$$E\left[\mathrm{Tr}(\hat{\Sigma}^2) - 2\mathrm{Tr}((E\hat{\Sigma})\Sigma)\right] = (1-\rho)^2 E\|\hat{S}\|_F^2 + 2\rho(1-\rho)E\|V\circ\hat{S}\|_F^2 + \rho^2 E\|W\circ\hat{S}\|_F^2 \\ - 2(1-\rho)\|\Sigma\|_F^2 - 2\rho\|V\circ\Sigma\|_F^2.$$

Differentiating this function w.r.t. ρ and finding its solution to zero, we immediately get (4.20).

Further, if it is assumed that $\{\mathbf{x}_i\}$ follow i.i.d. $N(\mathbf{0},\Sigma)$, by Wick's theorem, we have

$$\begin{aligned} E\hat{s}_{ij}^2 &= \frac{1}{n^2}E\left(\sum_{k=1} x_{ki}x_{kj}\right)^2 = \frac{1}{n^2}E\left(\sum_{k=1}^{n} x_{ki}^2x_{kj}^2 + \sum_{1\le k\ne k'\le n} x_{ki}x_{kj}x_{k'i}x_{k'j}\right) \\ &= \frac{1}{n}Ex_{ki}^2x_{kj}^2 + \frac{1}{n^2}\sum_{1\le k\ne k'\le n} E(x_{ki}x_{kj})E(x_{k'i}x_{k'j}) \\ &= \frac{1}{n}\left(\sigma_{ii}\sigma_{jj} + 2\sigma_{ij}^2\right) + \frac{n(n-1)}{n^2}\sigma_{ij}^2 = \frac{\sigma_{ii}\sigma_{jj}}{n} + \frac{n+1}{n}\sigma_{ij}^2. \end{aligned}$$

Now, it follows from direct calculation that

$$\begin{aligned} E\|V\circ\hat{S}\|_F^2 &= E\sum_{i=1}^{p}\sum_{j=1}^{p} w_{ij}\hat{s}_{ij}^2 = \sum_{i=1}^{p}\sum_{j=1}^{p} w_{ij}E\hat{s}_{ij}^2 \\ &= \sum_{i=1}^{p}\sum_{j=1}^{p} w_{ij}\left(\frac{n+1}{n}\sigma_{ij}^2 + \frac{1}{n}\sigma_{ii}\sigma_{jj}\right) = \frac{n+1}{n}\|V\circ\Sigma\|_F^2 + \frac{1}{n}\mathrm{Tr}(DV^2D), \end{aligned}$$

where $D = \mathrm{diag}(\Sigma)$; so similarly

$$E\|W\circ\hat{S}\|_F^2 = \frac{n+1}{n}\|W\circ\Sigma\|_F^2 + \frac{1}{n}\mathrm{Tr}(DW^2D).$$

Substituting into (4.20), we have

$$\hat{\rho}^{\text{STO}} = \frac{\|\Sigma\|_F^2 + \text{Tr}^2(\Sigma) - \|V \circ \Sigma\|_F^2 - \text{Tr}(DV^2D)}{(n+1)(\|\Sigma\|_F^2 + \|W \circ \Sigma\|_F^2 - 2\|V \circ \Sigma\|_F^2) + \text{Tr}^2(\Sigma) + \text{Tr}(DW^2D) - 2\text{Tr}(DV^2D)}. \tag{48}$$

If $\hat{\rho}^{\text{STO}} \notin [0,1]$, we modify $\hat{\rho}^{\text{STO}}$ to be either 0 or 1, whichever gives a smaller MSE since it is a quadratic function of ρ and therefore attains the minimum value at one of its boundary points. □

A.5 Proofs for Chapter 5

A.5.1 Proof of Lemma 5.2.1

Suppose $A \in \mathcal{G}_q(c_{n,p})$, for $\forall j$, we have

$$|a_{(i)j}|^q \leq \frac{c_{n,p}}{i}, \tag{49}$$

where (i) is the decreasingly ordered i-th index (in magnitude) of $\mathbf{a}_j$, the j-th column of A. Fix an $\varepsilon > 0$ and choose $B \in S_k$ such that

$$b_{ij} = \begin{cases} a_{ij}, & \text{for } |(i)| \leq k. \\ 0, & \text{otherwise.} \end{cases} \tag{50}$$

Then, it follows that

$$\begin{aligned} \|A - B\| &\leq \|A - B\|_{L^1} = \sup_j \sum_i |a_{ij} - b_{ij}| = \sup_j \sum_{i>k} |a_{(i)j}| \leq \sup_j \sum_{i>k} \left(\frac{c_{n,p}}{i}\right)^{1/q} \\ &= c_{n,p}^{1/q} \sup_j \sum_{i>k} i^{-\frac{1}{q}} \leq \frac{q}{1-q} c_{n,p}^{\frac{1}{q}} k^{1-\frac{1}{q}} \leq \varepsilon \end{aligned} \tag{51}$$

for k being large enough, since $-q^{-1} < -1$.

