©Copyright 2017
Sen Zhao

Hypothesis Testing With High-Dimensional Data

Sen Zhao

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Ali Shojaie, Chair

Marco Carone

Mathias Drton

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Hypothesis Testing With High-Dimensional Data

Sen Zhao

Chair of the Supervisory Committee:
Associate Professor Ali Shojaie
Biostatistics

In the past two decades, vast high-dimensional biomedical datasets have become the mainstay in various biomedical applications from genomics to neuroscience. These high-dimensional data enable researchers to answer scientific questions that are impossible to answer with classical, low-dimensional datasets. However, due to the "curse of dimensionality", such high-dimensional datasets also pose serious statistical challenges. Motivated by these emerging applications, statisticians have devoted much effort to developing estimation methods for high-dimensional linear models and graphical models. However, there is still little progress on quantifying the uncertainty of the estimates, e.g., obtaining p-values and confidence intervals, which are crucial for drawing scientific conclusions. While encouraging advances have been made in this area over the past couple of years, the majority of existing high-dimensional hypothesis testing methods still suffer from low statistical power or high computational intensity.

In this dissertation, we focus on developing hypothesis testing methods for high-dimensional linear and graphical models. In Chapter 2, we investigate a naïve and simple two-step hypothesis testing procedure for linear models. We show that, under appropriate conditions, such a simple procedure controls type-I error rate, and is closely connected to more complicated alternatives. We also show in numerical studies that such a simple procedure achieves similar performance as procedures that are computationally more intense. In Chapter 3, we

consider hypothesis testing for linear regression that incorporates external information about the relationship between variables represented by a graph, such as the gene regulatory network. We show in theory and numerical studies that by incorporating informative external information, our proposal is substantially more powerful than existing methods that ignore such information. We also propose a more robust procedure for settings where the external information is potentially inaccurate or imprecise. This robust procedure could adaptively choose the amount of external information to be incorporated based on the data. In Chapter 4, we shift our focus to Gaussian graphical models. We propose a novel procedure to test whether two Gaussian graphical models share the same edge set while controlling the false positive rate. In the case that two networks are different, our proposals could identify specific nodes and edges that show differential connectivity. In this chapter, we also demonstrate that when the goal is to identify differentially connected nodes and edges, the results from our proposal are more interpretable than existing procedures based on covariance or precision matrices. We finish the dissertation with a discussion in Chapter 5, in which we present viable future research directions, and discuss a possible extension of our proposals to vector autoregression models for time series.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

CDF: Cumulative Distribution Function.

CRAN: Comprehensive R Archive Network.

CV: Cross-Validation.

DNA: Deoxyribonucleic Acid.

FDR: False Discovery Rate.

FPR: False Positive Rate.

FMRI: Functional Magnetic Resonance Imaging.

FWER: Family-Wise Error Rate.

GGM: Gaussian Graphical Model.

GLM: Generalized Linear Model.

I.I.D.: Independent and Identically Distributed.

KEGG: Kyoto Encyclopedia of Genes and Genomes

OLS: Ordinary Least Squares.

PMSE: Prediction Mean Squared Error.

PSA: Prostate-Specific Antigen.

ROI: Region of Interest.

RSS: Residual Sum of Square.

TCGA: The Cancer Genome Atlas.

VAR: Vector Autoregression.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my dissertation advisor, Ali Shojaie, for his advisory and guidance throughout my Ph.D. study. His encouragement, insights and generous support have been instrumental to my development as a statistician. He has been not only an outstanding mentor, but also a great friend. I am also extremely grateful for the advices and mentorship of my dissertation committee members, Marco Carone, Mathias Drton, William Noble (Graduate School Reprsentative) and Daniela Witten.

Throughout my Ph.D. studies, I have met with many incredible advisors and collaborators, including Norman Breslow, Christiane Hampe, Christine Mac Donald and Timothy Randolph. They shared with me their perspectives, and also taught me the rigor of doing scientific research. I feel extremely blessed to have the opportunity to work with the late Norman Breslow, who was my first research advisor. Norm introduced me to modern statistical research, and taught me the art of being collaborative, yet firm on the scientific ground.

I would also like to thank Patrick Heagerty, Ken Rice, Paul Sampson, Noah Simon and Patricia Wahl for their research advices and kind encouragements.

Finally, I would like to thank Gitana Garofalo, cohort of the Biostatistics Entering Class of 2012 and members of Slab Lab, whose support and friendship have got me though these challenging, yet rewarding five years.

DEDICATION

to my supportive parents, Jingming and Lifang

Chapter 1

INTRODUCTION

1.1 Motivation

Recent technological advances have facilitated the collection of high-dimensional data in a number of scientific fields. The common feature of these data is that the number of variables for each sample, p, is usually much larger than the number of samples, n. For example, genetic datasets are usually high-dimensional, in the sense that we have measurements on more genes than individuals. Other common examples of high-dimensional data include phylogenetic data, brain imaging data, high-frequency financial market data, and internet user data.

Emerging high-dimensional data offer new opportunities for scientific discovery. For example, with high-dimensional gene sequencing data, we could find mutations of genes that are associated with the onset and progression of diseases, conditional on other deoxyribonucleic acid (DNA) sequences; using high-dimensional brain imaging data, we could find differences in brain connectivity between several hundred regions of interest (ROIs) between healthy individuals and patients with Alzheimer's disease; using high-dimensional internet user data, we could make more accurate item and content recommendations to users based on their browsing and purchasing history.

However, in addition to providing new opportunities, due to the "curse of dimensionality", high-dimensional data also pose various statistical challenges (see, e.g., Donoho, 2000, Fan and Li, 2006, Johnstone and Titterington, 2009, Fan et al., 2014a). In the past two decades, numerous methods have been proposed to estimate parameters in linear models and graphical models with high-dimensional data; see Section 1.2 for a brief introduction of recent proposals. The drawback of these estimation methods is that they only provide

point estimates, without quantifications of the uncertainty of estimates, e.g., p-values and confidence intervals, which are crucial for scientific reasoning. To address this issue, several hypothesis testing procedures for high-dimensional data have been recently developed. We present a short overview of those proposals in Section 1.3. Unfortunately, the majority of existing proposals are computationally intensive, hard to interpret, and/or have low statistical power.

In this dissertation, we propose new hypothesis testing methods for high-dimensional data and study their performances through theoretical derivations and numerical experiments.

1.2 Statistical Estimation With High-Dimensional Data

In this section, we present a brief summary of the recent development of statistical estimation methodologies in linear and graphical models with high-dimensional data.

1.2.1 High-Dimensional Linear Models

In a high-dimensional linear model,

$$\mathrm{E}\left[\boldsymbol{y}|\,\boldsymbol{X}\right] = \boldsymbol{X}\boldsymbol{\beta}, \tag{1.1}$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is a high-dimensional design matrix. Lasso (Tibshirani, 1996, Chen et al., 1998) is commonly used to estimate the regression coefficient $\boldsymbol{\beta}$. Compared to ordinary least squares (OLS), lasso further imposes an ℓ_1 penalty on the least squares criterion, i.e.,

$$\hat{\boldsymbol{\beta}}_\lambda \equiv \underset{\boldsymbol{b}\in\mathbb{R}^p}{\arg\min}\left\{\frac{1}{2n}\left\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}\right\|_2^2 + \lambda\left\|\boldsymbol{b}\right\|_1\right\}, \tag{1.2}$$

where $\lambda > 0$ is the lasso tuning parameter, and $\hat{\boldsymbol{\beta}}_\lambda$ is the lasso estimate of $\boldsymbol{\beta}$.

Theoretical properties of the lasso have been extensively studied in Osborne et al. (2000), Fu and Knight (2000), Greenshtein and Ritov (2004), Tropp (2006), Meinshausen and Bühlmann (2006), Greenshtein (2006), Zhao and Yu (2006), Rosset and Zhu (2007), Zou et al. (2007), Buena et al. (2007), Zhang and Huang (2008), Lounici (2008), Meinshausen and Yu (2009), Wainwright (2009), Bickel et al. (2009), Zhang (2009), van de Geer and Bühlmann (2009), Ye

and Zhang (2010), Dossal (2012), Tibshirani and Taylor (2012), Tibshirani (2013). Efficient computation methods for the lasso are proposed in Fu (1998), Efron et al. (2004), Friedman et al. (2007), Wu and Lange (2008), Friedman et al. (2010). Unlike earlier proposals, such as ridge regression (Hoerl and Kennard, 1970b,a), bridge regression (Frank and Friedman, 1993) and nonnegative garrote (Breiman, 1995), lasso is able to produce sparse solutions, in the sense that at most n of the p regression coefficient estimates are nonzero. This behavior greatly facilitates interpretation of the result, and makes lasso especially appropriate for the task of variable selection. However, without stringent conditions, lasso is likely unable to recover the true set of nonzero regression coefficients, $\mathcal{A} \equiv \{j : \beta_j \neq 0\}$ (Tropp, 2006, Meinshausen and Bühlmann, 2006, Zhao and Yu, 2006, Wainwright, 2009).

Several modifications of the lasso, including, to name a few, least angle regression (Efron et al., 2004), fused lasso (Tibshirani et al., 2005), elastic net (Zou and Hastie, 2005), group lasso (Yuan and Lin, 2006), adaptive lasso (Zou, 2006), Dantzig selector (Candes and Tao, 2007), graph lasso (Jacob et al., 2009), square-root lasso (Belloni et al., 2011) and scaled lasso (Sun and Zhang, 2012) have also been considered for improved estimation and prediction in various problems. In addition, Park and Casella (2008), Hans (2009, 2010) considered the Bayesian lasso regression. Although its posterior mean or median are not necessarily sparse, the Bayesian lasso provides a convenient way to quantify the uncertainty of parameters. See Tibshirani (2011) for a retrospective review of the lasso and its variants.

In addition to the lasso, which imposes an ℓ_1 penalty on the least-squares criterion, folded-concave penalties have also been considered to encourage sparsity in the estimates. They include the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001) and minimax concave penalty (MCP) (Zhang, 2010a). Due to the non-convexity of the objective function, computation with these penalties is more involved. Recent advances in computation algorithms include Hunter and Lange (2004), Hunter and Li (2005), Kim et al. (2008), Zou and Li (2008), Mazumder et al. (2011), Fan and Lv (2011). Theoretical properties of methods based on folded-concave penalization are studied in Fan and Peng (2004), Zhang (2010b), Fan and Lv (2011), Zhang and Zhang (2012), Wang et al. (2013), Zhang (2013),

Fan et al. (2014b), Wang et al. (2014). See Fan and Lv (2010) for a review of the progress in high-dimensional linear regression estimation method in the past two decades.

1.2.2 High-Dimensional Graphical Models

In Gaussian graphical models (GGMs), we assume the design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ follows a multivariate Gaussian distribution, i.e., $\boldsymbol{X} \sim_{i.i.d.} \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the population covariance matrix. Based on multivariate Gaussian theory, the support of inverse covariance matrices $\boldsymbol{\Omega} \equiv \boldsymbol{\Sigma}^{-1}$ corresponds to the conditional dependence relationship between variables. Thus, in GGM inference, we are often interested in the support of the population inverse covariance matrix.

Estimation methods for the support of the inverse covariance matrix include, among others, ℓ_1-penalized linear regression (Meinshausen and Bühlmann, 2006, Peng et al., 2009, Sun and Zhang, 2013), Dantzig-selector-based linear regression (Yuan, 2010), ℓ_1-penalized likelihood-based methods (Huang et al., 2006, Yuan and Lin, 2007, Banerjee et al., 2008, Friedman et al., 2008, Yuan, 2008, Rothman et al., 2008, Fan et al., 2009, Lam and Fan, 2009, Ravikumar et al., 2011, Hsieh et al., 2011), SCAD-based penalized likelihood methods (Fan et al., 2009, Lam and Fan, 2009) and constrained ℓ_1-minimization methods (Cai et al., 2011); see Drton and Maathuis (2017) for a review on GGM estimation methods. A number of recent proposals have also considered joint estimation of multiple GGMs in related populations. By leveraging similarities among parameters in multiple populations, joint GGM estimation methods can result in improved estimation performance of GGMs; they include Guo et al. (2011), Mohan et al. (2014), Danaher et al. (2014), Zhao et al. (2014), Peterson et al. (2015), Saegusa and Shojaie (2016), Cai et al. (2016),

1.3 Hypothesis Testing With High-Dimensional Data

In this section, we review recent proposals on hypothesis testing procedures with high-dimensional data.

For the linear model, early efforts to devise formal high-dimensional hypothesis tests for

$\boldsymbol{\beta}$ include least squares kernel machine (LSKM; Liu et al., 2007, Kwee et al., 2007) and sequence kernel association test (SKAT; Wu et al., 2010, 2011, Lee et al., 2012b,a, Ionita-Laza et al., 2013). These methods examine multiple entries in $\boldsymbol{\beta}$ together and are hence not appropriate to identify individual variables that are conditionally associated with the outcome. Early work on inference procedures for individual entries in $\boldsymbol{\beta}$ focused on bootstrapping (Bach, 2008, Chatterjee and Lahiri, 2011, 2013), and subsampling to control the family-wise error rate (FWER) (Meinshausen and Bühlmann, 2010, Shah and Samworth, 2013). However, bootstrapping and subsampling are computationally expensive, and their finite sample properties may not be desirable. More recently, two additional classes of high-dimensional tests have been proposed to examine the conditional association of individual variables using recent theoretical development for the lasso. The first class consists of debiased tests (Javanmard and Montanari, 2013b,a, Bühlmann, 2013, Zhang and Zhang, 2014, van de Geer et al., 2014, Javanmard and Montanari, 2014a,b, Ning and Liu, 2017), while the second consists of post-selection inference procedures (Wasserman and Roeder, 2009, Meinshausen et al., 2009, Weinstein et al., 2013, Berk et al., 2013, Belloni et al., 2014, Lee et al., 2016, Tibshirani et al., 2016). We describe these latter two classes of tests in more detail in Chapter 2.

Early work on high-dimensional hypothesis testing procedures for GGMs focused on the population covariance matrix $\boldsymbol{\Sigma}$. For example, the proposals in Ledoit and Wolf (2002), Srivastava (2005), Birke and Dette (2005), Bai et al. (2009), Chen et al. (2010), Jiang et al. (2012), Cai and Ma (2013) examine whether some specific entries in the population covariance matrix are equal to zero; the proposals in Schott (2007), Srivastava and Yanagihara (2010), Li and Chen (2012), Cai et al. (2013), Cai and Zhang (2016), on the other hand, focus on testing whether the covariance matrices in two populations are equal. More recently, inference on the inverse covariance matrix $\boldsymbol{\Omega}$ has also been considered. Specifically, Ren et al. (2015), Janková and van de Geer (2015, 2017), Xia and Li (2017) examine a single inverse covariance matrix, whereas Xia et al. (2015), Belilovsky et al. (2016) test the equality of two inverse covariance matrices.

Despite promising improvements over the last decade, existing procedures suffer from a

few weaknesses, which we will try to address in this dissertation.

1. Most of the hypothesis testing proposals described above are complicated and computationally intensive. In Chapter 2, we consider a simple and naïve two-step procedure for this task, in which we (i) fit a lasso model in order to obtain a subset of the variables; and (ii) fit an OLS model on the lasso-selected set. Conventional statistical wisdom tells us that we cannot make use of the standard statistical inference tools for the resulting OLS model (such as t-tests) since we have peeked at the data twice: once in running the lasso, and again in fitting the OLS model. However, in Chapter 2, we show that under certain assumptions, with high probability, the set of variables selected by the lasso is deterministic. Consequently, the naïve two-step approach can yield confidence intervals that have asymptotically correct coverage, as well as p-values with proper type-I error control. Furthermore, this two-step approach unifies two existing camps of work on high-dimensional inference: recent research on inference based on a sub-model selected by the lasso (lasso post selection inference), and research focusing on inference using a debiased version of the lasso estimator (lasso debiased tests).

2. Graph-constrained estimation methods, which incorporate external graphical information on the relationship between variables, can result in more accurate estimates, especially in high dimensions. Unfortunately, such valuable information is usually ignored in hypothesis testing. In Chapter 3, we present a new inference framework, called the Grace test, which produces coefficient estimates and corresponding p-values while incorporating the external graph information. We show, both theoretically and numerically, that the proposed method asymptotically controls the type-I error rate regardless of the informativeness of the graph. We also show that when the underlying graph is informative, the Grace test is asymptotically more powerful than similar tests that ignore the external information. We study the power properties of the proposed test when the graph is not fully informative and develop a more robust Grace-ridge test for such settings. Our numerical studies show that as long as the graph is reasonably

informative, the proposed inference procedures deliver improved statistical power over existing methods that ignore external information.

3. In Chapter 4, we focus on testing whether two Gaussian conditional dependence networks are the same. Existing methods try to accomplish this goal by either directly comparing their estimated structures without taking into account their uncertainty, or testing the null hypothesis that their inverse correlation matrices are equal. However, high-dimensional GGM estimation approaches rarely provide measures of uncertainty, e.g., p-values, which are crucial in drawing scientific conclusions, whereas the testing approaches might lead to misleading results in some cases, as we illustrate in Chapter 4. To address these shortcomings, we propose a two-step hypothesis testing framework, which tests the null hypothesis that *the edge sets in two networks are the same*. Our framework is especially appropriate if the goal is to identify nodes or edges that exhibit differential connectivity. We show, using theoretical and numerical analyses, that our proposal controls the type-I error rate. We demonstrate the applicability of our method in brain imaging and cancer genetics datasets.

1.4 Notations and Conventions

In this section, we introduce some of the notations used throughout the dissertation.