For part 2), we will use the following Lemma:

Lemma A.5.1. *For each $k \in \mathbb{N}, 0 \leq q < 1$, and $x_1 \geq x_2 \cdots \geq x_k \geq 0$, the following holds*

$$\Big(\sum_{i=1}^{k} i^{\frac{1}{q}-1} x_i\Big)^q \leq \sum_{i=1}^{k} x_i^q. \tag{52}$$

Proof. Without loss of generality, let us prove an equivalent statement with $y_i := x_i^q$, $q' := \frac{1}{q}$,

$$\sum_{i=1}^{k} i^{q'-1} y_i^{q'} \leq \Big(\sum_{i=1}^{k} y_i\Big)^{q'}.$$

This follows from

$$\begin{aligned}\Big(\sum_{i=1}^{k} y_i\Big)^{q'} &= \Big(\sum_{i=1}^{k-1} y_i + y_k\Big)^{q'} = \Big(\sum_{i=1}^{k-1} y_i\Big)^{q'} \Big(1 + \frac{y_k}{\sum_{i=1}^{k-1} y_i}\Big)^{q'} \\ &\geq \Big(\sum_{i=1}^{k-1} y_i\Big)^{q'} \Big[1 + \Big(1 + \frac{\sum_{i=1}^{k-1} y_i}{y_k}\Big)^{q'-1} \Big(\frac{y_k}{\sum_{i=1}^{k-1} y_i}\Big)^{q'}\Big].\end{aligned}$$

At the last line above, for $q' > 1$, we use

$$(1+x)^{q'} \geq 1 + (1 + \frac{1}{x})^{q'-1} x^{q'}, \tag{53}$$

which can be shown by integrating for $0 \leq t \leq x$,

$$q'(1+t)^{q'-1} \geq q' t^{q'-1} (1 + 1/x)^{q'-1}.$$

Therefore,

$$\Big(\sum_{i=1}^{k} y_i\Big)^{q'} \geq \Big(\sum_{i=1}^{k-1} y_i\Big)^{q'} + \Big(1 + \frac{\sum_{i=1}^{k-1} y_i}{y_k}\Big)^{q'-1} y_k^{q'} \geq \Big(\sum_{i=1}^{k-1} y_i\Big)^{q'} + k^{q'-1} y_k^{q'} \qquad (\text{since } \tfrac{y_i}{y_k} \geq 1).$$

The desired inequality follows from induction on k. □

Now consider two matrices $A = \big(a_{ij}\big)_{p \times p} \in \mathcal{G}_q(c_{n,p})$, $A' = \big(a'_{ij}\big)_{p \times p} \in \mathcal{G}_q(c'_{n,p})$. Fix

$1 \le j \le p$. Then, for $AA' = \left(a''_{ij}\right)_{p\times p}$,

$$\sum_i |a''_{ij}|^q \le \sum_i \left(\sum_\ell |a_{i\ell}a'_{\ell j}|\right)^q \le \sum_i \left(\sum_\ell |a_{i\sigma[\ell]}a'_{(\ell)j}|\right)^q,$$

where (ℓ) is the ordered index as in (49), and σ is a permutation of $\{1, 2, \cdots, p\}$ due to the reordering. Thus,

$$\sum_i |a''_{ij}|^q \le \sum_i \left[\sum_\ell \left(\frac{c'_{n,p}}{\ell}\right)^{\frac{1}{q}} |a_{i\sigma[\ell]}|\right]^q = c'_{n,p} \sum_i \left(\sum_\ell \ell^{-\frac{1}{q}} |a_{i\sigma[\ell]}|\right)^q.$$

Reorder $\ell^{-\frac{1}{q}} a_{i\sigma[\ell]}$'s according to their magnitude. Denote this new sequence by $z_{(\ell)}$: $z_{(1)} \ge z_{(2)} \ge z_{(3)} \ge \cdots \ge z_{(p)}$. Notice that since $\frac{1}{q} - 1 > 0$,

$$\frac{z_{(\ell)}}{\ell^{\frac{1}{q}-1}} \ge \frac{z_{(\ell+1)}}{(\ell+1)^{\frac{1}{q}-1}}, \qquad \ell = 1, 2, \cdots, p-1.$$

Therefore, we can use (52) to see

$$\begin{aligned}\sum_i \left(\sum_\ell \ell^{-\frac{1}{q}} |a_{i\sigma[\ell]}|\right)^q &= \sum_i \left(\sum_\ell z_{(\ell)}\right)^q = \sum_i \left(\sum_\ell \ell^{\frac{1}{q}-1} \frac{z_{(\ell)}}{\ell^{\frac{1}{q}-1}}\right)^q \\ &\le \sum_i \sum_\ell \frac{z^q_{(\ell)}}{\ell^{1-q}} = \sum_i \sum_\ell (\sigma'[\ell])^{q-1} \ell^{-1} |a_{i\sigma[\ell]}|^q\end{aligned}$$

for some permutation σ'. By interchanging the summations and using Hölder's inequality, this last term is bounded by

$$\begin{aligned}c_{n,p} \sum_\ell (\sigma'[\ell])^{q-1} \ell^{-1} &\le c_{n,p} \left(\sum_\ell (\sigma'[\ell])^{-2}\right)^{\frac{1-q}{2}} \left(\sum_\ell \ell^{-\frac{2}{1+q}}\right)^{\frac{1+q}{2}} \\ &= c_{n,p} \left(\sum_\ell \ell^{-2}\right)^{\frac{1-q}{2}} \left(\sum_\ell \ell^{-\frac{2}{1+q}}\right)^{\frac{1+q}{2}} \\ &\le c_{n,p} C(q)\end{aligned}$$

for some constant $C(q)$ depending only on q; here, $C(q) < \infty$ follows from $q < 1$.

Thus, we have the estimate

$$\sum_i |a''_{ij}|^q \leq C(q) c_{n,p} c'_{n,p}. \tag{54}$$

Then, we deduce by induction that for $A^r = (a_{ij}^{(r)})$,

$$\sup_j \sum_i \left| a_{ij}^{(r)} \right|^q \leq C(q)^r c_{n,p}^r. \tag{55}$$

Equivalently, we have $A^r \in \mathcal{G}_q(C(q)^r c_{n,p}^r)$ if $A \in \mathcal{G}_q(c_{n,p})$. Hence, we can construct $B' \in S_{k'}$ as in (50) such that $b'_{ij} = a_{ij}^{(r)}$ for $|(i)| \leq k'$ and $b'_{ij} = 0$ otherwise, where k' is defined in (5.17). The proof is then complete because$\|A^r - B'\| \leq \varepsilon$ for all

$$k' \geq \left[\frac{q}{(1-q)\varepsilon} C(q)^{\frac{r}{q}} c_{n,p}^{\frac{r}{q}} \right]^{\frac{q}{1-q}} = k_{\min} C(q)^{\frac{r}{1-q}} c_{n,p}^{\frac{r-1}{1-q}}.$$

□

A.5.2 Proof of Theorem 5.2.2

Suppose $\Sigma \in \mathcal{F}(q, m)$ and let $\Omega = \Sigma^{-1}$. Then $m \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq m^{-1}$. Since $\lambda_{\min}(\Sigma) = \lambda_{\max}^{-1}(\Omega)$ and $\lambda_{\max}(\Sigma) = \lambda_{\min}^{-1}(\Omega)$, we have

$$m \leq \lambda_{\min}(\Omega) \leq \lambda_{\max}(\Omega) \leq m^{-1}. \tag{56}$$

Thus, $\Omega \in \mathcal{U}(m)$. Write Ω in terms of its Neumann series as

$$\Omega = \Sigma^{-1} = B_r + R_r,$$

where B_r is defined in (5.10) and

$$R_r = \eta \sum_{j=r+1}^{\infty} (I - \eta\Sigma)^j. \tag{57}$$

It then follows from (5.9) that, for any $\varepsilon > 0$, there is an $r_0 = r_0(m, \varepsilon) \in \mathbb{N}$ such that $\|R_r\| \leq \varepsilon$ for all $r \geq r_0$. Also, we have by assumption that $\Sigma \in \mathcal{G}_q(c_{n,p})$ for some $c_{n,p} \geq 0$ which in turn implies by Lemma 5.2.1 that, for any $j \geq 1$, $(I - \eta\Sigma)^j \in \mathcal{G}_q\left((1 + \eta c_{n,p})^j C(q)^j\right)$. Then it is obtained (e.g. using (52)) that

$$B_{r_0} \in \mathcal{G}_q\left(\eta^q r_0 C(q)^{r_0}(1 + \eta c_{n,p})^{r_0}\right) \subset \mathcal{G}_q. \tag{58}$$