We use bold upper case fonts to denote matrices, bold lower case fonts to denote vectors, and normal fonts to denote scalars. Specifically, we use $\boldsymbol{I}$, $\mathbf{1}$ and $\mathbf{0}$ to denote the identity matrix, and vectors of ones and of zeros, respectively. For two symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we use $\boldsymbol{A} \preceq \boldsymbol{B}$ to denote that $\boldsymbol{B} - \boldsymbol{A}$ is positive semi-definite, and use $\boldsymbol{A} \prec \boldsymbol{B}$ to denote that $\boldsymbol{B} - \boldsymbol{A}$ is positive definite; we use $\phi^2_{\min}[\boldsymbol{A}]$ and $\phi^2_{\max}[\boldsymbol{A}]$ to denote its minimum and maximum eigenvalues, respectively. For any vector $\boldsymbol{b}$ and any matrix $\boldsymbol{M}$, $\|\boldsymbol{b}\|_k = (\sum_j b_j^k)^{1/k}$, $\|\boldsymbol{M}\|_k = \sup_{\boldsymbol{x}:\|\boldsymbol{x}\|_k=1} \|\boldsymbol{M}\boldsymbol{x}\|_k$ for $1 \leq k < \infty$; $\|\boldsymbol{b}\|_0 = |\text{supp}(\boldsymbol{b})|$, $\|\boldsymbol{b}\|_\infty = \max_j\{b_j\}$, $\|\boldsymbol{M}\|_\infty = \max_i\{\sum_j |M_{(i,j)}|\}$ and $\|\boldsymbol{M}\|_{\max} = \max_{i,j}\{|M_{(i,j)}|\}$. $\text{supp}(\cdot)$, $\text{sign}(\cdot)$, $\text{diag}(\cdot)$, $\text{tr}(\cdot)$ and $\text{vec}(\cdot)$ are used to denote the support of a vector, the vector of signs of entries in a vector, the

vector of diagonal entries in a matrix, the trace of a matrix and the vectorization of a matrix, respectively.

We use calligraphy fonts to denote sets. Given a set $\mathcal{S}$, $|\mathcal{S}|$ denotes its cardinality. We use "$\setminus$" and "$\triangle$" to denote the asymmetric and symmetric difference operators of two sets, respectively: $\mathcal{S}_1 \setminus \mathcal{S}_2 = \mathcal{S}_1 \cap \mathcal{S}_2^c$, $\setminus \mathcal{S}_2 = \mathcal{S}_2^c$ and $\mathcal{S}_1 \triangle \mathcal{S}_2 = (\mathcal{S}_1 \cup \mathcal{S}_2) \setminus (\mathcal{S}_1 \cap \mathcal{S}_2)$. For any vector $\boldsymbol{b}$, matrix $\boldsymbol{M}$, and index sets $\mathcal{S}_1$ and $\mathcal{S}_2$, we use $\boldsymbol{b}_{\mathcal{S}_1}$ to denote the sub-vector of $\boldsymbol{b}$ comprised of elements of $\mathcal{S}_1$, $\boldsymbol{M}_{\mathcal{S}_2}$ to denote the sub-matrix of $\boldsymbol{M}$ with columns of $\mathcal{S}_2$, and $\boldsymbol{M}_{(\mathcal{S}_1,\mathcal{S}_2)}$ to denote the sub-matrix of $\boldsymbol{M}$ with rows of $\mathcal{S}_1$ and columns of $\mathcal{S}_2$.

We use "$\equiv$" to denote equalities by definition, and "$\asymp$", "$\precsim$", "$\prec$","$\succsim$" and "$\succ$" to denote asymptotic orders, i.e., for any two sequences $\{a_n\}$ and $\{b_n\}$, $a_n \asymp b_n$ if $0 < \lim_{n\to\infty} a_n/b_n < \infty$, $a_n \precsim b_n$ if $\lim_{n\to\infty} a_n/b_n < \infty$ and $a_n \prec b_n$ if $\lim_{n\to\infty} a_n/b_n = 0$. We use "$\vee$" and "$\wedge$" denote the maximum and minimum of two real numbers, respectively; for a real number $a \in \mathbb{R}$, $a_+ \equiv a \vee 0$. We use $\mathbb{1}[\cdot]$ for the indicator function, and $\Phi_{\mathcal{N}}[\cdot]$ and $\Phi_{\mathcal{N}}^{-1}[\cdot]$ for the cumulative distribution function (CDF) and the quantile function of standard normal distribution, respectively. We use "$\perp\!\!\!\perp$" to denote independence of two random variables.

For the linear model

$$\mathrm{E}\left[\boldsymbol{y} \middle| \boldsymbol{X}\right] = \boldsymbol{X}\boldsymbol{\beta}, \tag{1.3}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the p-vector of coefficients, we use bar-symbols, e.g., $\bar{\boldsymbol{\beta}}$, to denote OLS estimates of $\boldsymbol{\beta}$:

$$\bar{\boldsymbol{\beta}} \equiv \underset{\boldsymbol{b}\in\mathbb{R}^p}{\arg\min} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 \right\} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}; \tag{1.4}$$

hat-symbols, e.g., $\hat{\boldsymbol{\beta}}_\lambda$, denote lasso estimates (Tibshirani, 1996, Chen et al., 1998) of $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}_\lambda \equiv \underset{\boldsymbol{b}\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + \lambda \|\boldsymbol{b}\|_1 \right\}; \tag{1.5}$$

check-symbols, e.g., $\check{\boldsymbol{\beta}}_\lambda$, are associated with the noiseless (and thus deterministic) counterpart of lasso estimates, $\hat{\boldsymbol{\beta}}_\lambda$:

$$\check{\boldsymbol{\beta}}_\lambda \equiv \underset{\boldsymbol{b}\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n} \mathrm{E}\left[\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 \right] + \lambda \|\boldsymbol{b}\|_1 \right\}. \tag{1.6}$$

Due to the presence of expectation, $\check{\boldsymbol{\beta}}_\lambda$ is usually unknown; it is mainly used as a tool to analyze the behavior of $\hat{\boldsymbol{\beta}}_\lambda$. Finally, we use tilde-symbols, e.g., $\tilde{\boldsymbol{\beta}}^{\boldsymbol{L}}_{\lambda_{\boldsymbol{L}}}$ and $\tilde{\boldsymbol{\beta}}^{\boldsymbol{L},\boldsymbol{I}}_{\lambda_{\boldsymbol{L}},\lambda_{\boldsymbol{I}}}$, to denote ridge-like estimates (Hoerl and Kennard, 1970b,a) of $\boldsymbol{\beta}$, e.g.,

$$\tilde{\boldsymbol{\beta}}^{\boldsymbol{L}}_{\lambda_{\boldsymbol{L}}} \equiv \underset{\boldsymbol{b}\in\mathbb{R}^p}{\arg\min} \left\{ \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}\|_2^2 + \lambda_{\boldsymbol{L}} \boldsymbol{b}^\top \boldsymbol{L}\boldsymbol{b} \right\}, \tag{1.7}$$

$$\tilde{\boldsymbol{\beta}}^{\boldsymbol{L},\boldsymbol{I}}_{\lambda_{\boldsymbol{L}},\lambda_{\boldsymbol{I}}} \equiv \underset{\boldsymbol{b}\in\mathbb{R}^p}{\arg\min} \left\{ \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}\|_2^2 + \lambda_{\boldsymbol{L}} \boldsymbol{b}^\top \boldsymbol{L}\boldsymbol{b} + \lambda_{\boldsymbol{I}} \boldsymbol{b}^\top \boldsymbol{I}\boldsymbol{b} \right\}. \tag{1.8}$$

In Chapter 2, we also use symbols which are superscripted by a set surrounded by round parentheses, e.g., $\boldsymbol{\beta}^{(\mathcal{S})}$, $\bar{\boldsymbol{\beta}}^{(\mathcal{S})}$ and $\hat{\boldsymbol{\beta}}^{(\mathcal{S})}$. These notations are used to denote regression coefficients and estimates in the sub-model $(\boldsymbol{y}, \boldsymbol{X}_{\mathcal{S}})$, e.g.,

$$\bar{\boldsymbol{\beta}}^{(\mathcal{S})} \equiv \underset{\boldsymbol{b}\in\mathbb{R}^{|\mathcal{S}|}}{\arg\min} \left\{ \|\boldsymbol{y}-\boldsymbol{X}_{\mathcal{S}}\boldsymbol{b}\|_2^2 \right\}, \tag{1.9}$$

$$\hat{\boldsymbol{\beta}}^{(\mathcal{S})}_{\lambda} \equiv \underset{\boldsymbol{b}\in\mathbb{R}^{|\mathcal{S}|}}{\arg\min} \left\{ \frac{1}{2n} \|\boldsymbol{y}-\boldsymbol{X}_{\mathcal{S}}\boldsymbol{b}\|_2^2 + \lambda \|\boldsymbol{b}\|_1 \right\}. \tag{1.10}$$

These notations should not be confused with $\boldsymbol{\beta}_{\mathcal{S}}$, $\bar{\boldsymbol{\beta}}_{\mathcal{S}}$ and $\hat{\boldsymbol{\beta}}_{\mathcal{S}}$, which represent entries of $\mathcal{S}$ in the full-model $(\boldsymbol{y}, \boldsymbol{X})$ regression coefficients and estimates.

Chapter 2

IN DEFENSE OF THE INDEFENSIBLE: A VERY NAÏVE APPROACH TO HIGH-DIMENSIONAL INFERENCE

2.1 Introduction

In this chapter, we consider the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p]$ is an $n \times p$ deterministic design matrix, $\boldsymbol{\epsilon}$ is a vector of independent and identically distributed (i.i.d.) errors with $\mathrm{E}[\epsilon_i] = 0$ and $\mathrm{Var}[\epsilon_i] = \sigma^2_{\boldsymbol{\epsilon}}$, and $\boldsymbol{\beta}$ is a p-vector of regression coefficients. Without loss of generality, we assume that the columns of $\boldsymbol{X}$ are centered and standardized, such that $\mathbf{1}^\top \boldsymbol{X}_k = 0$ and $\boldsymbol{X}_k^\top \boldsymbol{X}_k = n$ for $k = 1, \ldots, p$.

When the number of variables p is much smaller than the sample size n, estimation and inference for the vector $\boldsymbol{\beta}$ are straightforward. For instance, estimation can be performed using OLS, i.e., (1.4), and inference can be conducted using classical approaches, such as t-tests and F-tests (see, e.g., Draper and Smith, 1998, Gelman and Hill, 2006, Weisberg, 2013).

As the scope and scale of data collection have increased across virtually all fields, there is an increase in datasets that are *high-dimensional*, in the sense that the number of variables, p, is larger than the number of observations, n (see, e.g., Clarke et al., 2008). In this setting, classical approaches for estimation and inference of $\boldsymbol{\beta}$ cannot be directly applied. In the past 20 years, a vast statistical literature has focused on estimating $\boldsymbol{\beta}$ in high dimensions. In particular, penalized regression methods, such as the lasso (Tibshirani, 1996), defined in (1.2), can be used to estimate $\boldsymbol{\beta}$. However, the topic of inference in the high-dimensional setting remains relatively unexplored, despite promising recent work in this area (see, e.g.,

Taylor and Tibshirani, 2015, Dezeure et al., 2015). Roughly speaking, recent work on inference in the high-dimensional setting falls into two classes: (i) methods that examine the null hypothesis $H_{0,j} : \beta_j = 0$; and (ii) methods that make inference based on a sub-model. We will review these two classes of methods in turn.

First, we review methods that examine the null hypothesis $H_{0,j} : \beta_j = 0$, i.e., the variable $\boldsymbol{X}_j$ is unassociated with the outcome $\boldsymbol{y}$, conditional on *all other variables*. It might be tempting to estimate $\boldsymbol{\beta}$ using the lasso in (1.2), and then (for instance) to construct a confidence interval around the lasso estimate $\hat{\beta}_{\lambda,j}$. Unfortunately, such an approach is problematic, because $\hat{\boldsymbol{\beta}}_\lambda$ is a biased estimate of $\boldsymbol{\beta}$. To remedy this problem, we can apply a one-step adjustment to $\hat{\boldsymbol{\beta}}_\lambda$, such that under appropriate assumptions, the resulting *debiased estimate* is asymptotically unbiased for $\boldsymbol{\beta}$. Then, p-values and confidence intervals can be constructed around this debiased estimate. For example, such an approach is taken by the low dimensional projection estimator (LDPE; Zhang and Zhang, 2014, van de Geer et al., 2014), the debiased lasso test with unknown population covariance (SSLasso; Javanmard and Montanari, 2013a, 2014a), the debiased lasso test with known population covariance (SDL; Javanmard and Montanari, 2014b), and the decorrelated score test (dScore; Ning and Liu, 2017). See Dezeure et al. (2015) for a review of such procedures. In what follows, we will refer to these and related approaches for testing $H_{0,j} : \beta_j = 0$ as *debiased lasso tests*.

Now, we review recent work that makes statistical inference based on a sub-model. Recall that the challenge in high dimensions stems from the fact that when $p > n$, classical statistical methods cannot be applied; for instance, we cannot even perform OLS. This suggests a simple approach: given an index set $\mathcal{M} \subseteq \{1, \ldots, p\}$, we can consider performing inference *based on the sub-model* composed only of the variables in the index set $\mathcal{M}$. That is, rather than considering the model (2.1), we consider the sub-model

$$\boldsymbol{y} = \boldsymbol{X}_\mathcal{M} \boldsymbol{\beta}^{(\mathcal{M})} + \boldsymbol{\epsilon}^{(\mathcal{M})}. \tag{2.2}$$

In (2.2), the notation $\boldsymbol{\beta}^{(\mathcal{M})}$ and $\boldsymbol{\epsilon}^{(\mathcal{M})}$ emphasizes that the true sub-model regression coefficients and corresponding noise are functions of the set $\mathcal{M}$.

Now, provided that $|\mathcal{M}| < n$, we can perform estimation and inference on the vector $\boldsymbol{\beta}^{(\mathcal{M})}$ using classical statistical approaches. For instance, we can consider building confidence intervals $\text{CI}_j^{(\mathcal{M})}$ such that for any $j \in \mathcal{M}$,

$$\Pr\left[\beta_j^{(\mathcal{M})} \in \text{CI}_j^{(\mathcal{M})}\right] \geq 1 - \alpha. \tag{2.3}$$

At first blush, the problems associated with high dimensionality have been solved!

Of course, there are some problems with the aforementioned approach. The first problem is that the coefficients in the sub-model (2.2) typically are not the same as the coefficients in the original model (2.1) (see, e.g., Berk et al., 2013). Roughly speaking, the problem is that the coefficients in the model (2.1) quantify the linear association between a given variable and the outcome, *conditional on the other* $p-1$ *variables*, whereas the coefficients in the model (2.2) quantify the linear association between a variable and the outcome, *conditional on the other* $|\mathcal{M}|-1$ *variables in the sub-model*. The true regression coefficients in the sub-model are of the form

$$\boldsymbol{\beta}^{(\mathcal{M})} = \left(\boldsymbol{X}_{\mathcal{M}}^\top \boldsymbol{X}_{\mathcal{M}}\right)^{-1} \boldsymbol{X}_{\mathcal{M}}^\top \boldsymbol{X}\boldsymbol{\beta}. \tag{2.4}$$

Thus, $\boldsymbol{\beta}^{(\mathcal{M})} \neq \boldsymbol{\beta}_{\mathcal{M}}$ unless $\boldsymbol{X}_{\mathcal{M}}^\top \boldsymbol{X}_{\mathcal{M}^c}\boldsymbol{\beta}_{\mathcal{M}^c} = \boldsymbol{0}$. To see this more concretely, consider the following example with $p = 4$ deterministic variables. We let

$$\boldsymbol{G} \equiv \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X} = \begin{bmatrix} 1 & 0 & 0.6 & 0 \\ 0 & 1 & 0.6 & 0 \\ 0.6 & 0.6 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Set $\boldsymbol{\beta} = [1, 1, 0, 0]^\top$. If $\mathcal{M} = \{2, 3\}$, then it is easy to verify that

$$\begin{aligned}
\boldsymbol{\beta}^{(\mathcal{M})} &= \left[\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right]^{-1}_{(\{2,3\},\{2,3\})}\left[\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right]_{(\{2,3\},\{1,2,3,4\})}\boldsymbol{\beta} \\
&= \begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}^{-1}\begin{bmatrix} 0 & 1 & 0.6 & 0 \\ 0.6 & 0.6 & 1 & 0 \end{bmatrix}\begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix}^\top \\
&= \begin{bmatrix} 0.4375 \\ 0.9375 \end{bmatrix} \neq \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \boldsymbol{\beta}_{\mathcal{M}}.
\end{aligned}$$

The second problem that arises in restricting our attention to the sub-model (2.2) is that in practice, the index set $\mathcal{M}$ is not pre-specified. Instead, it is typically chosen based on the data. For instance, one might take $\mathcal{M}$ to equal the support of the lasso estimate,

$$\hat{\mathcal{A}}_\lambda \equiv \text{supp}(\hat{\boldsymbol{\beta}}_\lambda) \equiv \left\{j : \hat{\beta}_{\lambda,j} \neq 0\right\}. \tag{2.5}$$

The problem is that if we construct the index set $\mathcal{M}$ based on the data, and then apply classical inference approaches on the vector $\boldsymbol{\beta}^{(\mathcal{M})}$, the resulting p-values and confidence intervals will not be valid (see, e.g., Pötscher, 1991, Kabaila, 1998, Leeb and Pötscher, 2003, 2005, 2006a,b, 2008, Kabaila, 2009, Berk et al., 2013). This is because we peeked at the data twice: once to determine which variables to include in $\mathcal{M}$, and then again to test hypotheses associated with those variables. Consequently, an extensive recent body of literature has focused on the task of performing inference on $\boldsymbol{\beta}^{(\mathcal{M})}$ in (2.2) given that $\mathcal{M}$ was chosen based on the data. Cox (1975) proposed the idea of sample-splitting to break up the dependence of variable selection and hypothesis testing, whereas Wasserman and Roeder (2009) studied this proposal in application to the lasso, marginal regression and forward step-wise regression. Meinshausen et al. (2009) extended the single-splitting proposal of Wasserman and Roeder (2009) to multi-splitting, which improved statistical power and reduced the number of falsely selected variables. Berk et al. (2013) instead considered simultaneous inference, which is univesally valid under all possible model selection procedures without sample-splitting. More recently, Lee et al. (2016), Tibshirani et al. (2016) studied the geometry of the lasso and sequential regression, respectively, and proposed post-selection inference methods conditional

on the random set of selected variables. See Taylor and Tibshirani (2015) for a review of post-selection inference procedures.