The second claim of the theorem follows from (5.17). □

A.5.3 Proof of Proposition 5.3.1

By definition, $\sigma^\star_{jk} = n^{-1}\sum_{i=1}^n x_{ij}x_{ik}$. We begin with

$$\begin{aligned} P(|\sigma^\star_{jk} - \sigma_{jk}| \geq t) &= P\left(\frac{1}{n}\left|\sum_{i=1}^n x_{ij}x_{ik} - \sigma_{jk}\right| \geq t\right) \\ &= P\left(\frac{1}{n}\left|\sum_{i=1}^n \tilde{x}_{ij}\tilde{x}_{ik} - \rho_{jk}\right| \geq \frac{t}{\sqrt{\sigma_{jj}\sigma_{kk}}}\right), \end{aligned} \tag{59}$$

where $\tilde{x}_{ij} = x_{ij}/\sqrt{\sigma_{jj}}$ (similar definition for $\tilde{x}_{ik}$) and $\rho_{jk} = \sigma_{jk}/\sqrt{\sigma_{jj}\sigma_{kk}}$. Because

$$\begin{pmatrix} \tilde{x}_{ij} \\ \tilde{x}_{ik} \end{pmatrix} \overset{\text{i.i.d.}}{\sim} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{jk} \\ \rho_{kj} & 1 \end{pmatrix}\right),$$

we deduce that

$$\begin{aligned} \tilde{x}_{ij} + \tilde{x}_{ik} &\overset{\text{i.i.d.}}{\sim} N(0, 2(1 + \rho_{jk})), \\ \tilde{x}_{ij} - \tilde{x}_{ik} &\overset{\text{i.i.d.}}{\sim} N(0, 2(1 - \rho_{jk})). \end{aligned}$$

Therefore, from the polarization identity

$$\sum_i (\tilde{x}_{ij}\tilde{x}_{ik} - \rho_{ij}) = \frac{1}{4}\sum_i \left[(\tilde{x}_{ij} + \tilde{x}_{ik})^2 - 2(1 + \rho_{jk}) - (\tilde{x}_{ij} - \tilde{x}_{ik})^2 + 2(1 - \rho_{jk})\right],$$

the expression (59) can be bounded by

$$\begin{aligned}
\leq & P\left(\frac{1}{4n}\left|\sum_i (\tilde{x}_{ij}+\tilde{x}_{ik})^2 - 2(1+\rho_{jk})\right| \geq \frac{t}{2\sqrt{\sigma_{jj}\sigma_{kk}}}\right) \\
& + P\left(\frac{1}{4n}\left|\sum_i (\tilde{x}_{ij}-\tilde{x}_{ik})^2 - 2(1-\rho_{jk})\right| \geq \frac{t}{2\sqrt{\sigma_{jj}\sigma_{kk}}}\right) \\
\leq & 2P\left(\frac{1}{n}\left|\sum_i \left(V_i^2-1\right)\right| \geq \frac{t}{(1+|\rho_{jk}|)\sqrt{\sigma_{jj}\sigma_{kk}}}\right),
\end{aligned}$$

where V_i are i.i.d. $N(0,1)$. Note that $(1+|\rho_{jk}|)\sqrt{\sigma_{jj}\sigma_{kk}} \leq 2m^{-1}$ since $\Sigma \in \mathcal{U}(m)$ and V_i^2 are $\chi^2(1)$ with exponential tail, it follows from [120, Corollary 17, page 16] that there are constants $C_4, C_5 > 0$, depending only on C_1, C_2 and m, such that

$$\begin{aligned}
& P\left(\frac{1}{n}\left|\sum_i \left(V_i^2-1\right)\right| \geq \frac{t}{(1+|\rho_{jk}|)\sqrt{\sigma_{jj}\sigma_{kk}}}\right) \\
& \leq P\left(\frac{1}{n}\left|\sum_i \left(V_i^2-1\right)\right| \geq \frac{mt}{2}\right) \\
& \leq 2\exp\left[-C_4 \min\left(\frac{t^2}{C_5^2}, \frac{t}{C_5}\right) n\right].
\end{aligned}$$

The proposition is therefore proved. □

A.5.4 Proof of Theorem 5.3.3

By choosing $\varepsilon = O(c_{n,p}(\log p/n)^{(1-q)/2})$ in the forth step of our algorithm, it suffices to show $\left\|\tilde{\Omega} - \Omega\right\|$ obeys the upper bound (5.24). By (5.23) and in [11, Theorem 1],

$$\left\|\tilde{\Sigma}_n - \Sigma\right\| \leq C c_{n,p} \left(\frac{\log p}{n}\right)^{(1-q)/2} \tag{60}$$

with probability greater than $(1 - C_6 p^{-8\tau^2/C_4^2+2})$ that approaches to 1 whenever $\tau > C_4/2$. With such a high probability,

$$\begin{aligned}
\left\|\tilde{\Omega} - \Omega\right\| = \eta & \left\|\sum_{j=0}^{r}(I - \eta\tilde{\Sigma}_n)^j - \sum_{j=0}^{\infty}(I - \eta\Sigma)^j\right\| \\
\leq & \eta \left\|\sum_{j=0}^{r}\left[(I - \eta\tilde{\Sigma}_n)^j - (I - \eta\Sigma)^j\right]\right\| + \eta \sum_{j=r+1}^{\infty} \left\|(I - \eta\Sigma)^j\right\| \\
\leq & \eta^2 \sum_{j=1}^{r} C(j,\tau)\left\|\tilde{\Sigma}_n - \Sigma\right\| + \frac{1}{m}\left(\frac{1-m^2}{1+m^2}\right)^{r+1} \\
\leq & C(q,m,\tau)\left\|\tilde{\Sigma}_n - \Sigma\right\| + \frac{1}{m}\left(\frac{1-m^2}{1+m^2}\right)^{r+1}, \qquad (61)
\end{aligned}$$

where the last two inequalities follow from Lemma A.5.2 (see below) and $\|I - \eta\Sigma\| = (1 - m^2)/(1 + m^2) \leq \delta < 1$. Note that $c_{n,p}(\log p/n)^{(1-q)/2} \to 0$ as $n \to \infty$, thus from (60) the theorem follows. □

The following lemma says that the matrix power operation A^r is a contraction mapping for $\|A\|$ uniformly bounded up by 1 on appropriate subsets.

Lemma A.5.2. *Fix a $\delta \in (0, 1)$ and $\varepsilon > 0$. Let A and B be any two square matrices such that $\|B\| \leq \delta$ and $\|A - B\| \leq \varepsilon$. Then for all $r \in \mathbb{N}$, there exists a constant, $C(r, \delta + \varepsilon)$, depending only on r and $\delta + \varepsilon$ such that*

$$\|A^r - B^r\| \leq C(r, \delta + \varepsilon) \|A - B\|. \qquad (62)$$

Moreover, if $\delta + \varepsilon < 1$, then the sequence $\{C(r, \delta + \varepsilon)\}_{r=1}^{\infty}$ is summable; in particular,

$$\sum_{r=1}^{\infty} \|A^r - B^r\| \lesssim \|A - B\|, \qquad (63)$$

where the constant here depends on $(\delta + \varepsilon)$.

Proof. The proof is standard and is based on induction. By assumptions, $\|A\| \leq \delta + \varepsilon < 1$. For $r = 1$, select $C(1, \delta + \varepsilon) = 1$. Suppose (62) holds for r. Observe the following identity

$$A^{r+1} - B^{r+1} = A(A^r - B^r) + (A - B)B^r, \tag{64}$$

which in turn implies that

$$\begin{aligned} \left\|A^{r+1} - B^{r+1}\right\| \leq & \|A\| \|A^r - B^r\| + \|A - B\| \|B\|^r \\ \leq & (\delta + \varepsilon)C(r, \delta + \varepsilon) \|A - B\| + \delta^r \|A - B\| \\ = & [(\delta + \varepsilon)C(r, \delta + \varepsilon) + \delta^r] \|A - B\| , \end{aligned} \tag{65}$$

where the induction hypothesis is used in the second inequality. Choose

$$C(r + 1, \delta + \varepsilon) = (\delta + \varepsilon)C(r, \delta + \varepsilon) + \delta^r \tag{66}$$

and (62) follows. Now, we analyze the property of $C(r, \delta + \varepsilon)$. By (66), simple recursion yields

$$C(r, \delta + \varepsilon) = \sum_{k=0}^{r-1} (\delta + \varepsilon)^k \delta^{r-1-k} \leq r(\delta + \varepsilon)^{r-1}.$$

Since $\delta + \varepsilon < 1$, the lemma is hence proved. □

A.5.5 Proof of Proposition 5.3.5

The proof essentially is similar to the proof of Theorem 5.3.3 except changing the norm. Thus, we only sketch the important steps and emphasize the differences. First, we invoke a simple fact about the equivalence between the spectral and entry-∞ norms of a matrix $A_{n\times m}$; that is,

$$\|A\|_\infty \leq \|A\| \leq \sqrt{mn} \|A\|_\infty . \tag{67}$$

For the proposition, we only need the left inequality and thus prove it. Indeed, let (i_0, j_0) be the index pair with its corresponding entry attaining $\|A\|_\infty$. Let $\mathbf{x}_0 = (0, \cdots, 1, \cdots, 0)^T$ where 1 is in the j_0-th position. Then by definition of the spectral norm, we have

$$\|A\| = \sup_{\|\mathbf{x}\|=1} \|A\mathbf{x}\| \geq \|A\mathbf{x}_0\| = \|\mathbf{a}_{j_0}\| \geq \|\mathbf{a}_{j_0}\|_\infty = |a_{i_0 j_0}| = \|A\|_\infty ,$$

where $\mathbf{a}_{j_0}$ is the j_0-th column of A. Furthermore, by (5.23) and the union bound

$$\|\Sigma_n^\star - \Sigma\|_\infty \leq \tau \sqrt{\frac{\log p}{n}} \tag{68}$$

with probability greater than $1 - p^{-8\tau^2/C_4^2+2}$ that approaches to 1 whenever $\tau > C_4/2$. But this implies that