In a recent *Statistical Science* paper, Leeb et al. (2015) performed a simulation study, in which they obtained a set $\mathcal{M}$ using variable selection, and then calculated *"naïve" confidence intervals* for $\boldsymbol{\beta}^{(\mathcal{M})}$ using OLS, *without accounting for the fact that the set* $\mathcal{M}$ *was chosen based on the data*. Of course, conventional statistical wisdom dictates that the resulting confidence intervals will be much too narrow. In fact, this is what Leeb et al. (2015) found, when they used best subset selection to construct the set $\mathcal{M}$. However, surprisingly, when the lasso was used to construct the set $\mathcal{M}$, the confidence intervals had approximately correct coverage, in the sense that (2.3) holds. This is in stark contrast to the existing literature!

In this chapter, we present a theoretical justification for the empirical finding in Leeb et al. (2015). We show that selecting a set $\mathcal{M}$ based on the lasso and then constructing naïve confidence intervals based on the selected set could lead to asymptotically valid inference of the vector $\boldsymbol{\beta}^{(\mathcal{M})}$. Furthermore, to make inference on $\boldsymbol{\beta}$, we utilize our theoretical findings in order to develop the *naïve score test*, a simple procedure for testing the null hypothesis $H_{0,j} : \beta_j = 0$ for $j = 1, \ldots, p$.

The rest of this chapter is organized as follows. In Sections 2.2 and 2.3, we focus on post-selection inference: we seek to perform inference on $\boldsymbol{\beta}^{(\mathcal{M})}$ in (2.2), where $\mathcal{M}$ is selected based on the lasso, i.e., $\mathcal{M} = \hat{\mathcal{A}}_\lambda$, where $\hat{\mathcal{A}}_\lambda$ is defined in (2.5). In Section 2.2, we point out a previously overlooked fact: although $\hat{\mathcal{A}}_\lambda$ is random, it converges in probability to a deterministic and non-data-dependent set. This result implies that we can use classical methods for inference on $\boldsymbol{\beta}^{(\mathcal{M})}$, when $\mathcal{M} = \hat{\mathcal{A}}_\lambda$. In Section 2.3, we provide empirical evidence in support of these theoretical findings. In Sections 2.4 to 2.5, we instead focus on the task of performing inference on $\boldsymbol{\beta}$ in (2.1). We propose the naïve score test in Section 2.4, and study its empirical performance in Section 2.5. In Section 2.6, we establish that our naïve post-selection inference procedures are in fact closely related to debiased lasso tests, and thus unify the frameworks of lasso post-selection inference and debiased lasso tests. We end with a discussion of future research directions in Section 2.7. Technical proofs are collected

in Appendix A.

2.2 Theoretical Justification for Naïve Confidence Intervals

Recall that the sub-model regression coefficient $\boldsymbol{\beta}^{(\hat{\mathcal{A}}_\lambda)}$ was defined in (2.4). The simulation results of Leeb et al. (2015) suggest that if we perform OLS using the variables contained in the support set of the lasso, $\hat{\mathcal{A}}_\lambda$, then the classical confidence intervals associated with the OLS estimator,

$$\bar{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda)} \equiv \left(\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}\right)^{-1} \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top \boldsymbol{y}, \tag{2.6}$$

have approximately correct coverage, where correct coverage means that for all $j \in \hat{\mathcal{A}}_\lambda$,

$$\Pr\left[\beta_j^{(\hat{\mathcal{A}}_\lambda)} \in \mathrm{CI}_j^{(\hat{\mathcal{A}}_\lambda)}\right] \geq 1 - \alpha. \tag{2.7}$$

We reiterate that in (2.7), $\mathrm{CI}_j^{(\hat{\mathcal{A}}_\lambda)}$ is the confidence interval output by standard OLS software applied to the data $(\boldsymbol{y}, \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda})$. This goes against our statistical intuition: it seems that by fitting a lasso model and then performing OLS on the selected set, we are peeking at the data twice, and thus we would expect the confidence interval $\mathrm{CI}_j^{(\hat{\mathcal{A}}_\lambda)}$ to be much too narrow.

In this section, we present a theoretical result that suggests that, in fact, this "double-peeking" might not be so bad. Our key insight is as follows: under appropriate assumptions, *the set of variables selected by the lasso is deterministic and non-data-dependent with high probability*. Thus, fitting an OLS model on the variables selected by the lasso does not really constitute peeking at the data twice: effectively, with high probability, we are only peeking at them once. This means that the naïve confidence intervals obtained from OLS will have approximately correct coverage, in the sense of (2.7).

We first introduce sufficient conditions for our theoretical result.

(**M1**) The design matrix $\boldsymbol{X}$ is deterministic, with columns in *general position*; see Definition 2.2.1. Furthermore, the columns of $\boldsymbol{X}$ are centered and standardized, i.e., for any $j = 1, \ldots, p$, $\mathbf{1}^\top \boldsymbol{X}_j = 0$, $\boldsymbol{X}_j^\top \boldsymbol{X}_j = n$.

(**M2**) The error $\boldsymbol{\epsilon}$ in (2.1) has i.i.d. entries and sub-Gaussian tails. That is, there exist some constant $h > 0$ such that for all $x > 0$, we have $\Pr[|\epsilon_i| > x] < \exp(1 - hx^2)$, for all $i = 1, \ldots, n$.

(**M3**) The sample size n, dimension p and lasso tuning parameter λ satisfy

$$\sqrt{\frac{\log(p)}{n}}\frac{1}{\lambda} \to 0. \tag{2.8}$$

(**M4**) Let $\mathcal{A} \equiv \text{supp}(\boldsymbol{\beta})$, and $\mathcal{S} \equiv \{j : |\beta_j| > 3\lambda\sqrt{q}/\phi^2\}$, where $q \equiv |\mathcal{A}| \equiv |\text{supp}(\boldsymbol{\beta})|$, and ϕ is defined in (**E**). The signal strength satisfies

$$\left\|\boldsymbol{\beta}_{\check{\mathcal{A}}_\lambda \setminus \mathcal{S}}\right\|_\infty = \mathcal{O}\left(\sqrt{\frac{\log(p)}{n}}\right), \tag{2.9}$$

and

$$\left\|\boldsymbol{X}_{\mathcal{A}\setminus(\check{\mathcal{A}}_\lambda \cup \mathcal{S})}\boldsymbol{\beta}_{\mathcal{A}\setminus(\check{\mathcal{A}}_\lambda \cup \mathcal{S})}\right\|_2 = \mathcal{O}\left(\sqrt{\log(p)}\right), \tag{2.10}$$

where $\check{\mathcal{A}}_\lambda \equiv \text{supp}(\check{\boldsymbol{\beta}}_\lambda)$, with $\check{\boldsymbol{\beta}}_\lambda$ defined in (1.6).

(**E**) Let $\check{q}_\lambda \equiv |\check{\mathcal{A}}_\lambda|$. There exists a constant $\phi^2 > 0$, such that for $q \geq 1$, and for any index set $\mathcal{I}$ with $|\mathcal{I}| \leq q \vee \check{q}_\lambda$, and all $\boldsymbol{a} \in \mathbb{R}^p$ that satisfy $\|\boldsymbol{a}_{\setminus\mathcal{I}}\|_1 \leq \|\boldsymbol{a}_{\mathcal{I}}\|_1$,

$$\liminf_{n\to\infty} \frac{\|\boldsymbol{X}\boldsymbol{a}\|_2^2}{n\|\boldsymbol{a}_{\mathcal{B}}\|_2^2} \geq \phi^2 > 0, \tag{2.11}$$

for any index set $\mathcal{B}$ such that $\mathcal{B} \supseteq \mathcal{I}$, $|\mathcal{B}\setminus\mathcal{I}| \leq q \vee \check{q}_\lambda$ and $\|\boldsymbol{a}_{\setminus\mathcal{B}}\|_\infty \leq \min_{j\in(\mathcal{B}\setminus\mathcal{I})} |a_j|$.

(**T**) Define the scaled Gramian matrix $\boldsymbol{G} \equiv \boldsymbol{X}^\top\boldsymbol{X}/n$, and the sub-gradient of the noiseless lasso $\check{\boldsymbol{\tau}}_\lambda$ based on the stationary condition (Karush, 1939, Kuhn and Tucker, 1951) of (1.6),

$$\check{\boldsymbol{\tau}}_\lambda = \frac{1}{\lambda}\boldsymbol{G}\left(\boldsymbol{\beta} - \check{\boldsymbol{\beta}}_\lambda\right). \tag{2.12}$$

Then,

$$\limsup_{n\to\infty} \left\|\check{\boldsymbol{\tau}}_{\lambda,\setminus\check{\mathcal{A}}_\lambda}\right\|_\infty \leq 1 - \delta \quad \text{and} \quad \frac{\sqrt{\log(p)/n}/\lambda}{\min_{j\in\check{\mathcal{A}}_\lambda\setminus\mathcal{S}}\left|\left(\boldsymbol{G}_{(\check{\mathcal{A}}_\lambda,\check{\mathcal{A}}_\lambda)}\right)^{-1}\check{\boldsymbol{\tau}}_{\lambda,\check{\mathcal{A}}_\lambda}\right|_j} \to 0,$$

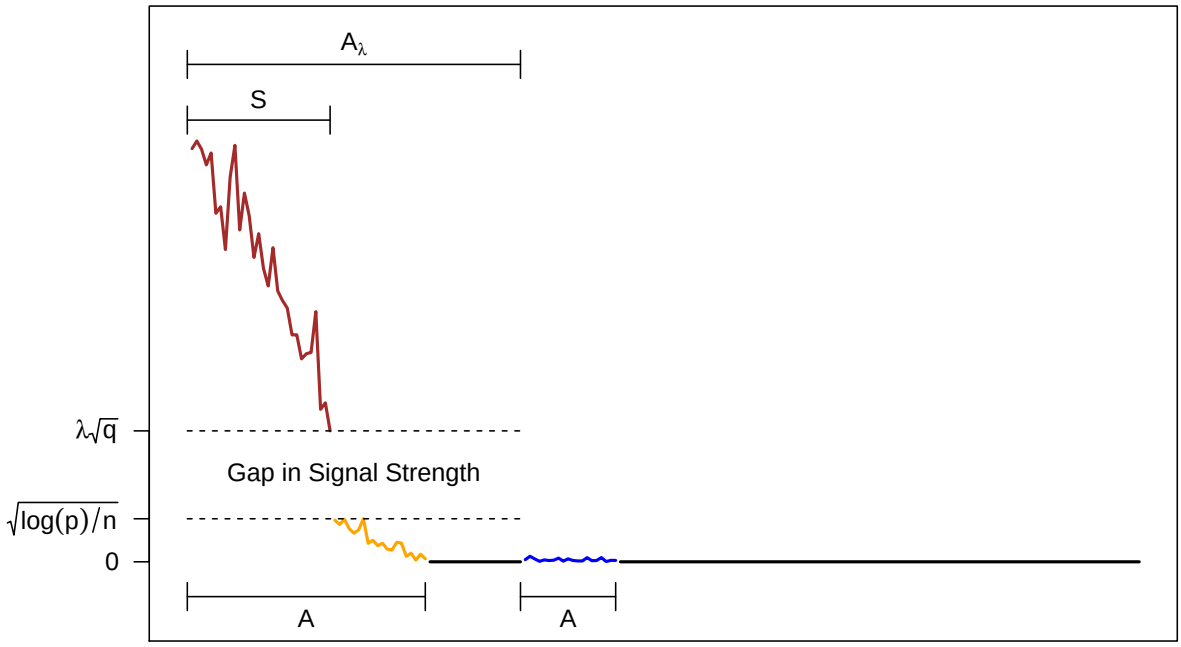
such that $\sqrt{\log(p)/n}/(\lambda\delta) \to 0$.

Definition 2.2.1 (Definition 1.1 in Dossal (2012)). *Let* $(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^p)$ *be* p *points in* $\mathbb{R}^n$. *These points are said in general position if all affine subspaces of* $\mathbb{R}^n$ *of dimension* $k < n \vee p$ *contain at most* $k+1$ *points in* $(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^p)$. *Columns of matrix* $\boldsymbol{X}$ *are in general position if for all sign vectors* $\boldsymbol{s} \in \{-1, 1\}^p$, *points* $(s_1\boldsymbol{X}_1, \ldots, s_p\boldsymbol{X}_p)$ *are in general position.*

Condition (**M1**), presented in Rosset et al. (2004), Dossal (2012), Tibshirani (2013), is a mild assumption that guarantees the uniqueness of $\check{\boldsymbol{\beta}}_\lambda$ and $\hat{\boldsymbol{\beta}}_\lambda$. Condition (**M2**) enables the dimension p to grow at an exponential rate relative to the sample size n, i.e., $p = \mathcal{O}(\exp(n))$. Condition (**M3**) requires the lasso tuning parameter λ to approach zero at a slightly slower rate than the ℓ_2 estimation and prediction consistent rate $\lambda \asymp \sqrt{\log(p)/n}$ (see, e.g., Bickel et al., 2009, van de Geer and Bühlmann, 2009); this helps further control the randomness of the error $\boldsymbol{\epsilon}$. Unfortunately, this condition complicates the task of tuning parameter selection; we further discuss this issue in Sections 2.3 and 2.7.

In (**M4**), the requirements that $\mathcal{S} \equiv \{j : |\beta_j| > 3\lambda\sqrt{q}/\phi^2\}$ and $\|\boldsymbol{\beta}_{\check{\mathcal{A}}_\lambda \setminus \mathcal{S}}\|_\infty = \mathcal{O}(\sqrt{\log(p)/n})$ indicate that the regression coefficients of variables in $\check{\mathcal{A}}_\lambda$ are either asymptotically no smaller than $\lambda\sqrt{q}$, or else no larger than $\sqrt{\log(p)/n}$. Given that λ is asymptotically slightly larger than $\sqrt{\log(p)/n}$ by (**M3**), these requirements imply that there needs to be a gap in signal strength of order at least $\sqrt{q}$ between the *strong signal variables* (those in $\mathcal{S}$) and the *weak signal* variables (those in $\mathcal{A}\setminus\mathcal{S}$). We note that this is substantially milder than the $\boldsymbol{\beta}$-min condition that is commonly used to establish model selection consistency (i.e., $\Pr[\hat{\mathcal{A}}_\lambda = \mathcal{A}] \to 1$) or the variable screening property (i.e., $\Pr[\hat{\mathcal{A}}_\lambda \supseteq \mathcal{A}] \to 1$) of the lasso (see, e.g., Meinshausen and Bühlmann, 2006, Zhao and Yu, 2006, Wainwright, 2009, Bühlmann and van de Geer, 2011), which does not allow for the presence of weak signal variables. If we impose additional stringent conditions on the design matrix (Buena et al., 2007, Zhang, 2009, Candès and Plan, 2009), we could allow for the signal strength of $\boldsymbol{\beta}_{\check{\mathcal{A}}_\lambda \setminus \mathcal{S}}$ to achieve $\sqrt{\log(p)q/n}$ in magnitude, in which case there is no gap in signal strength. $\|\boldsymbol{X}_{\mathcal{A}\setminus(\check{\mathcal{A}}_\lambda \cup \mathcal{S})}\boldsymbol{\beta}_{\mathcal{A}\setminus(\check{\mathcal{A}}_\lambda \cup \mathcal{S})}\|_2 = \mathcal{O}(\sqrt{\log(p)})$ of (**M4**) implies that the total signal strength of weak signal variables that are not selected by the noiseless lasso cannot be too large.

Figure 2.1: Conceptual illustration of the required magnitude of $|\beta_j|$ in (**M4**). Brown, orange and blue lines represent the signal strength of strong signal variables, weak signal variables in $\check{\mathcal{A}}_\lambda$ and weak signal variables in $\check{\mathcal{A}}_\lambda^c$, respectively. Black lines represent noise variables.



A conceptual illustration of (**M4**) is presented in Figure 2.1, in which lines show the required magnitude of β_j. Specifically, it shows that (**M4**) allows for the presence of both strong signal and weak signal variables, with a gap in signal strength.

- Brown line represents the signal strength of strong signal variables, $|\beta_j| \gtrsim \lambda\sqrt{q}$, $j \in \mathcal{S}$.

- Orange line represents the signal strength of weak signal variables that are selected by the noiseless lasso, $|\beta_j| \precsim \sqrt{\log(p)/n}$, $j \in \check{\mathcal{A}}_\lambda \backslash \mathcal{S}$.

- Blue line represents the signal strength of weak signal variables that are not selected by the noiseless lasso, $\|\boldsymbol{X}_{\mathcal{A}\backslash(\check{\mathcal{A}}_\lambda \cup \mathcal{S})}\boldsymbol{\beta}_{\mathcal{A}\backslash(\check{\mathcal{A}}_\lambda \cup \mathcal{S})}\|_2 \precsim \sqrt{\log(p)}$.