$$\left\|\tilde{\Sigma}_n - \Sigma\right\|_\infty \leq \|T_t\Sigma_n^\star - T_t\Sigma\|_\infty + \|T_t\Sigma - \Sigma\|_\infty \leq \|\Sigma_n^\star - \Sigma\|_\infty + 2\tau\sqrt{\frac{\log p}{n}} = O_p\left(\sqrt{\frac{\log p}{n}}\right), \tag{69}$$

since $\|T_t(A) - T_t(B))\|_\infty \leq \|A - B\|_\infty + t$. Now, the argument proceeds as in Theorem 5.3.3. Since $\|I - \eta\Sigma\|_\infty \leq \|I - \eta\Sigma\| < 1$, we thus have, with high probability,

$$\left\|\tilde{\Omega} - \Omega\right\|_\infty \leq C(q, m, \tau)\left\|\tilde{\Sigma}_n - \Sigma\right\|_\infty + \frac{1}{m}\left(\frac{1 - m^2}{1 + m^2}\right)^{r+1} \leq C(q, m, \tau)\left(\sqrt{\frac{\log p}{n}} + \delta^{r+1}\right). \tag{70}$$

□

A.5.6 Proof of Theorem 5.3.6

Before proving Theorem 5.3.6, we present a few technical lemmas that bound the magnitudes $E\left\|\tilde{\Sigma}_n - \Sigma\right\|^r$ on some "bad set". Define

$$B = \left\{\left\|\tilde{\Sigma}_n - \Sigma\right\| > Cc_{n,p}\left(\frac{\log p}{n}\right)^{(1-q)/2}\right\}. \tag{71}$$

Lemma A.5.3. *Under the assumptions in Theorem 5.3.6, there is a constant* $C(q,m,\tau) > 0$ *independent of* $r, n,$ *and* p*, such that we have*

$$E\left(\left\|\tilde{\Sigma}_n - \Sigma\right\|^r; B\right) \lesssim \left[C(q,m,\tau)c_{n,p}\left(\frac{\log p}{n}\right)^{(1-q)/2}\right]^r. \tag{72}$$

Proof. By the proof of [11, Theorem 1], we extract the consequence that

$$P\left(\left\|\hat{\Sigma}_n - \Sigma\right\| \le C_1 c_{n,p} t^{1-q}\right) \ge 1 - C_2 p^2 \exp(-C_3 n t^2), \tag{73}$$

where $t = \tau\sqrt{\log p/n}$ is the thresholding parameter (tending to 0); or equivalently, write

$$P\left(\left\|\hat{\Sigma}_n - \Sigma\right\| > t'\right) < C_2 p^2 \exp\left[-C_4 n \left(\frac{t'}{c_{n,p}}\right)^{2/(1-q)}\right], \tag{74}$$

for $t' = C_1 c_{n,p} t^{1-q}$. Applying Lemma A.5.4 with $X = \left\|\hat{\Sigma}_n - \Sigma\right\|$ and t', we therefore have that

$$\begin{aligned} E\left(\left\|\tilde{\Sigma}_n - \Sigma\right\|^r; B\right) &= t'^r P(X > t') + \int_{t'^r}^{C_5^r} P(X > u^{1/r})\, du + \int_{C_5^r}^{\infty} P(X > u^{1/r})\, du \\ &\overset{\text{def}}{=} R_1 + R_2 + R_3. \end{aligned}$$

We now develop upper bounds for R_1, R_2, and R_3. For R_1, we see from (74) that

$$P(X > t') \le C_2 p^{-C_5+2} \to 0 \tag{75}$$

and therefore it follows that for large enough τ

$$R_1 = o\left(\left(C_1 c_{n,p} t^{1-q}\right)^r\right). \tag{76}$$

Writing R_2 by definition and applying Lemma A.5.5 with $\alpha = 2/[r(1-q)]$ and

$C_2 = C_4 c_{n,p}^{-2/(1-q)}$, we deduce that

$$\begin{aligned} R_2 &\leq C_2 p^2 \int_{t'^r}^{\infty} \exp\left[-C_4 n \left(\frac{u^{1/r}}{c_{n,p}} \right)^{2/(1-q)} \right] du \\ &= C_2 p^2 \frac{r(1-q)}{2} c_{n,p}^r (C_4 n)^{-r(1-q)/2} \int_{C_6 \log p}^{\infty} v^{\frac{r(1-q)}{2}-1} e^{-v} \, dv, \end{aligned} \tag{77}$$

where the last term is an upper incomplete gamma function

$$\Gamma\left(\frac{r(1-q)}{2}; C_6 \log p \right) \asymp p^{-C_6} (C_6 \log p)^{\frac{r(1-q)}{2}-1},$$

as $p \to \infty$. But this now implies that

$$R_2 \lesssim \left[C_7 c_{n,p} \left(\frac{\log p}{n} \right)^{\frac{1-q}{2}} \right]^r, \tag{78}$$

since $p \geq n^{\xi}$ and $r = O(\log n)$. Invoking LemmaA.5.5 twice, we see that R_3 can be bounded in a similar means as in R_2, except for the difference that the tail is sub-exponential rather than sub-Gaussian. The lemma now follows from (76) and (78). □