The condition in (**M4**) for weak signal variables differs based on whether the variable is in $\check{\mathcal{A}}_\lambda$, which may seem unintuitive. In fact, the results in this chapter can be obtained with (**M4**) replaced with (**M4a**) and (**M4b**). In the new conditions, and in particular, in (**M4a**), the condition on signal strength no longer depends on $\check{\mathcal{A}}_\lambda$.

(**M4a**) Recall that $\mathcal{A} \equiv \text{supp}(\boldsymbol{\beta})$ and $\mathcal{S} \equiv \{j : |\beta_j| > 3\lambda\sqrt{q}/\phi^2\}$, where $q \equiv |\mathcal{A}| \equiv |\text{supp}(\boldsymbol{\beta})|$, and ϕ is defined in (**E**). The signal strength satisfies

$$\left\|\boldsymbol{\beta}_{\mathcal{A}\setminus\mathcal{S}}\right\|_\infty = \mathcal{O}\left(\sqrt{\frac{\log(p)}{n}}\right). \tag{2.13}$$

(**M4b**) Recall that $\boldsymbol{G} \equiv \boldsymbol{X}^\top\boldsymbol{X}/n$. $\boldsymbol{X}$, $\mathcal{A}$ and $\check{\mathcal{A}}_\lambda$ satisfy

$$\left\|\left(\boldsymbol{G}_{(\check{\mathcal{A}}_\lambda,\check{\mathcal{A}}_\lambda)}\right)^{-1}\boldsymbol{G}_{(\check{\mathcal{A}}_\lambda,\mathcal{A}\setminus\check{\mathcal{A}}_\lambda)}\right\|_\infty = \mathcal{O}(1). \tag{2.14}$$

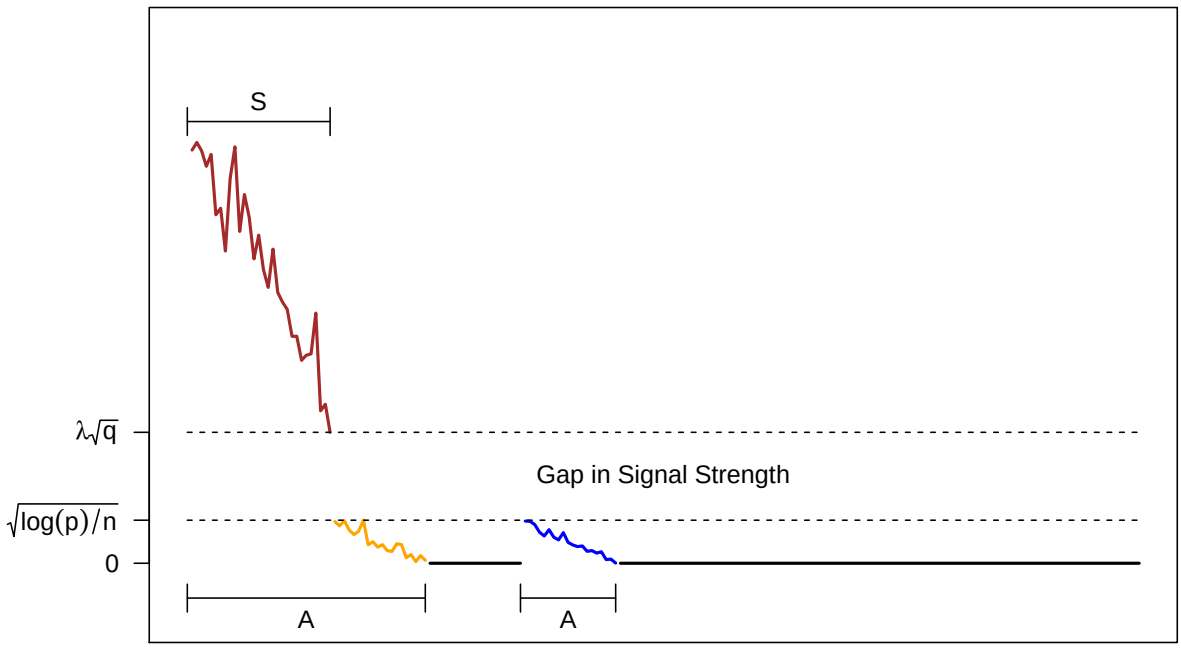
Condition (**M4a**) unifies the condition on $\|\boldsymbol{\beta}_{\check{\mathcal{A}}_\lambda\setminus\mathcal{S}}\|_\infty$ and on $\|\boldsymbol{X}_{\mathcal{A}\setminus(\check{\mathcal{A}}_\lambda\cup\mathcal{S})}\boldsymbol{\beta}_{\mathcal{A}\setminus(\check{\mathcal{A}}_\lambda\cup\mathcal{S})}\|_2$, and states that the orange and blue lines in Figure 2.1 can be of the same magnitude; see Figure 2.2 for a conceptual illustration of (**M4a**). In other words, compared to (**M4**), (**M4a**) no longer depends on $\check{\mathcal{A}}_\lambda$.

Condition (**M4b**) is similar to the mutual incoherence condition (see, e.g., Fuchs, 2005, Tropp, 2006, Wainwright, 2009), which requires that

$$\limsup_{n\to\infty}\left\|\left(\boldsymbol{G}_{(\mathcal{A},\mathcal{A})}\right)^{-1}\boldsymbol{G}_{(\mathcal{A},\setminus\mathcal{A})}\right\|_\infty < 1. \tag{2.15}$$

However, (**M4b**) is considerably milder: first, requiring a value to be bounded above is much milder than requiring it to be smaller than one. Second, when the model is sparse, i.e., $q/p \to 0$, and when n is large, $|\mathcal{A}^c|$ is on the same order as p. In this case, $(\boldsymbol{G}_{(\check{\mathcal{A}}_\lambda,\check{\mathcal{A}}_\lambda)})^{-1}\boldsymbol{G}_{(\check{\mathcal{A}}_\lambda,\mathcal{A}\setminus\check{\mathcal{A}}_\lambda)}$ is of dimension no larger than $n \times q$, which is substantially smaller than the dimension of $(\boldsymbol{G}_{(\mathcal{A},\mathcal{A})})^{-1}\boldsymbol{G}_{(\mathcal{A},\setminus\mathcal{A})}$, i.e., approximately $q \times p$. Thus, $\|(\boldsymbol{G}_{(\check{\mathcal{A}}_\lambda,\check{\mathcal{A}}_\lambda)})^{-1}\boldsymbol{G}_{(\check{\mathcal{A}}_\lambda,\mathcal{A}\setminus\check{\mathcal{A}}_\lambda)}\|_\infty$ is expected to be much smaller than $\|(\boldsymbol{G}_{(\mathcal{A},\mathcal{A})})^{-1}\boldsymbol{G}_{(\mathcal{A},\setminus\mathcal{A})}\|_\infty$. In other words, the left-hand side of (**M4b**) is smaller than the left-hand side of the mutual incoherence condition, and the right-hand side of (**M4b**) is larger than the right-hand side of the mutual incoherence condition.

Figure 2.2: Conceptual illustration of the required magnitude of $|\beta_j|$ in (**M4a**). Brown line represents the signal strength of strong signal variables; orange and blue lines represent the signal strength of weak signal variables. Black lines represent noise variables.



Condition (**E**) is the $(q \vee \check{q}_\lambda, q \vee \check{q}_\lambda, 1)$-restricted eigenvalue condition (Bickel et al., 2009, van de Geer and Bühlmann, 2009). Note that although (**E**) depends on q_λ, which could be larger than q, with some mild conditions, $\check{q}_\lambda \asymp q$ (see Theorem 3 in Belloni and Chernozhukov, 2013). This makes (**E**) not much more stringent than the $(q, q, 1)$-restricted eigenvalue condition, which is a standard condition in the lasso literature.

The first part of (**T**) requires that δ converges to zero at a slower rate than $\sqrt{\log(p)/n}/\lambda$, which means that λ does not converge to a transition point too fast, at which some variable enters or leaves $\check{\mathcal{A}}_\lambda$. Since $\sqrt{\log(p)/n}/\lambda \to 0$ by (**M3**), the second part of (**T**) requires that $\min_{j \in \check{\mathcal{A}}_\lambda \setminus \mathcal{S}} |(\boldsymbol{G}_{(\check{\mathcal{A}}_\lambda, \check{\mathcal{A}}_\lambda)})^{-1} \check{\boldsymbol{\tau}}_{\lambda, \check{\mathcal{A}}_\lambda}|_j$ does not converge to zero too fast.

Conditions (**M4**) and (**T**) critically depend on the set of variables selected by the noiseless lasso, $\check{\mathcal{A}}_\lambda$, which could be hard to interpret. However, as shown in Remarks 2.2.2 and 2.2.3, with some simple designs, we can simplify these two conditions to make them more interpretable. These two remarks are proven in Appendix A.3.

Remark 2.2.2. *If the design matrix $\boldsymbol{X}$ is orthonormal, i.e., $\boldsymbol{G} \equiv \boldsymbol{X}^\top \boldsymbol{X}/n = \boldsymbol{I}$, then $\check{\boldsymbol{\beta}}_\lambda = \text{sign}(\boldsymbol{\beta})(|\boldsymbol{\beta}| - \lambda)_+$ can be obtained by soft-thresholding $\boldsymbol{\beta}$ with threshold λ. It then follows that for n sufficiently large, $\check{\mathcal{A}}_\lambda = \mathcal{S}$. Furthermore, in this case,*

$$\min_{j \in \check{\mathcal{A}}_\lambda \setminus \mathcal{S}} \left| \left[\boldsymbol{G}_{(\check{\mathcal{A}}_\lambda, \check{\mathcal{A}}_\lambda)} \right]^{-1} \check{\boldsymbol{\tau}}_{\lambda, \check{\mathcal{A}}_\lambda} \right|_j = 1.$$

*When combined with (**M3**), this implies that the second part of (**T**) is satisfied:*

$$\frac{\sqrt{\log(p)/n}/\lambda}{\min_{j \in \check{\mathcal{A}}_\lambda \setminus \mathcal{S}} \left| \left[\boldsymbol{G}_{(\check{\mathcal{A}}_\lambda, \check{\mathcal{A}}_\lambda)} \right]^{-1} \check{\boldsymbol{\tau}}_{\lambda, \check{\mathcal{A}}_\lambda} \right|_j} \to 0.$$

*On the other hand, (**M3**) and (**M4**) imply that $\lambda \succ \sqrt{\log(p)/n}$ and $\|\boldsymbol{\beta}_{\check{\mathcal{A}}_\lambda \setminus \mathcal{S}}\|_\infty = \mathcal{O}(\sqrt{\log(p)/n})$, respectively. Together, these imply that $\check{\mathcal{A}}_\lambda = \mathcal{S}$, where $\mathcal{S} \equiv \{j : |\beta_j| \succeq \lambda\sqrt{q}\}$. Thus, in this special case, Conditions (**M4**) and (**T**) take a much simpler form, wherein $\check{\mathcal{A}}_\lambda$ is replaced with $\mathcal{S}$.*

Remark 2.2.3. *If the covariance of design matrix $\boldsymbol{X}$ is block-diagonal, such that $\boldsymbol{G}_{(\mathcal{A}, \setminus \mathcal{A})} = \boldsymbol{0}$, then $\mathcal{S} \subseteq \check{\mathcal{A}}_\lambda \subseteq \mathcal{A}$. In this case, if we write Conditions (M4) and (T) to hold for any set*

M such that $\mathcal{A} \supseteq \mathcal{M} \supseteq \mathcal{S}$*, then they automatically hold for* $\check{\mathcal{A}}_\lambda$ *and hence become much more interpretable.*

We now present Proposition 2.2.4 and Corollary 2.2.5, which are proven in Appendices A.1 and A.2, respectively.

Proposition 2.2.4. *Suppose conditions (***M1***)-(***M4***), (***E***) and (***T***) hold. Then, we have* $\lim_{n\to\infty} \Pr\left[\hat{\mathcal{A}}_\lambda = \check{\mathcal{A}}_\lambda\right] = 1$*, where* $\check{\mathcal{A}}_\lambda \equiv \text{supp}(\check{\boldsymbol{\beta}}_\lambda)$*, with* $\check{\boldsymbol{\beta}}_\lambda$ *defined in* (1.6)*.*

Corollary 2.2.5. *Suppose conditions (***M1***)-(***M3***), (***M4a***), (***M4b***), (***E***) and (***T***) hold. Then, we have* $\lim_{n\to\infty} \Pr\left[\hat{\mathcal{A}}_\lambda = \check{\mathcal{A}}_\lambda\right] = 1$*.*

It is important to emphasize the difference between the result $\lim_{n\to\infty} \Pr[\hat{\mathcal{A}}_\lambda = \check{\mathcal{A}}_\lambda] = 1$ and variable selection consistency of the lasso. Variable selection consistency asserts that $\Pr[\hat{\mathcal{A}}_\lambda = \mathcal{A}] \to 1$, and requires the stringent irrepresentable and $\boldsymbol{\beta}$-min conditions with the lasso (see, e.g., Meinshausen and Bühlmann, 2006, Zhao and Yu, 2006, Wainwright, 2009). For estimation methods with folded-concave penalties (e.g., Fan and Li, 2001, Zhang, 2010a), the irrepresentable condition may be relaxed. However, to achieve variable selection consistency, they still do not allow the existence of weak signal variables. Furthermore, similar to the lasso, these estimation methods are unable to provide confidence intervals, which are crucial in scientific reasoning. In contrast, Proposition 2.2.4 and Corollary 2.2.5 assert that under milder conditions, $\hat{\mathcal{A}}_\lambda$ converges with high probability to a deterministic set $\check{\mathcal{A}}_\lambda$ with cardinality smaller than n, *which is likely different from* $\mathcal{A} \equiv \text{supp}(\boldsymbol{\beta})$. Based on Proposition 2.2.4 and Corollary 2.2.5, we could build asymptotically valid confidence intervals, as shown in Theorem 2.2.6.

Proposition 2.2.4 and Corollary 2.2.5 imply that asymptotically, we "pay no price" for peeking at our data by performing the lasso: we should be able to perform downstream analyses on the subset of variables in $\hat{\mathcal{A}}_\lambda$ *as though we had obtained that subset without looking at the data*. This intuition will be formalized in Theorem 2.2.6.

Theorem 2.2.6, which is proven in Appendix A.4, shows that $\bar{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda)}$ in (2.6) is asymptotically normal, with mean and variance suggested by classical OLS theory: that is, *the fact*

that $\hat{\mathcal{A}}_\lambda$ *was selected by peeking at the data has no effect on the asymptotic distribution of* $\bar{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda)}$. This result requires that λ be chosen in a non-data-adaptive way. Otherwise, $\check{\mathcal{A}}_\lambda$ will be affected by the random error $\boldsymbol{\epsilon}$ through λ, which complicates the distribution of $\bar{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda)}$. Theorem 2.2.6 requires Condition (**W**), which is used to apply the Lindeberg-Feller Central Limit Theorem. This condition can be relaxed if the noise $\boldsymbol{\epsilon}$ is normally distributed.

(**W**) λ, $\boldsymbol{\beta}$ and $\boldsymbol{X}$ are such that $\lim_{n\to\infty} \|\boldsymbol{r}^w\|_\infty / \|\boldsymbol{r}^w\|_2 \to 0$, where

$$\boldsymbol{r}^w \equiv \boldsymbol{e}^j \left(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top \boldsymbol{X}_{\check{\mathcal{A}}_\lambda} \right)^{-1} \boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top,$$

and $\boldsymbol{e}^j$ is the row vector of length $|\check{\mathcal{A}}_\lambda|$ with the entry corresponding to β_j equal to one, and zero otherwise.

Theorem 2.2.6. *Suppose (***M1***)–(***M4***) (or with (***M4***) replaced by (***M4a***) and (***M4b***)), (***E***), (***T***) and (***W***) hold. Then, for any* $j \in \hat{\mathcal{A}}_\lambda$,

$$\frac{\bar{\beta}_j^{(\hat{\mathcal{A}}_\lambda)} - \beta_j^{(\hat{\mathcal{A}}_\lambda)}}{\sigma_{\boldsymbol{\epsilon}} \sqrt{\left[\left(\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda} \right)^{-1} \right]_{(j,j)}}} \to_d \mathcal{N}(0,1), \tag{2.16}$$

where $\bar{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda)}$ *is defined in* (2.6) *and* $\boldsymbol{\beta}^{(\hat{\mathcal{A}}_\lambda)}$ *in* (2.4)*, and* $\sigma_{\boldsymbol{\epsilon}}$ *is the standard deviation of* $\boldsymbol{\epsilon}$ *in the linear model* (2.1)*.*

The error standard deviation $\sigma_{\boldsymbol{\epsilon}}$ in (2.16) is usually unknown. It can be estimated using various high-dimensional estimation methods, e.g., the scaled lasso (Belloni et al., 2011, Sun and Zhang, 2012), cross-validation (CV) based methods (Fan et al., 2012) or method-of-moments based methods (Dicker, 2014); see a comparison study of high dimensional error variance estimation methods in Reid et al. (2016). Alternatively, Theorem 2.2.7 shows that we could also consistently estimate the error variance using the post-selection OLS residual sum of square (RSS).

Theorem 2.2.7. *Suppose (***M1***)–(***M4***), (***E***) and (***T***) hold, and* $\log(p)/(n - \check{q}_\lambda) \to 0$*, where* $\check{q}_\lambda \equiv |\check{\mathcal{A}}_\lambda|$*. Then*

$$\frac{1}{n - \hat{q}_\lambda} \left\| \boldsymbol{y} - \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda} \bar{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda)} \right\|_2^2 \to_p \sigma_{\boldsymbol{\epsilon}}^2, \tag{2.17}$$

where $\hat{q}_\lambda \equiv |\hat{\mathcal{A}}_\lambda|$.

Theorem 2.2.7 is proven in Appendix A.5. In (2.17), $\boldsymbol{y} - \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda} \bar{\boldsymbol{\beta}}^{\hat{\mathcal{A}}_\lambda}$ is the fitted OLS residual on the sub-model (2.2). $\log(p)/(n - \check{q}_\lambda) \to 0$ is a weak condition. Since $\log(p)/n \to 0$, $\log(p)/(n - \check{q}_\lambda) \to 0$ is satisfied if $\lim_{n\to\infty} \check{q}_\lambda / n < 1$.

To summarize, in this section, we have provided a theoretical justification for a procedure that seems, intuitively, to be statistically unjustifiable:

1. Perform the lasso in order to obtain the support set $\hat{\mathcal{A}}_\lambda$;

2. Use OLS to fit the sub-model containing just the variables in $\hat{\mathcal{A}}_\lambda$;

3. Use the classical confidence intervals from that OLS model, without accounting for the fact that $\hat{\mathcal{A}}_\lambda$ was obtained by peeking at the data.

Theorem 2.2.6 guarantees that the naïve confidence intervals in Step 3 will indeed have approximately correct coverage, in the sense of (2.7).

2.3 Numerical Examination of Naïve Confidence Intervals

In this section, we perform simulation studies to examine the coverage probability (2.7) of the *naïve* confidence intervals obtained by applying standard OLS software to the data $(\boldsymbol{y}, \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda})$.

Recall from Section 2.1 that (2.7) involves the probability that the confidence interval contains the quantity $\boldsymbol{\beta}^{(\hat{\mathcal{A}}_\lambda)}$, which in general does not equal the population regression coefficient vector $\boldsymbol{\beta}_{\hat{\mathcal{A}}_\lambda}$. Inference for $\boldsymbol{\beta}$ is discussed in Sections 2.4 and 2.5.

The results in this section complement simulation findings in Leeb et al. (2015).

2.3.1 Methods for Comparison

Following Theorem 2.2.6, for $\bar{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda)}$ defined in (2.6), and for each $j \in \hat{\mathcal{A}}_\lambda$, the level $1-\alpha$, $\alpha < 1$, naïve confidence interval takes the form

$$\mathrm{CI}_j^{(\hat{\mathcal{A}}_\lambda)} \equiv \left(\bar{\beta}_j^{(\hat{\mathcal{A}}_\lambda)} - \Phi_{\mathcal{N}}^{-1}\left[1-\frac{\alpha}{2}\right] \hat{\sigma}_{\boldsymbol{\epsilon}} \sqrt{\left[\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}\right]_{(j,j)}},\ \bar{\beta}_j^{(\hat{\mathcal{A}}_\lambda)} + \Phi_{\mathcal{N}}^{-1}\left[1-\frac{\alpha}{2}\right] \hat{\sigma}_{\boldsymbol{\epsilon}} \sqrt{\left[\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}\right]_{(j,j)}} \right), \tag{2.18}$$

where $\Phi_{\mathcal{N}}^{-1}[\cdot]$ is the quantile function of the standard normal distribution. In order to obtain the set $\hat{\mathcal{A}}_\lambda$, we must apply the lasso using some value of λ. Note that (**M3**) requires that $\lambda \succ \sqrt{\log(p)/n}$, which is slightly larger than the prediction optimal rate, $\lambda \asymp \sqrt{\log(p)/n}$ (Bickel et al., 2009, van de Geer and Bühlmann, 2009). Thus, we propose to use the tuning parameter value λ_{1SE}, which is the largest value of λ for which the 10-fold CV prediction mean squared error (PMSE) is within one standard error of the minimum CV PMSE (see Section 7.10.1 in Hastie et al., 2009). We leave the optimal choice of tuning parameters to future research.

As a comparison, we also report the confidence intervals for $\boldsymbol{\beta}^{(\hat{\mathcal{A}}_\lambda)}$ by the exact lasso post-selection inference procedure proposed in Lee et al. (2016); this procedure is implemented in the `selectiveInference` R package. For exact lasso post-selection confidence intervals, we adopt the procedure in Section 7 of Lee et al. (2016) to choose its tuning parameter: we let $\lambda = 2\mathrm{E}[\|\boldsymbol{X}^\top \boldsymbol{e}\|_\infty]/n$, where we simulate $\boldsymbol{e} \sim \mathcal{N}_n(\boldsymbol{0}, \hat{\sigma}_{\boldsymbol{\epsilon}}^2 \boldsymbol{I})$, and approximate its expectation based on the average of 1000 replicates. For fairer comparisons, we also report the performance of naïve confidence intervals with λ_{SUP}. Unlike λ_{1SE}, λ_{SUP} does not depend on the randomness in $\boldsymbol{y}$.

In both approaches, the standard deviation of errors, $\sigma_{\boldsymbol{\epsilon}}$ in (2.1), is estimated using the scaled lasso (Sun and Zhang, 2012).

2.3.2 Simulation Set-Up

For the simulations, we consider two partial correlation settings for $\boldsymbol{X}$, generated based on (i) a scale-free graph and (ii) a stochastic block model (see, e.g., Kolaczyk, 2009), each

containing $p = 100$ nodes. These settings are relaxations of the simple orthogonal and block-diagonal settings considered in Section 2.2, and are displayed in Figure 2.3.

In the scale-free graph setting, we use the `igraph` package in R to simulate an undirected, scale-free network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with power-law exponent parameter 5, and edge density 0.05. Such graphs are commonly found in cell biology (Wolf et al., 2002, Albert, 2005). Here, $\mathcal{V} = \{1, \ldots, p\}$ is the set of nodes in the graph, and $\mathcal{E}$ is the set of edges. This results in a total of $|\mathcal{E}| = 247$ edges in the graph. We then order the indices of the nodes in the graph so that the first, second, third, fourth, and fifth nodes correspond to the 10th, 20th, 30th, 40th, and 50th least-connected nodes in the graph.

In the stochastic block model setting, we first generate two dense Erdős-Rényi graphs (Erdős and Rényi, 1959, Gilbert, 1959) with five nodes and 95 nodes, respectively. In each graph, the edge density is 0.3. We then add edges randomly between these two graphs to achieve an inter-graph edge density of 0.05. The indices of the nodes are ordered so that the nodes in the five-node graph precede the remaining nodes.

Next, for both graph settings, we define the weighted adjacency matrix, $\boldsymbol{A}$, as follows:

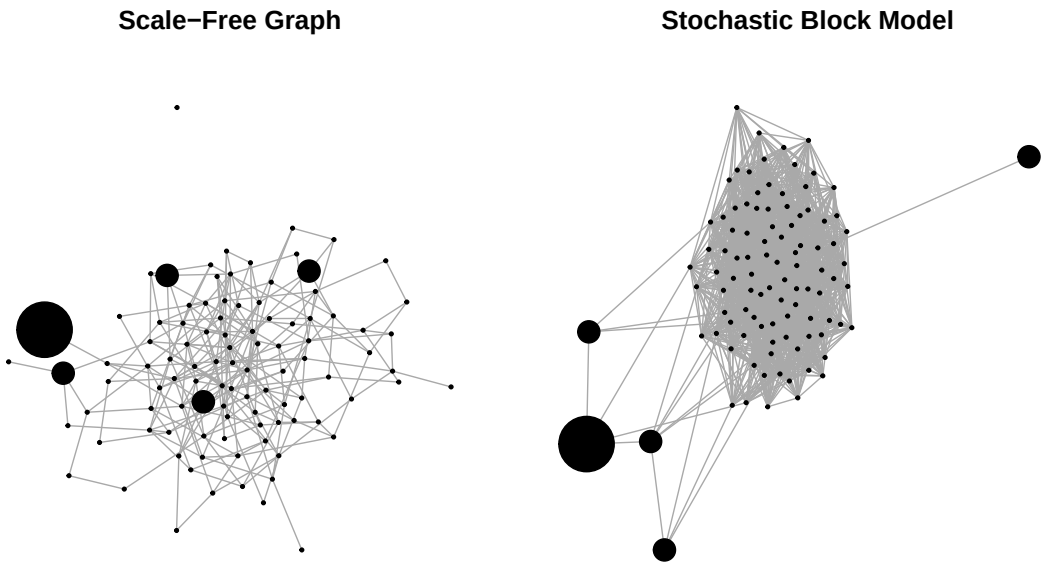
$$A_{(j,k)} = \begin{cases} 1 & \text{for } j = k \\ \rho & \text{for } (j,k) \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases}, \tag{2.19}$$

where $\rho \in \{0.2, 0.6\}$. We then set $\boldsymbol{\Sigma} = \boldsymbol{A}^{-1}$, and standardize $\boldsymbol{\Sigma}$ so that $\Sigma_{(j,j)} = 1$, for all $j = 1, \ldots, p$. We simulate observations $\boldsymbol{X} \sim_{i.i.d.} \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, and generate the outcome $\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I})$, $n \in \{300, 400, 500\}$, where

$$\beta_j = \begin{cases} 1 & \text{for } j = 1 \\ 0.1 & \text{for } 2 \leq j \leq 5 \\ 0 & \text{otherwise} \end{cases}.$$

A range of error variances $\sigma^2_{\boldsymbol{\epsilon}}$ are used to produce signal-to-noise ratios, SNR $\equiv (\boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta})/\sigma^2_{\boldsymbol{\epsilon}} \in \{0.1, 0.3, 0.5\}$.

Figure 2.3: Graph structures in the scale-free graph and stochastic block model settings. The size of nodes corresponds to the magnitude of corresponding elements in $\boldsymbol{\beta}$.



Throughout the simulations, $\boldsymbol{\Sigma}$ and $\boldsymbol{\beta}$ are held fixed over $R = 1000$ repetitions of the simulation study, while $\boldsymbol{X}$ and $\boldsymbol{y}$ vary.

2.3.3 Simulation Results

We calculate the average length and coverage proportion of the 95% naïve confidence intervals, where the coverage proportion is defined as

$$\text{Coverage Proportion} \equiv \sum_{r=1}^{R} \sum_{j \in \hat{\mathcal{A}}_\lambda^r} \mathbb{1}\left[\beta_j^{(\hat{\mathcal{A}}_\lambda^r)} \in \text{CI}_j^{(\hat{\mathcal{A}}_\lambda^r),r}\right] \Big/ \left|\hat{\mathcal{A}}_\lambda^r\right|, \tag{2.20}$$

where $\hat{\mathcal{A}}_\lambda^r$ and $\text{CI}_j^{(\hat{\mathcal{A}}_\lambda^r),r}$ are the set of variables selected by the lasso in the rth repetition, and the 95% naïve confidence interval in (2.18) for the jth variable in the rth repetition, respectively; $\beta_j^{(\hat{\mathcal{A}}_\lambda^r)}$ was defined in (2.4). In order to calculate the coverage proportion associated with the exact lasso post selection procedure of Lee et al. (2016), we replace $\text{CI}_j^{(\hat{\mathcal{A}}_\lambda^r),r}$

in (2.20) with the confidence interval output by the `selectiveInference` R package.

Tables 2.1 and 2.2 show the coverage proportion of 95% naïve confidence intervals and 95% exact lasso post-selection confidence intervals under the scale-free graph and stochastic block model settings, respectively. The result shows that the exact confidence intervals control the coverage probability (2.7) better than the naïve confidence intervals when the sample size is small. However, when the sample size is large, the coverage probability of the naïve confidence intervals is approximately correct. This corroborates the findings in Leeb et al. (2015), in which the authors consider settings with $n = 30$ and $p = 10$. The coverage probability of the naïve confidence intervals with tuning parameter λ_{1SE} is slightly too small. Tables 2.1 and 2.2 also show that naïve confidence intervals are narrower than exact lasso post-selection confidence intervals, especially when the signal is weak.

In addition, to evaluate whether $\hat{\mathcal{A}}_\lambda$ is deterministic, among repetitions that $\hat{\mathcal{A}}^r_\lambda \neq \emptyset$ (there is no confidence interval if $\hat{\mathcal{A}}^r_\lambda = \emptyset$), we also calculate the proportion of $\hat{\mathcal{A}}^r_\lambda = \mathcal{D}$, where $\mathcal{D}$ is the most common $\hat{\mathcal{A}}^r_\lambda$, $b = 1, \ldots, 1000$. The result is summarized in Table 2.3, which shows that $\hat{\mathcal{A}}_\lambda$ is almost deterministic with tuning parameter λ_{SUP}. With λ_{1SE}, due to the randomness in the tuning parameter, $\hat{\mathcal{A}}_\lambda$ is less deterministic, which may explain the result that the coverage probability is slightly smaller than the desired level in this case.

2.4 Inference for β With the Naïve Score Test

Sections 2.2 and 2.3 focused on the task of developing confidence intervals for $\boldsymbol{\beta}^{(\mathcal{M})}$ in (2.2), where $\mathcal{M} = \hat{\mathcal{A}}_\lambda$, the set of variables selected by the lasso. However, recall from (2.4) that typically $\boldsymbol{\beta}^{(\mathcal{M})} \neq \boldsymbol{\beta}_{\mathcal{M}}$, where $\boldsymbol{\beta}$ was introduced in (2.1).

In this section, we shift our focus to performing inference on $\boldsymbol{\beta}$. We will exploit Proposition 2.2.4 to develop a simple approach for testing $H_{0,j} : \beta_j = 0$, for $j = 1, \ldots, p$.

Recall that in the low-dimensional setting, the classical score statistic for the hypothesis $H_{0,j} : \beta_j = 0$ is proportional to $\boldsymbol{X}_j^T(\boldsymbol{y} - \bar{\boldsymbol{y}}^{(\setminus j)})$, where $\bar{\boldsymbol{y}}^{(\setminus j)}$ is the vector of fitted values that results from OLS of $\boldsymbol{y}$ onto the $p-1$ variables $\boldsymbol{X}_{\setminus j}$. In order to adapt the classical score test statistic to the high-dimensional setting, we define the *naïve score test statistic* for testing

Table 2.1: Coverage proportions and average lengths of 95% naïve confidence intervals with tuning parameters λ_{SUP} and λ_{1SE}, and 95% exact post-selection confidence intervals under the scale-free graph partial correlation setting with $\rho \in \{0.2, 0.6\}$, sample size $n \in \{300, 400, 500\}$, dimension $p = 100$ and signal-to-noise ratio SNR $\in \{0.1, 0.3, 0.5\}$.

	ρ	0.2								
	n	300			400			500		
	SNR	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
Coverage	exact λ_{SUP}	0.949	0.949	0.949	0.950	0.959	0.959	0.956	0.964	0.965
	naïve λ_{SUP}	0.951	0.948	0.948	0.977	0.958	0.958	0.969	0.964	0.964
	naïve λ_{1SE}	0.922	0.936	0.928	0.944	0.937	0.930	0.950	0.938	0.930
Length	exact λ_{SUP}	1.902	0.427	0.327	1.148	0.368	0.284	0.815	0.327	0.254
	naïve λ_{SUP}	0.714	0.418	0.325	0.623	0.363	0.282	0.561	0.325	0.252
	naïve λ_{1SE}	0.719	0.418	0.324	0.625	0.363	0.282	0.559	0.325	0.252
	ρ	0.6								
	n	300			400			500		
	SNR	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
Coverage	exact λ_{SUP}	0.960	0.954	0.954	0.956	0.955	0.956	0.953	0.945	0.945
	naïve λ_{SUP}	0.962	0.950	0.950	0.972	0.954	0.954	0.966	0.944	0.942
	naïve λ_{1SE}	0.918	0.939	0.932	0.974	0.937	0.932	0.968	0.936	0.928
Length	exact λ_{SUP}	1.669	0.426	0.326	1.081	0.365	0.283	0.781	0.326	0.255
	naïve λ_{SUP}	0.711	0.418	0.324	0.624	0.363	0.281	0.559	0.324	0.251
	naïve λ_{1SE}	0.716	0.416	0.323	0.623	0.361	0.280	0.556	0.324	0.251

Table 2.2: Coverage proportions and average lengths of 95% naïve confidence intervals with tuning parameters λ_{SUP} and λ_{1SE}, and 95% exact post-selection confidence intervals under the stochastic block model setting. Details are as in Table 2.1.

	ρ	0.2								
	n	300			400			500		
	SNR	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
Coverage	exact λ_{SUP}	0.957	0.948	0.948	0.948	0.952	0.952	0.949	0.958	0.958
	naïve λ_{SUP}	0.958	0.946	0.946	0.969	0.952	0.952	0.972	0.956	0.956
	naïve λ_{1SE}	0.907	0.938	0.935	0.938	0.939	0.933	0.947	0.928	0.919
Length	exact λ_{SUP}	1.843	0.427	0.325	1.177	0.363	0.285	0.812	0.323	0.251
	naïve λ_{SUP}	0.704	0.415	0.322	0.615	0.360	0.279	0.554	0.322	0.250
	naïve λ_{1SE}	0.710	0.414	0.321	0.617	0.359	0.279	0.554	0.322	0.250
	ρ	0.6								
	n	300			400			500		
	SNR	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
Coverage	exact λ_{SUP}	0.961	0.959	0.959	0.946	0.957	0.958	0.947	0.946	0.945
	naïve λ_{SUP}	0.969	0.958	0.958	0.961	0.956	0.956	0.968	0.945	0.945
	naïve λ_{1SE}	0.925	0.941	0.934	0.919	0.927	0.921	0.959	0.944	0.940
Length	exact λ_{SUP}	1.759	0.420	0.323	1.133	0.361	0.280	0.778	0.323	0.250
	naïve λ_{SUP}	0.703	0.412	0.320	0.614	0.358	0.278	0.552	0.321	0.249
	naïve λ_{1SE}	0.705	0.411	0.319	0.612	0.358	0.278	0.553	0.321	0.249

Table 2.3: The proportion of selected set that equals the most common selected set among repetitions, under the scale-free graph and stochastic block model settings with tuning parameters λ_{SUP} and λ_{1SE}. Details are as in Table 2.1.