Lemma A.5.4. *Let X be a non-negative r.v., $t \geq 0$, and $r \in \mathbb{N}$. Then*

$$E\left(X^r; X > t \right) = t^r P(X > t) + \int_{t^r}^{\infty} P(X > u^{1/r}) \, du. \tag{79}$$

Proof. The lemma is an easy consequence of Fubini's theorem. □

Lemma A.5.5. *Let X be a non-negative r.v. and $\alpha > 0$. Suppose that there are absolute constants C_1 and C_2 such that*

$$P(X > t) \leq C_1 \exp\left(-C_2 n t^{\alpha} \right) \tag{80}$$

for all $t \in [a, b]$ where $0 \leq a \leq b \leq \infty$. Then we have

$$\int_a^b P(X > u)\, du \leq C_1 \alpha^{-1} (C_2 n)^{-1/\alpha} \int_{C_2 n a^\alpha}^{C_2 n b^\alpha} v^{\alpha^{-1}-1} e^{-v}\, dv. \tag{81}$$

Proof. Let $v = C_2 n u^\alpha$. Then the lemma follows from a direct application of the change of variables. □

Proof of Theorem 5.3.6: By Theorem 5.3.3 and Lemma A.5.3, we have

$$\begin{aligned} E\left\|\hat{\Omega} - \Omega\right\|^2 &= E\left(\left\|\hat{\Omega} - \Omega\right\|^2 ; B^c\right) + E(\left(\left\|\hat{\Omega} - \Omega\right\|^2 ; B\right) \\ &\leq C(q, m, \tau)\left[c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q} + \delta^{2(r+1)}\right] + E(\left(\left\|\hat{\Omega} - \Omega\right\|^2 ; B\right). \qquad (82) \\ &\leq C(q, m, \tau)\left[c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q} + \delta^{2(r+1)}\right] + O\left(\sum_{j=2}^{2r} C^j c_{p,n}^j \left(\frac{\log p}{n}\right)^{j(1-q)/2}\right) \\ &= O\left(c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q} + \delta^{2(r+1)}\right), \qquad (83) \end{aligned}$$

since $c_{n,p}(\log p/n)^{(1-q)/2} \to 0$. This is the content of theorem. □

A.5.7 Proof of Theorem 5.3.9

Considering the general lower bound in Proposition 5.3.8 with $\psi(\theta) = (\Sigma(\theta))^{-1}$, similarly as in [22], we have

$$\begin{aligned} \inf_{\hat{\Omega}} \sup_{\Sigma \in \mathcal{G}_q(c_{n,p}) \cap \mathcal{U}(m)} E\left\|\hat{\Omega} - \Sigma^{-1}\right\|^2 &\geq \inf_{\hat{\Omega}} \max_{\theta \in \Theta} E_\theta \left\|\hat{\Omega} - \Sigma(\theta)^{-1}\right\|^2 \\ &\geq \frac{\alpha}{4} \frac{p}{4} \min_{1 \leq i \leq J} \left\|\bar{P}_{i,0} \wedge \bar{P}_{i,1}\right\|, \end{aligned} \tag{84}$$

where

$$\alpha = \min_{(\theta, \theta') : H(\gamma(\theta), \gamma(\theta')) \geq 1} \frac{\|\Sigma(\theta)^{-1} - \Sigma(\theta')^{-1}\|^2}{H(\gamma(\theta), \gamma(\theta'))}. \tag{85}$$

Since $\Sigma(\theta)$ and $\Sigma(\theta')$ belong to $\mathcal{U}(m)$, it easily follows that

$$\left\|\Sigma(\theta)^{-1} - \Sigma(\theta')^{-1}\right\| \geq m^2 \left\|\Sigma(\theta) - \Sigma(\theta')\right\|. \tag{86}$$

Now, the theorem follows from [22, Lemma 5 and Lemma 6]: that is, we have

$$\inf_{\hat{\Omega}} \sup_{\Sigma \in \mathcal{G}_q(c_{n,p}) \cap \mathcal{U}(m)} E \left\|\hat{\Omega} - \Sigma^{-1}\right\|^2 \geq C' c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q} \tag{87}$$

for some constant $C' > 0$. □

A.6 Proofs for Chapter 6

A.6.1 Proof of Theorem 6.2.1

The proof of the estimation and sign consistency for the group robust lasso estimator is based on the argument developed in [31, Theorem IV.1].