ρ	0.2								
n	300			400			500		
SNR	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
Scale-free λ_{SUP}	1.000	0.999	0.999	1.000	0.999	0.999	0.999	1.000	1.000
Scale-free λ_{1SE}	0.976	0.960	0.943	0.978	0.981	0.967	0.990	0.978	0.958
Stochastic block λ_{SUP}	0.999	0.997	0.997	1.000	0.999	0.999	1.000	1.000	0.999
Stochastic block λ_{1SE}	0.966	0.969	0.963	0.984	0.968	0.959	0.990	0.980	0.962
ρ	0.6								
n	300			400			500		
SNR	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
Scale-free λ_{SUP}	0.999	0.996	0.996	0.999	1.000	1.000	1.000	0.999	0.996
Scale-free λ_{1SE}	0.969	0.970	0.958	0.986	0.978	0.965	0.992	0.986	0.967
Stochastic block λ_{SUP}	1.000	0.999	0.999	1.000	0.999	0.999	1.000	1.000	1.000
Stochastic block λ_{1SE}	0.972	0.963	0.953	0.987	0.969	0.955	0.997	0.979	0.967

$H_{0,j} : \beta_j = 0$ as

$$S^j \equiv \boldsymbol{X}_j^T \left(\boldsymbol{y} - \bar{\boldsymbol{y}}^{(\hat{\mathcal{A}}_\lambda \setminus \{j\})} \right) \equiv \boldsymbol{X}_j^\top \left(\boldsymbol{I} - \boldsymbol{P}^{(\hat{\mathcal{A}}_\lambda \setminus \{j\})} \right) \boldsymbol{y}, \tag{2.21}$$

where

$$\bar{\boldsymbol{y}}^{(\hat{\mathcal{A}}_\lambda \setminus \{j\})} \equiv \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda \setminus \{j\}} \bar{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda \setminus \{j\})},$$

and

$$\boldsymbol{P}^{(\hat{\mathcal{A}}_\lambda \setminus \{j\})} \equiv \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda \setminus \{j\}} \left(\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda \setminus \{j\}}^\top \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda \setminus \{j\}} \right)^{-1} \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda \setminus \{j\}}^\top$$

is the orthogonal projection matrix onto the set of variables in $\hat{\mathcal{A}}_\lambda \setminus \{j\}$. $\bar{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda \setminus \{j\})}$ is defined in (2.6), which is the post-selection sub-model OLS estimate.

In Theorem 2.4.1, we derive the asymptotic distribution for S^j under $H_{0,j} : \beta_j = 0$. We first introduce two new conditions.

First, we require that the total signal strength of variables not selected by the noiseless lasso, (1.6), is small.

(**M4***) Recall that $\mathcal{A} \equiv \text{supp}(\boldsymbol{\beta})$ and $\mathcal{S} \equiv \{j : |\beta_j| > 3\lambda\sqrt{q}/\phi^2\}$, where $q \equiv |\mathcal{A}| \equiv |\text{supp}(\boldsymbol{\beta})|$, and ϕ is defined in (**E**). The signal strength satisfies

$$\left\| \boldsymbol{\beta}_{\check{\mathcal{A}}_\lambda \setminus \mathcal{S}} \right\|_\infty = \mathcal{O}\left(\sqrt{\frac{\log(p)}{n}} \right),$$

and

$$\left\| \boldsymbol{X}_{\mathcal{A} \setminus (\check{\mathcal{A}}_\lambda \cup \mathcal{S})} \boldsymbol{\beta}_{\mathcal{A} \setminus (\check{\mathcal{A}}_\lambda \cup \mathcal{S})} \right\|_2 = o(1),$$

where $\check{\mathcal{A}}_\lambda \equiv \text{supp}(\check{\boldsymbol{\beta}}_\lambda)$, with $\check{\boldsymbol{\beta}}_\lambda$ defined in (1.6).

Condition (**M4***) closely resembles (**M4**), which was required for Theorem 2.2.6 in Section 2.2. The only difference between the two is that (**M4***) requires $\|\boldsymbol{X}_{\mathcal{A} \setminus (\check{\mathcal{A}}_\lambda \cup \mathcal{S})} \boldsymbol{\beta}_{\mathcal{A} \setminus (\check{\mathcal{A}}_\lambda \cup \mathcal{S})}\|_2 = o(1)$, whereas (**M4**) requires only that $\|\boldsymbol{X}_{\mathcal{A} \setminus (\check{\mathcal{A}}_\lambda \cup \mathcal{S})} \boldsymbol{\beta}_{\mathcal{A} \setminus (\check{\mathcal{A}}_\lambda \cup \mathcal{S})}\|_2 = \mathcal{O}(\sqrt{\log(p)})$. In other words, testing the population regression parameter $\boldsymbol{\beta}$ in (2.1) requires more stringent assumptions than constructing confidence intervals for the parameters in the sub-model (2.2).

The following condition, required to apply the Lindeberg-Feller Central Limit Theorem, can be relaxed if the noise $\boldsymbol{\epsilon}$ in (2.1) is normally distributed.

(**S**) λ, $\boldsymbol{\beta}$ and $\boldsymbol{X}$ satisfy $\lim_{n\to\infty} \|\boldsymbol{r}^s\|_\infty / \|\boldsymbol{r}^s\|_2 = 0$, where $\boldsymbol{r}^s \equiv (\boldsymbol{I} - \boldsymbol{P}^{(\check{\mathcal{A}}_\lambda \setminus \{j\})})\boldsymbol{X}_j$.

We now present Theorem 2.4.1, which is proven in Appendix A.6.

Theorem 2.4.1. *Suppose (***M1***)–(***M3***), (***M4*****), (***E***), (***T***) and (***S***) hold. For any $j = 1, \ldots, p$, under the null hypothesis $H_{0,j} : \beta_j = 0$,*

$$T \equiv \frac{S^j}{\sigma_{\boldsymbol{\epsilon}} \sqrt{\boldsymbol{X}_j^\top \left(\boldsymbol{I} - \boldsymbol{P}^{(\hat{\mathcal{A}}_\lambda \setminus \{j\})}\right) \boldsymbol{X}_j}} \to_d \mathcal{N}(0,1), \tag{2.22}$$

where S^j was defined in (2.21)*, and where $\sigma_{\boldsymbol{\epsilon}}$ is the standard deviation of $\boldsymbol{\epsilon}$ in* (2.1)*.*

Theorem 2.4.1 states that the distribution of the naïve score test statistic S^j is asymptotically the same as if $\hat{\mathcal{A}}_\lambda$ were a deterministic set, as opposed to being selected by fitting a lasso model on the data. Based on (2.22), we reject the null hypothesis $H_{0,j} : \beta_j = 0$ at level $\alpha > 0$ if $|T| > \Phi^{-1}_{\mathcal{N}}[1 - \alpha/2]$, where $\Phi^{-1}_{\mathcal{N}}[\cdot]$ is the quantile function of the standard normal distribution.

We emphasize that Theorem 2.4.1 holds for any variable $j = 1, \ldots, p$, and thus can be used to test $H_{0,j} : \beta_j = 0$, for all $j = 1, \ldots, p$. This is in contrast to Theorem 2.2.6, which concerns confidence intervals for the parameters in the sub-model (2.2) consisting of the variables in $\hat{\mathcal{A}}_\lambda$, and hence holds only for $j \in \hat{\mathcal{A}}_\lambda$.

2.5 Numerical Examination of the Naïve Score Test

In this section, we compare the performance of the naïve score test (2.21) to three recent proposals from the literature for testing $H_{0,j} : \beta_j = 0$: namely, LDPE (Zhang and Zhang, 2014, van de Geer et al., 2014), SSLasso (Javanmard and Montanari, 2014a), and the decorrelated score test (dScore; Ning and Liu, 2017). LDPE is implemented in the `hdi` R package. R codes for SSLasso and dScore are provided by the authors. For the naïve score test, we estimate $\sigma_{\boldsymbol{\epsilon}}$, the standard deviation of the errors in (2.1), using the scaled lasso (Sun and Zhang, 2012).

All four of these methods require us to select the value of the lasso tuning parameter. For LDPE, SSLasso, and dScore, we use 10-fold cross-validation to select the tuning parameter value that produces the smallest cross-validated mean square error, λ_{MIN}. As in the numerical study of the naïve confidence intervals in Section 2.3, we implement the naïve score test using the tuning parameter value λ_{1SE} and λ_{SUP}.

Unless otherwise noted, all tests are performed at a significance level of 0.05.

In Section 2.5.1, we investigate the powers and type-I errors of the above tests in simulation experiments. Section 2.5.2 contains an analysis of a glioblastoma gene expression dataset.

2.5.1 Power and Type-I Error

In this section, we adapt the scale-free graph and the stochastic block model presented in Section 2.3.2 to have $p = 500$.

In the scale-free graph setting, we generate a scale-free graph with power-law exponent parameter 5, edge density 0.05, and $p = 500$ nodes. The resulting graph has $|\mathcal{E}| = 6237$ edges. We order the nodes in the graph so that jth node is the $(30 \times j)$-th least-connected node in the graph, for $1 \leq j \leq 10$. For example, the 4th node is the 120th least-connected node in the graph.

In the stochastic block model setting, we generate two dense Erdős-Rényi graphs with ten nodes and 490 nodes, respectively; each has an intra-graph edge density of 0.3. The node indices are ordered so that the nodes in the smaller graph precede those in the larger graph. We then randomly connect nodes between the two graphs in order to obtain an inter-graph edge density of 0.05.

Next, for both graph settings, we generate $\boldsymbol{A}$ as in (2.19), where $\rho \in \{0.2, 0.6\}$. We then set $\boldsymbol{\Sigma} = \boldsymbol{A}^{-1}$, and standardize $\boldsymbol{\Sigma}$ so that $\Sigma_{(j,j)} = 1$, for all $j = 1, \ldots, p$. We simulate observations $\boldsymbol{X} \sim_{i.i.d.} \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, and generate the outcome $\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I})$, $n \in \{100, 200, 400\}$,

where

$$\beta_j = \begin{cases} 1 & \text{for } 1 \leq j \leq 3 \\ 0.1 & \text{for } 4 \leq j \leq 10 \\ 0 & \text{otherwise} \end{cases}.$$

A range of error variances $\sigma^2_{\boldsymbol{\epsilon}}$ are used to produce signal-to-noise ratios, SNR $\equiv (\boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta})/\sigma^2_{\boldsymbol{\epsilon}} \in \{0.1, 0.3, 0.5\}$.

We hold $\boldsymbol{\Sigma}$ and $\boldsymbol{\beta}$ fixed over $B = 100$ repetitions of the simulation, while $\boldsymbol{X}$ and $\boldsymbol{y}$ vary.

For each test, the average power on the strong signal variables, the average power on the weak signal variables, and the average type-I error rate are defined as

$$\text{Power}_{\text{strong}} \equiv \frac{1}{B}\frac{1}{3}\sum_{r=1}^{R}\sum_{j:\beta_j=1} \mathbb{1}\left[p_j^r < 0.05\right], \tag{2.23}$$

$$\text{Power}_{\text{weak}} \equiv \frac{1}{B}\frac{1}{7}\sum_{r=1}^{R}\sum_{j:\beta_j=0.1} \mathbb{1}\left[p_j^r < 0.05\right], \tag{2.24}$$

$$\text{Type-I Error} \equiv \frac{1}{B}\frac{1}{490}\sum_{r=1}^{R}\sum_{j:\beta_j=0} \mathbb{1}\left[p_j^r < 0.05\right], \tag{2.25}$$

respectively. In (2.23)–(2.25), p_j^r is the p-value associated with null hypothesis $H_{0,j} : \beta_j = 0$ in the rth simulated data set. In the simulations, the graphs and $\boldsymbol{\beta}$ are held fixed over $R = 100$ repetitions of the simulation study, while $\boldsymbol{X}$ and $\boldsymbol{y}$ vary.

Tables 2.4 and 2.5 summarize the results in the two simulation settings. All four methods have approximate control over type-I error rate, even when $p \gg n$. Moreover, due to the similarity between these four approaches, as discussed in Section 2.6, all methods have comparable power.

2.5.2 Application to Glioblastoma Data

We investigate a glioblastoma gene expression data set previously studied in Horvath et al. (2006). For each of 130 patients, a survival outcome is available; we remove the twenty patients who were still alive at the end of the study. This results in a data set with $n = 110$

Table 2.4: Average power and type-I error rate for the hypotheses $H_{0,j} : \beta_j = 0$ for $j = 1, \ldots, p$, as defined in (2.23)–(2.25), under the scale-free graph setting with $p = 500$. Results are shown for various values of ρ, n, SNR. Methods for comparison include LDPE, SSLasso, dScore, and the naïve score test with tuning parameter λ_{1SE} and λ_{SUP}.

	ρ	0.2								
	n	100			200			400		
	SNR	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
$\text{Power}_{\text{strong}}$	LDPE λ_{MIN}	0.400	0.773	0.910	0.627	0.973	1.000	0.923	1.000	1.000
	SSLasso λ_{MIN}	0.410	0.770	0.950	0.650	0.970	1.000	0.910	1.000	1.000
	dScore λ_{MIN}	0.330	0.643	0.857	0.547	0.957	1.000	0.887	1.000	1.000
	nScore λ_{1SE}	0.427	0.763	0.893	0.677	0.977	1.000	0.957	1.000	1.000
	nScore λ_{SUP}	0.403	0.847	0.960	0.727	0.990	0.997	0.940	1.000	1.000
$\text{Power}_{\text{weak}}$	LDPE λ_{MIN}	0.064	0.083	0.056	0.054	0.059	0.079	0.070	0.079	0.113
	SSLasso λ_{MIN}	0.081	0.087	0.060	0.066	0.061	0.086	0.069	0.086	0.113
	dScore λ_{MIN}	0.044	0.056	0.036	0.039	0.039	0.060	0.046	0.056	0.093
	nScore λ_{1SE}	0.080	0.077	0.059	0.060	0.061	0.061	0.083	0.076	0.101
	nScore λ_{SUP}	0.061	0.091	0.109	0.070	0.109	0.107	0.097	0.103	0.114
T1 Error	LDPE λ_{MIN}	0.051	0.052	0.051	0.049	0.051	0.047	0.050	0.051	0.049
	SSLasso λ_{MIN}	0.056	0.056	0.056	0.054	0.055	0.053	0.053	0.054	0.054
	dScore λ_{MIN}	0.035	0.040	0.040	0.033	0.036	0.034	0.035	0.037	0.034
	nScore λ_{1SE}	0.061	0.057	0.048	0.056	0.055	0.040	0.060	0.046	0.046
	nScore λ_{SUP}	0.069	0.082	0.095	0.064	0.083	0.079	0.065	0.068	0.050
	ρ	0.6								
	n	100			200			400		
	SNR	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
$\text{Power}_{\text{strong}}$	LDPE λ_{MIN}	0.330	0.783	0.947	0.627	0.980	1.000	0.887	1.000	1.000
	SSLasso λ_{MIN}	0.347	0.790	0.957	0.623	0.987	1.000	0.867	1.000	1.000
	dScore λ_{MIN}	0.270	0.673	0.883	0.533	0.960	0.993	0.863	1.000	1.000
	nScore λ_{1SE}	0.357	0.767	0.887	0.677	0.980	0.997	0.937	1.000	1.000
	nScore λ_{SUP}	0.430	0.790	0.933	0.707	0.977	1.000	0.923	1.000	1.000
$\text{Power}_{\text{weak}}$	LDPE λ_{MIN}	0.031	0.046	0.063	0.064	0.074	0.076	0.054	0.077	0.119
	SSLasso λ_{MIN}	0.047	0.063	0.076	0.063	0.090	0.099	0.053	0.076	0.121
	dScore λ_{MIN}	0.021	0.037	0.047	0.036	0.060	0.044	0.034	0.050	0.083
	nScore λ_{1SE}	0.039	0.060	0.050	0.076	0.074	0.066	0.070	0.067	0.104
	nScore λ_{SUP}	0.071	0.089	0.136	0.081	0.121	0.104	0.114	0.113	0.123
T1 Error	LDPE λ_{MIN}	0.050	0.051	0.051	0.050	0.049	0.051	0.051	0.050	0.047
	SSLasso λ_{MIN}	0.056	0.056	0.056	0.054	0.055	0.053	0.053	0.054	0.054
	dScore λ_{MIN}	0.033	0.036	0.034	0.031	0.031	0.035	0.036	0.035	0.033
	nScore λ_{1SE}	0.056	0.060	0.045	0.061	0.051	0.040	0.058	0.048	0.047
	nScore λ_{SUP}	0.065	0.080	0.093	0.064	0.084	0.088	0.070	0.071	0.054

Table 2.5: Average power and type-I error rate for the hypotheses $H_{0,j} : \beta_j = 0$ for $j = 1, \ldots, p$, as defined in (2.23)–(2.25), under the stochastic block model setting with $p = 500$. Details are as in Table 2.4.

	ρ	0.2								
	n	100			200			400		
	SNR	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
$\text{Power}_{\text{strong}}$	LDPE λ_{MIN}	0.370	0.793	0.937	0.687	0.990	1.000	0.914	1.000	1.000
	SSLasso λ_{MIN}	0.393	0.803	0.933	0.687	0.990	1.000	0.892	1.000	1.000
	dScore λ_{MIN}	0.333	0.783	0.917	0.693	0.993	1.000	0.905	1.000	1.000
	nScore λ_{1SE}	0.400	0.797	0.903	0.713	0.997	1.000	0.910	1.000	1.000
	nScore λ_{SUP}	0.473	0.857	0.953	0.697	0.993	0.993	0.943	1.000	1.000
$\text{Power}_{\text{weak}}$	LDPE λ_{MIN}	0.041	0.044	0.051	0.057	0.050	0.071	0.050	0.093	0.071
	SSLasso λ_{MIN}	0.054	0.056	0.074	0.071	0.056	0.089	0.071	0.101	0.101
	dScore λ_{MIN}	0.037	0.044	0.057	0.060	0.046	0.077	0.056	0.101	0.094
	nScore λ_{1SE}	0.047	0.059	0.060	0.059	0.047	0.059	0.062	0.106	0.105
	nScore λ_{SUP}	0.059	0.071	0.107	0.043	0.083	0.070	0.069	0.094	0.106
T1ER	LDPE λ_{MIN}	0.051	0.049	0.048	0.050	0.050	0.050	0.051	0.050	0.049
	SSLasso λ_{MIN}	0.057	0.056	0.058	0.054	0.054	0.054	0.054	0.053	0.054
	dScore λ_{MIN}	0.043	0.040	0.041	0.041	0.044	0.042	0.042	0.042	0.041
	nScore λ_{1SE}	0.062	0.058	0.048	0.056	0.052	0.040	0.054	0.047	0.046
	nScore λ_{SUP}	0.064	0.074	0.090	0.059	0.076	0.076	0.060	0.060	0.49
	ρ	0.6								
	n	100			200			400		
	SNR	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
$\text{Power}_{\text{strong}}$	LDPE λ_{MIN}	0.327	0.827	0.960	0.700	0.983	0.997	0.968	1.000	1.000
	SSLasso λ_{MIN}	0.350	0.853	0.957	0.687	0.990	0.997	0.945	0.996	1.000
	dScore λ_{MIN}	0.297	0.787	0.937	0.697	0.987	0.993	0.968	0.996	1.000
	nScore λ_{1SE}	0.350	0.800	0.927	0.717	0.980	1.000	0.968	1.000	1.000
	nScore λ_{SUP}	0.420	0.870	0.957	0.720	0.987	1.000	0.947	1.000	1.000
$\text{Power}_{\text{weak}}$	LDPE λ_{MIN}	0.043	0.049	0.046	0.041	0.077	0.063	0.053	0.066	0.083
	SSLasso λ_{MIN}	0.061	0.054	0.070	0.053	0.086	0.083	0.067	0.099	0.105
	dScore λ_{MIN}	0.044	0.047	0.046	0.040	0.081	0.069	0.063	0.077	0.098
	nScore λ_{1SE}	0.059	0.056	0.044	0.054	0.087	0.074	0.067	0.086	0.103
	nScore λ_{SUP}	0.054	0.073	0.093	0.063	0.093	0.094	0.061	0.094	0.096
T1 Error	LDPE λ_{MIN}	0.049	0.049	0.049	0.049	0.050	0.049	0.049	0.047	0.048
	SSLasso λ_{MIN}	0.057	0.056	0.056	0.053	0.054	0.054	0.053	0.053	0.053
	dScore λ_{MIN}	0.033	0.039	0.036	0.031	0.033	0.033	0.032	0.030	0.031
	nScore λ_{1SE}	0.057	0.051	0.047	0.056	0.049	0.039	0.055	0.045	0.046
	nScore λ_{SUP}	0.063	0.079	0.089	0.062	0.077	0.075	0.060	0.062	0.048

observations. The gene expression measurements are normalized using the method of Gautier et al. (2004). We limit our analysis to $p = 3600$ highly-connected genes (Zhang and Horvath, 2005, Horvath and Dong, 2008). We log-transform the survival outcome and center it to have mean zero. Furthermore, we log-transform the expression data, and then standardize each gene to have mean zero and standard deviation one across the $n = 110$ observations.

Our goal is to identify individual genes whose expression levels are associated with survival time, after adjusting for the other 3599 genes in the data set. With FWER controlled at level 0.1 using the Holm procedure (Holm, 1979), the naïve score test identifies three such genes: CKS2, H2AFZ, and RPA3. You et al. (2015) observed that CKS2 is highly expressed in glioma. Vardabasso et al. (2014) found that histone genes, of which H2AFZ is one, are related to cancer progression. Jin et al. (2015) found that RPA3 is associated with glioma development. As a comparison, SSLasso finds two genes associated with patient survival: PPAP2C and RGS3. LDPE and dScore identify no genes at FWER of 0.1.

2.6 Connections of the Naïve Score Test With Existing Approaches

In the section, we show that although derived from different perspectives, the naïve post-selection prcedures are closely related to existing lasso debiased tests. Specifically, in Section 2.6.1, we show that the post-selection OLS estimator is closely related to the recently-proposed debiased lasso estimators, such as LDPE (Zhang and Zhang, 2014, van de Geer et al., 2014), and SSLasso (Javanmard and Montanari, 2013a, 2014a). These estimators are obtained by inverting the stationary condition of (1.2), and can be characterized as (see, e.g., Javanmard and Montanari, 2013b, van de Geer et al., 2014, Javanmard and Montanari, 2014a)

$$\hat{\boldsymbol{\beta}}_\lambda^{(debiased)} \equiv \hat{\boldsymbol{\beta}}_\lambda + \frac{1}{n}\boldsymbol{M}\boldsymbol{X}^\top \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda\right), \tag{2.26}$$

where $\hat{\boldsymbol{\beta}}_\lambda$ is the lasso estimator, defined in (1.2). Let $\boldsymbol{G} \equiv \boldsymbol{X}^\top\boldsymbol{X}/n$ be the scaled Gramian matrix. For the matrix $\boldsymbol{M}$, Javanmard and Montanari (2014a) showed that when $\|\boldsymbol{M}\boldsymbol{G} - \boldsymbol{I}\|_\infty$ is small, the asymptotic bias of $\sqrt{n}\hat{\boldsymbol{\beta}}_\lambda^{(debiased)}$ could be negligible. Note that after debias-

ing, $\hat{\boldsymbol{\beta}}_\lambda^{(debiased)}$ is not necessarily sparse. With low-dimensional data, we could take $\boldsymbol{M} = \boldsymbol{G}^{-1}$, and asymptotically debias lasso estimates. However, with high-dimensional data, $\boldsymbol{G}$ is singular, in which case $\boldsymbol{M}$ is usually derived from computationally intense procedures.

In Section 2.6.2, we show that a slightly modified version of the naïve score test is closely related to the decorrelated score test (dScore; Ning and Liu, 2017), which is based on the debiased score function

$$S^{(debiased)} = \boldsymbol{X}_j^\top \left(\boldsymbol{y} - \boldsymbol{X}_{\setminus j}\hat{\boldsymbol{\beta}}_\lambda^{(\setminus j)}\right) - \frac{1}{n}\boldsymbol{X}_j^\top \boldsymbol{X}_{\setminus j}\boldsymbol{M}^{(\setminus j)}\boldsymbol{X}_{\setminus j}^\top \left(\boldsymbol{y} - \boldsymbol{X}_{\setminus j}\hat{\boldsymbol{\beta}}_\lambda^{(\setminus j)}\right), \tag{2.27}$$

where $\hat{\boldsymbol{\beta}}_\lambda^{(\setminus j)}$ is the lasso estimator of $\boldsymbol{\beta}^{(\setminus j)}$, defined in (2.4):

$$\hat{\boldsymbol{\beta}}_\lambda^{(\setminus j)} = \underset{\boldsymbol{b}\in\mathbb{R}^{p-1}}{\arg\min}\left\{\frac{1}{2n}\left\|\boldsymbol{y} - \boldsymbol{X}_{\setminus j}\boldsymbol{b}\right\|_2^2 + \lambda\left\|\boldsymbol{b}\right\|_1\right\}. \tag{2.28}$$

Under the null hypothesis $H_{0,j} : \beta_j = 0$, $\boldsymbol{\beta}^{(\setminus j)} = \boldsymbol{\beta}_{\setminus j}$. Therefore, $\hat{\boldsymbol{\beta}}_\lambda^{(\setminus j)}$ is also the lasso estimate of $\boldsymbol{\beta}_{\setminus j}$ under the null hypothesis $H_{0,j} : \beta_j = 0$. Similar to $\boldsymbol{M}$, $\boldsymbol{M}^{(\setminus j)}$ is a matrix such that if $\|\boldsymbol{M}^{(\setminus j)}\boldsymbol{G}_{(\setminus j,\setminus j)} - \boldsymbol{I}\|_\infty$ is small, $S^{(debiased)}$ could be asymptotically unbiased. With high-dimensional data, the computation of $\boldsymbol{M}^{(\setminus j)}$ in Ning and Liu (2017) is based on the lasso or Dantzig selector (Candes and Tao, 2007), which is computationally intense.

Given that it is reasonable to debias the lasso estimate $\hat{\boldsymbol{\beta}}_\lambda$ using $\boldsymbol{M}\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda)/n$ in (2.26), it is intuitive that one can also debias $\boldsymbol{X}_j^\top \boldsymbol{X}_{\setminus j}\hat{\boldsymbol{\beta}}_\lambda^{(\setminus j)}$ using $\boldsymbol{X}_j^\top \boldsymbol{X}_{\setminus j}\boldsymbol{M}^{(\setminus j)}\boldsymbol{X}_{\setminus j}^\top(\boldsymbol{y} - \boldsymbol{X}_{\setminus j}\hat{\boldsymbol{\beta}}_\lambda^{(\setminus j)})/n$ in (2.27). In other words, although the debiased lasso estimator in (2.26) and the debiased score function in (2.27) are developed from different perspectives – the former results from inverting the stationary condition of (1.2), and the latter from a Taylor expansion of the score function – they are closely related.

2.6.1 Connection Between the Post-Selection OLS Estimator and Debiased Lasso Estimators

We will now argue that the post-selection OLS estimator $\bar{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda)}$ in (2.6), used by the naïve confidence intervals and the naïve score test, can be written in the form (2.26).

Lemma 2.6.1. *Suppose (***M1***) holds. Without loss of generality, assume that variables are ordered such that* $\hat{\mathcal{A}}_\lambda = \{1, \ldots, |\hat{\mathcal{A}}_\lambda|\}$. *Then we have*

$$\begin{bmatrix} \bar{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda)} \\ \mathbf{0} \end{bmatrix} = \hat{\boldsymbol{\beta}}_\lambda + \frac{1}{n} \begin{bmatrix} \left(\boldsymbol{G}_{(\hat{\mathcal{A}}_\lambda, \hat{\mathcal{A}}_\lambda)}\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \boldsymbol{X}^\top \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda\right). \tag{2.29}$$

The proof of Lemma 2.6.1 is given in Appendix A.7. By definition, $\hat{\boldsymbol{\beta}}_{\lambda, \setminus \hat{\mathcal{A}}_\lambda} = \mathbf{0}$. So we can rewrite (2.29) as

$$\begin{bmatrix} \bar{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda)} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_{\lambda, \hat{\mathcal{A}}_\lambda} \\ \mathbf{0} \end{bmatrix} + \frac{1}{n} \begin{bmatrix} \left(\boldsymbol{G}_{(\hat{\mathcal{A}}_\lambda, \hat{\mathcal{A}}_\lambda)}\right)^{-1} \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top \left(\boldsymbol{y} - \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda} \hat{\boldsymbol{\beta}}_{\lambda, \hat{\mathcal{A}}_\lambda}\right) \\ \mathbf{0} \end{bmatrix}. \tag{2.30}$$

Thus, $\bar{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda)}$ in (2.6) is a debiased lasso estimator that only debiases $\hat{\boldsymbol{\beta}}_{\lambda, \hat{\mathcal{A}}_\lambda}$, while leaving $\hat{\boldsymbol{\beta}}_{\lambda, \setminus \hat{\mathcal{A}}_\lambda}$ unchanged. Instead of requiring estimating the entire $p \times p$ matrix $\boldsymbol{M}$, as in LDPE and SSLasso – a computationally costly proposition – the naïve approaches require the much smaller matrix $(\boldsymbol{G}_{(\hat{\mathcal{A}}_\lambda, \hat{\mathcal{A}}_\lambda)})^{-1}$. This produces an asymptotically unbiased estimator without de-sparsifying the original lasso estimator.

2.6.2 Connection Between a Modification of the Naïve Score Test and the Decorrelated Score Test

In this section, we show that a slight modification of naïve score test can be written in the form (2.26). In Section 2.4, we defined the naïve score test statistic to be proportional to $S^j \equiv \boldsymbol{X}_j^T (\boldsymbol{I} - \boldsymbol{P}^{(\hat{\mathcal{A}}_\lambda \setminus \{j\})}) \boldsymbol{y}$ in (2.21). Alternatively, we could have defined the naïve score test statistic to be proportional to

$$S^0 \equiv \boldsymbol{X}_j^\top \left(\boldsymbol{I} - \boldsymbol{P}^{(\hat{\mathcal{A}}_\lambda^{(\setminus j)})}\right) \boldsymbol{y}, \tag{2.31}$$

where $\hat{\mathcal{A}}_\lambda^{(\setminus j)} \equiv \text{supp}(\hat{\boldsymbol{\beta}}_\lambda^{(\setminus j)})$, defined in (2.28). This alternative definition of the naïve score test statistic would have led to a result similar to that in Theorem 2.4.1, under suitable assumptions. However, the definition of the naïve score test presented in Section 2.4 has the benefit of greater computational simplicity: testing $H_{0,j} : \beta_j = 0$ for all $j = 1, \ldots, p$ requires

fitting one single lasso model to obtain $\hat{\mathcal{A}}_\lambda$. In contrast, if we defined the naïve score test using S^0 in (2.31), then we would have to fit p lasso models to get p $\hat{\mathcal{A}}_\lambda^{(\backslash j)}$, in order to test p hypotheses.

Lemma 2.6.2 shows that S^0 in (2.31) is closely related to the decorrelated score test (dScore) statistic of Ning and Liu (2017).

Lemma 2.6.2. *Suppose that (***M1***) holds. Then*

$$S^0 = \boldsymbol{X}_j^\top \left(\boldsymbol{y} - \boldsymbol{X}_{\backslash j}\hat{\boldsymbol{\beta}}_\lambda^{(\backslash j)}\right) - \frac{1}{n}\boldsymbol{X}_j^\top \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda^{(\backslash j)}} \left(\boldsymbol{G}_{(\hat{\mathcal{A}}_\lambda^{(\backslash j)}, \hat{\mathcal{A}}_\lambda^{(\backslash j)})}\right)^{-1} \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda^{(\backslash j)}}^\top \left(\boldsymbol{y} - \boldsymbol{X}_{\backslash j}\hat{\boldsymbol{\beta}}_\lambda^{(\backslash j)}\right). \quad (2.32)$$

The proof of Lemma 2.6.2 is given in Appendix A.8.

2.7 Discussion

In this chapter, we examined a very naïve two-step approach to high-dimensional inference:

1. Perform the lasso in order to select a small set of variables, $\hat{\mathcal{A}}_\lambda$.

2. Fit an OLS model using just the variables in $\hat{\mathcal{A}}_\lambda$, and make use of standard regression inference tools. Make no adjustment for the fact that $\hat{\mathcal{A}}_\lambda$ was selected based on the data.

It seems clear that this naïve approach is problematic, since we have peeked at the data twice, but are not accounting for this double-peeking in our analysis.

In this chapter, we have shown that under appropriate assumptions, $\hat{\mathcal{A}}_\lambda$ converges with high probability to a deterministic set, $\check{\mathcal{A}}_\lambda$. This key insight allows us to establish that the confidence intervals resulting from the aforementioned naïve two-step approach have asymptotically correct coverage, in the sense of (2.3). This constitutes a theoretical justification for the recent simulation findings of Leeb et al. (2015). Furthermore, we used this key insight in order to establish that the score test that results from the naïve two-step approach has asymptotically the same distribution as though the selected set of variables had been fixed in advance; thus, it can be used to test the null hypothesis $H_{0,j} : \beta_j = 0$, $j = 1, \ldots, p$.

Our simulation results corroborate our theoretical findings. In fact, we find essentially no difference between the empirical performance of these naïve proposals, and a host of other recent proposals in the literature for high-dimensional inference (Javanmard and Montanari, 2014a, Zhang and Zhang, 2014, van de Geer et al., 2014, Lee et al., 2016, Ning and Liu, 2017).

From a bird's-eye view, the recent literature on high-dimensional inference falls into two camps. The work of Wasserman and Roeder (2009), Meinshausen et al. (2009), Berk et al. (2013), Lee et al. (2016), Tibshirani et al. (2016) focuses on performing inference on the sub-model (2.2), whereas the work of Javanmard and Montanari (2013a, 2014a,b), Zhang and Zhang (2014), van de Geer et al. (2014), Ning and Liu (2017) focuses on testing hypotheses associated with (2.1). In this chapter, we have shown that the confidence intervals that result from the naïve approach can be used to perform inference on the sub-model (2.2), whereas the score test that results from the naïve approach can be used to test hypotheses associated with (2.1). Furthermore, we show that our post-selection approaches are closely related to the debiased tests in Section 2.6. Thus, the naïve approach to inference considered in this chapter serves to unify these two camps of high-dimensional inference.

In the era of big data, simple analyses that are easy to apply and easy to understand are especially attractive to scientific investigators. Therefore, a careful investigation of such simple approaches is worthwhile, in order to determine which have the potential to yield accurate results, and which do not. We do not advocate applying the naïve two-step approach described above in most practical data analysis settings: we are confident that in practice, our intuition is correct, and this approach will perform poorly when the sample size is small or moderate, and/or the assumptions are not met. However, in very large data settings, our results suggest that this naïve approach may indeed be viable for high-dimensional inference, and warrants further investigation.