Let Z_n be the objective function to minimized. Define

$$\begin{aligned} V_n(\mathbf{u}) &= \delta \sum_{i=1}^{n} \left(\mathbf{u}^* \frac{\mathbf{x}_i \mathbf{x}_i^*}{n} \mathbf{u}\right) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{u}^* \mathbf{x}_i Y_i \\ &+ (1-\delta) \sum_{i=1}^{n} Z_i + \lambda_n \sum_{g=1}^{G} \left[\left\|\boldsymbol{\beta}_g + \frac{\mathbf{u}_g}{\sqrt{n}}\right\|_2 - \left\|\boldsymbol{\beta}_g\right\|_2\right], \end{aligned} \tag{88}$$

where

$$Y_i = (1-\delta)(1(e_i < 0) - 1(e_i \geq 0)) - 2\delta e_i$$

and

$$Z_i = 2 \times \left(\frac{\mathbf{u}^* \mathbf{x}_i}{\sqrt{n}} - e_i\right) 1 \left(0 \leq e_i < \frac{\mathbf{u}^* \mathbf{x}_i}{\sqrt{n}}\right).$$

Then it follows that $\sqrt{n}(\hat{\boldsymbol{\beta}}_n^H - \boldsymbol{\beta})$ minimizes V_n and the first three terms together converge in distribution to

$$(\delta + (1-\delta) f(0)) \mathbf{u}^* C \mathbf{u} + \mathbf{u}^* \mathbf{W},$$

where $\mathbf{W} \sim N(\mathbf{0}, ((1-\delta)^2 + 4\delta^2\sigma^2 + 4\delta(1-\delta)M_{10})\, C)$. For the last term, we divide into two cases. For $\boldsymbol{\beta}_g \neq \mathbf{0}$, we have

$$\sqrt{n}\left(\left\|\boldsymbol{\beta}_g + \frac{\mathbf{u}_g}{\sqrt{n}}\right\|_2 - \left\|\boldsymbol{\beta}_g\right\|_2\right) \rightarrow \frac{{\mathbf{u}_g}^T\boldsymbol{\beta}_g}{\left\|\boldsymbol{\beta}_g\right\|_2} \tag{89}$$

by identifying the derivative of the Euclidean norm. So it follows that

$$\underbrace{\frac{\lambda_n}{\sqrt{n}}}_{\rightarrow 0} \times \underbrace{\sqrt{n}\left(\left|\beta_j + \frac{u_j}{\sqrt{n}}\right| - |\beta_j|\right)}_{\rightarrow \frac{{\mathbf{u}_g}^T\boldsymbol{\beta}_g}{\left\|\boldsymbol{\beta}_g\right\|_2}} \xrightarrow{P} 0. \tag{90}$$

where we have used our assumption on the shrinkage rates $\lambda_n/\sqrt{n} \rightarrow 0$. For $\boldsymbol{\beta}_g = \mathbf{0}$, it is obvious that

$$\left\|\boldsymbol{\beta}_g + \frac{\mathbf{u}_g}{\sqrt{n}}\right\|_2 - \left\|\boldsymbol{\beta}_g\right\|_2 = n^{-\frac{1}{2}}\left\|\mathbf{u}_g\right\|_2$$

so we obtain that

$$\lambda_n n^{-\frac{1}{2}}\left\|\mathbf{u}_g\right\|_2 = \underbrace{\lambda_n n^{\frac{\gamma-1}{2}}}_{\rightarrow\infty} \underbrace{\left\|\sqrt{n}\hat{\boldsymbol{\beta}}_{ng}^{LS}\right\|_2^{-\gamma}}_{O_p(1)} \left\|\mathbf{u}_g\right\|_2 \xrightarrow{P} \infty. \tag{91}$$

Putting all the terms together and applying Slutsky's lemma, we deduce that $V_n(\mathbf{u}) \Rightarrow V(\mathbf{u})$ for each $\mathbf{u} \in \mathbb{R}^p$ where

$$V(\mathbf{u}) = \begin{cases} (\delta + (1-\delta)f(0))\mathbf{u}^* C\mathbf{u} + \mathbf{u}^*\mathbf{W} & \text{if } \mathbf{u}_g = \mathbf{0} \forall g \notin A, \\ \infty & \text{otherwise.} \end{cases} \tag{92}$$

Since V_n is convex and V has unique minimum, it follows from the standard epi-convergence results that $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) = \arg\min(V_n) \Rightarrow \arg\min(V)$ which is equivalent to say that $\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{nA} - \boldsymbol{\beta}_A\right) \Rightarrow \mathbf{C}_{AA}^{-1}\mathbf{W}_A$ and $\sqrt{n}\hat{\boldsymbol{\beta}}_{nA^c} \xrightarrow{P} \mathbf{0}$ with $\mathbf{u}_A$ is the restriction of $\mathbf{W}$ to the support of the true coefficient vector. Now the theorem follows. □