We close with some technical comments. One reviewer brought up an interesting comment: Methods with folded-concave penalties (e.g., Fan and Li, 2001, Zhang, 2010a) require milder conditions to achieve variable selection consistency than the lasso, i.e., $\Pr[\hat{\mathcal{A}}_\lambda = \mathcal{A}] \rightarrow$

1. Inspired by this observation, we wonder whether Fan and Li (2001), Zhang (2010a) also require milder conditions to achieve $\Pr[\hat{\mathcal{A}}_\lambda = \check{\mathcal{A}}_\lambda] \to 1$. If so, then we could replace lasso with Fan and Li (2001), Zhang (2010a) in the variable selection step, and improve the robustness of the naïve approaches. We believe this could be a fruitful area of future research. In addition, extending the proposed theory and methods to generalized linear models and M-estimators may also be fruitful areas for future research.

Chapter 3

A SIGNIFICANCE TEST FOR GRAPH-CONSTRAINED ESTIMATION

3.1 Introduction

Interactions among genes, proteins and metabolites shed light into underlying biological mechanisms, and clarify their roles in carrying out cellular functions (Barabási and Oltvai, 2004, Zhu et al., 2007). This has motivated the development of many statistical methods to incorporate existing knowledge of biological networks into data analysis (see, e.g., Kong et al., 2006, Wei and Pan, 2008, Shojaie and Michailidis, 2009, 2010c, Michailidis, 2012). Such methods can lead to identification of novel biological mechanisms associated with the onset and progression of complex diseases (see, e.g., Barabási et al., 2011, Khatri et al., 2012).

External network information may be summarized using an undirected weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \boldsymbol{W})$, whose node set $\mathcal{V} = \{1, ..., p\}$ corresponds to p variables. The edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ of the graph encodes similarities among variables, in the sense that two vertices $j, k \in \mathcal{V}$ are connected with an edge $(j, k) \in \mathcal{E}$ if variables j and k are "similar" to each other. The level of similarity between neighboring nodes $(j, k) \in \mathcal{E}$ is captured by entries in the weight matrix, $W_{(j,k)} = W_{(k,j)}$. Such similarities can for instance correspond to interactions between genes or phylogenetic proximities of species.

A popular approach for incorporating network information is to encourage smoothness in coefficient estimates corresponding to neighboring nodes in the network using a *network smoothing penalty* (Li and Li, 2008, Slawski et al., 2010, Pan et al., 2010, Li and Li, 2010, Huang et al., 2011, Yang et al., 2012, Shen et al., 2012, Zhu et al., 2013). This approach can also be generalized to induce smoothness among similar variables defined based on a distance

matrix or "kernel" (Randolph et al., 2012), which, for instance, capture similarities among microbial communities according to lineages of a phylogenetic tree (see, e.g., Lozupone and Knight, 2005, Lozupone et al., 2007, Purdom, 2011, Fukuyama et al., 2012).

The smoothness induced by the network smoothing penalty can result in more accurate parameter estimations, particularly when the sample size n is small compared to the number of variables p. Sparsity-inducing penalties, e.g., the ℓ_1 penalty (Li and Li, 2008, Slawski et al., 2010, Li and Li, 2010) or the minimum convex penalty (MCP) (Huang et al., 2011), can then be used to select a subset of variables $\boldsymbol{X}$ associated with the outcome $\boldsymbol{y}$ for improved interpretability and reduced variability. It has been shown that, under appropriate assumptions, the combination of network smoothing and sparsity-inducing penalties can consistently select the subset of variables associated with the outcome (Huang et al., 2011). However, such procedures do not account for the uncertainty of the estimates, and in particular, do not provide p-values.

As mentioned in Chapter 2, a number of approaches have recently been proposed for formal hypothesis testing in penalized regression, including resampling and subsampling approaches (see, e.g., Bach, 2008, Meinshausen and Bühlmann, 2010, Chatterjee and Lahiri, 2011, Shah and Samworth, 2013), ridge test with deterministic design matrices (Bühlmann, 2013), debiased lasso tests (Javanmard and Montanari, 2013b,a, Zhang and Zhang, 2014, van de Geer et al., 2014, Javanmard and Montanari, 2014a,b, Ning and Liu, 2017) and post-selection lasso tests (Lee et al., 2016, Tibshirani et al., 2016). However, there are currently no inference procedures available for methods that incorporate external information using smoothing penalties. Inference procedures for least squares kernel machine (LSKM; Liu et al., 2007, Kwee et al., 2007) and sequence kernel association test (SKAT; Wu et al., 2010, 2011, Lee et al., 2012b,a, Ionita-Laza et al., 2013), on the other hand, test the global association of the outcome with variables and are hence not appropriate for testing the association of individual variables.

Another limitation of existing approaches that incorporate external network information, including those using network smoothing penalties, is their implicit assumption that the

network is accurate and informative. However, existing networks may be incomplete or inaccurate (Hart et al., 2006). As shown in Shojaie and Michailidis (2010b), such inaccuracies can severely impact the performance of network-based methods. Moreover, even if the network is accurate and complete, it is often unclear whether network connectivities correspond to similarities among corresponding coefficients, which is necessary for methods based on network smoothing penalties.

To address the above shortcomings, we propose a testing framework, the *Grace test*, which incorporates external network information into high-dimensional regression and corresponding inferences. The proposed framework builds upon the graph-constrained estimation (Grace) procedure of Li and Li (2008), Slawski et al. (2010), Li and Li (2010), and utilizes recent theoretical developments for the lasso estimator (Bickel et al., 2009, van de Geer and Bühlmann, 2009) and the ridge test (Bühlmann, 2013). As part of our theoretical development, we generalize the ridge test with fixed design to the setting with random design matrices $\boldsymbol{X}$. This generalization was suggested in the discussion of Bühlmann (2013) as a possible extension of the ridge test.

Our theoretical analysis shows that the proposed testing framework controls type-I error rate, regardless of the informativeness or accuracy of the incorporated network. We also show, both theoretically and using simulation experiments, that if the network is accurate and informative, the Grace test offers improved power over existing approaches that ignore such information. Finally, we propose an extension of the Grace test, called the Grace-ridge or *GraceR* test, for settings where the network may potentially be inaccurate or uninformative.

The rest of this chapter is organized as follows. In Section 3.2, we introduce the Grace estimation procedure and the Grace test. We also formally define the "informativeness" of a network. Section 3.3 investigates the power of the Grace test, in comparison to its competitors. In Section 3.4, we propose the Grace-ridge (GraceR) test for robust estimation and inference with potentially uninformative networks. We apply our methods to simulated data in Section 3.5 and to data from the Cancer Genome Atlas (TCGA) in Section 3.6. We end with a discussion in Section 3.7. Proofs of theoretical results are gathered in Appendix B.

3.2 The Grace Estimation Procedure and the Grace Test

3.2.1 The Grace Estimation Procedure

Let $\boldsymbol{L}$ be the matrix encoding the external information in an undirected weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \boldsymbol{W})$. In general, $\boldsymbol{L}$ can be any positive semi-definite matrix, or kernel, capturing the "similarity" between variables. In this chapter, however, we focus on the case where $\boldsymbol{L}$ is the graph Laplacian matrix,

$$L_{(j,k)} \equiv \begin{cases} d_j & \text{if } j = k \\ -W_{(j,k)} & \text{if } (j,k) \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases},$$

with $d_j \equiv \sum_{(j,k)\in\mathcal{E}} W_{(j,k)}$ denoting the degree of node j. We also assume that weights $W_{(j,k)}$ are nonnegative. However, the definition of Laplacian and the analysis in this chapter can be generalized to also accommodate negative weights (Chung, 1997).

Let $\boldsymbol{X} = [\boldsymbol{X}_1, ..., \boldsymbol{X}_p] \in \mathbb{R}^{n\times p}$ be the $n \times p$ design matrix and $\boldsymbol{y} \in \mathbb{R}^n$ be the outcome vector in the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}_n\left(\boldsymbol{0}, \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I}\right), \quad \boldsymbol{X} \sim_{i.i.d.} \mathcal{N}_p\left(\boldsymbol{0}, \boldsymbol{\Sigma}\right) \text{ for } i = 1, ..., n. \tag{3.1}$$

Multivariate normality of variables is commonly assumed in the analysis of biological networks, particularly, when estimating interactions among genes or proteins using GGMs (see, e.g., Dobra et al., 2004, de la Fuente et al., 2004). Interestingly, the underlying assumption of network smoothing penalties – that connected variables after scaling have similar associations with the outcome – is also related to the assumption of multivariate normality (Shojaie and Michailidis, 2010c). In this chapter, we assume $\boldsymbol{y}$ is centered and columns of $\boldsymbol{X}$ are centered and scaled, i.e., $\boldsymbol{1}^\top\boldsymbol{y} = 0$ and $\boldsymbol{1}^\top\boldsymbol{X}_j = 0$, $\boldsymbol{X}_j^\top\boldsymbol{X}_j = n$ for $j = 1, ..., p$. We denote the scaled Gramian matrix by $\boldsymbol{G} \equiv \boldsymbol{X}^\top\boldsymbol{X}/n$.

For a non-negative Grace tuning parameter, $\lambda_{\boldsymbol{L}}$, Grace solves the following optimization

problem:

$$\tilde{\boldsymbol{\beta}}^{\boldsymbol{L}}_{\lambda_{\boldsymbol{L}}} = \underset{\boldsymbol{b}\in\mathbb{R}^p}{\arg\min} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + \lambda_{\boldsymbol{L}} \boldsymbol{b}^\top \boldsymbol{L}\boldsymbol{b} \right\} = (n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L})^{-1} \boldsymbol{X}^\top \boldsymbol{y}. \tag{3.2}$$

When $\boldsymbol{L}$ is the Laplacian matrix, $\boldsymbol{b}^\top \boldsymbol{L}\boldsymbol{b} = \sum_{(j,k)\in\mathcal{E}} (b_j - b_k)^2 W_{(j,k)}$ (Huang et al., 2011). Hence, the Grace penalty $\boldsymbol{b}^\top \boldsymbol{L}\boldsymbol{b}$ encourages smoothness in coefficients of connected variables, according to weights of edges. Henceforth, we call $\boldsymbol{L}$ the penalty weight matrix.

For any tuning parameter $\lambda_{\boldsymbol{L}} > 0$, (3.2) has a unique solution if $(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L})$ is invertible. However, if $p > n$ and $rank(\boldsymbol{L}) < p$, this condition may not hold. With a Gaussian design $\boldsymbol{X} \sim_{i.i.d.} \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, it follows from Bai (1999) that if $\liminf_{n\to\infty} \phi^2_{\min}[\boldsymbol{\Sigma}] > 0$, and if there exists a sequence of index sets $\mathcal{C}_n \subset \{1, ..., p\}$, $\lim_{n\to\infty} |\mathcal{C}_n|/n < 1$, such that $\liminf_{n\to\infty} \phi^2_{\min}[\boldsymbol{L}_{(\setminus\mathcal{C}_n, \setminus\mathcal{C}_n)}] > 0$, then $(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L})$ is invertible with high probability. In this section, we hence assume that $(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L})$ is invertible. This condition is relaxed in Section 3.4, where we propose the more general Grace-ridge (GraceR) test.

As mentioned in the Section 3.1, several methods have been proposed to select the subset of relevant variables for Grace. For example, Li and Li (2008), Slawski et al. (2010), Li and Li (2010) added an ℓ_1 penalty to the Grace objective function. Huang et al. (2011) instead added the MCP and proposed the sparse Laplacian shrinkage (SLS) estimator. While these methods perform automatic variable selection, they do not provide measures of uncertainty, i.e., confidence intervals or p-values. In this chapter, we instead propose an inference procedure that provides p-values for estimated coefficients from (3.2). The resulting p-values can then be used to assess the significance of individual variables, and select a subset of relevant variables.

3.2.2 The Grace Test

Before introducing the Grace test, we present a lemma that characterizes the bias of the Grace estimation procedure.

Lemma 3.2.1. *For any $\lambda_{\boldsymbol{L}} > 0$, assume $(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L})$ is invertible. Then, given $\boldsymbol{X}$, $\tilde{\boldsymbol{\beta}}^{\boldsymbol{L}}_{\lambda_{\boldsymbol{L}}}$ as*

formulated in (3.2) *is an unbiased estimator of* $\boldsymbol{\beta}$ *if and only if* $\boldsymbol{L}\boldsymbol{\beta} = \mathbf{0}$. *Moreover,*

$$\left\| \mathbf{Bias}\left[\tilde{\boldsymbol{\beta}}^{\boldsymbol{L}}_{\lambda_{\boldsymbol{L}}} | \boldsymbol{X}\right] \right\|_2 \leq \frac{\lambda_{\boldsymbol{L}} \left\| \boldsymbol{L}\boldsymbol{\beta} \right\|_2}{\phi^2_{\min}\left[n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right]}. \tag{3.3}$$

Lemma 3.2.1 is proven in Appendix B.1. Because the bias of the Grace estimator depends directly on the magnitude of $\boldsymbol{L}\boldsymbol{\beta}$, we consider $\boldsymbol{L}$ to be informative if $\boldsymbol{L}\boldsymbol{\beta}$ is small. According to Lemma 3.2.1, the Grace estimator will be unbiased only if $\boldsymbol{\beta}$ lies in the space spanned by the eigenvectors of $\boldsymbol{L}$ with 0 eigenvalues. In reality, however, this condition cannot be checked from data. Thus, to control type-I error rate, we must adjust for this potential estimation bias.

Our testing procedure is motivated by the ridge test proposed in Bühlmann (2013), which we briefly discuss next. First, note that ridge is also a biased estimator of $\boldsymbol{\beta}$, and its *estimation bias* is negligible only if the ridge tuning parameter is close to zero. In addition to the estimation bias, Bühlmann (2013) also accounted for the *projection bias* of ridge regression for a *fixed* design matrix $\boldsymbol{X}$. This is because for fixed design matrices with $p > n$, $\boldsymbol{\beta}$ is not uniquely identifiable, as there are infinitely many $\boldsymbol{b} \in \mathbb{R}^p$ such that $\mathrm{E}[\boldsymbol{y}] = \boldsymbol{X}\boldsymbol{b}$. Using ridge regression, $\boldsymbol{\beta}$ is only estimable if it lies in the row space of $\boldsymbol{X}$, which is a proper subspace of $\mathbb{R}^p$ when $p > n$. If $\boldsymbol{\beta}$ does not lie in this subspace, the ridge estimated regression coefficient is indeed the projection of $\boldsymbol{\beta}$ onto the row space of $\boldsymbol{X}$, which is not identical to $\boldsymbol{\beta}$. This gives rise to the projection bias.

To account for these two types of biases, Bühlmann (2013) proposed to shrink the ridge estimation bias to zero by shrinking the ridge tuning parameter to zero, while controlling the projection bias using a stochastic bias bound derived from a lasso initial estimator. A side effect of shrinking the ridge tuning parameter to zero is that the variance of variables with high multi-collinearity could become large; this would hurt the statistical power of the ridge test.

In this chapter, we develop a test for random design matrices, which was suggested in the discussion of Bühlmann (2013) as a potential extension. With random design matrices, we do not incur any projection bias. This is because the regression coefficients in this case

are uniquely identifiable as $\boldsymbol{\Sigma}^{-1}\text{Cov}[\boldsymbol{X}, \boldsymbol{y}]$ under the joint distribution of $(\boldsymbol{y}, \boldsymbol{X})$. Here, $\boldsymbol{\Sigma}$ denotes the population covariance matrix of variables $\boldsymbol{X}$, and $\text{Cov}[\boldsymbol{X}, \boldsymbol{y}]$ is the population covariance between variables and the outcome; see Shao and Deng (2012) for a more elaborate discussion of identifiability for fixed and random design matrices.

To control type-I error rate of the Grace test, we adjust for the potential estimation bias using a stochastic bias bound derived from an initial estimator. By adjusting for the estimation bias using a stochastic upper bound, the Grace tuning parameter needs not be very small. Thus, the variance of Grace estimator is less likely to be unreasonably large; this results in improved power for the Grace test. Power properties of the Grace test are more formally investigated in Section 3.3. Next, we formally introduce our testing procedure.

Consider the null hypothesis $H_{0,j} : \beta_j = 0$ for some $j \in \{1, ..., p\}$. Let $\hat{\boldsymbol{\beta}}^{(init)}$ be an initial estimator of $\boldsymbol{\beta}$. We define the Grace test statistic as

$$\boldsymbol{z}^G = \tilde{\boldsymbol{\beta}}^{\boldsymbol{L}}_{\lambda_{\boldsymbol{L}}} + \lambda_{\boldsymbol{L}} \left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1} \boldsymbol{L}\hat{\boldsymbol{\beta}}^{(init)}, \tag{3.4}$$

where $\tilde{\boldsymbol{\beta}}^{\boldsymbol{L}}_{\lambda_{\boldsymbol{L}}}$ is the Grace estimator from (3.2) with tuning parameter $\lambda_{\boldsymbol{L}}$. Plugging in (3.2) and adding and subtracting $\lambda_{\boldsymbol{L}}(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L})^{-1}\boldsymbol{L}\boldsymbol{\beta}$, we can write

$$z^G_j = \beta_j + Z^G_j + \gamma^G_j, \qquad j = 1, ..., p, \tag{3.5}$$

where

$$Z^G_j | \boldsymbol{X} \sim \mathcal{N}\left(0, n\sigma^2_{\boldsymbol{\epsilon}} \left[\left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1} \boldsymbol{G} \left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1}\right]_{(j,j)}\right), \tag{3.6}$$

$$\boldsymbol{\gamma}^G \equiv \lambda_{\boldsymbol{L}} \left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1} \boldsymbol{L} \left(\hat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}\right). \tag{3.7}$$

In other words, $\boldsymbol{\gamma}^G$ is the vector of unknown bias of the Grace test statistic. Under the null hypothesis $H_{0,j} : \beta_j = 0$, $\text{E}[z^G_j | \boldsymbol{X}] = \gamma^G_j$, which is not necessarily equal to zero. Thus, we have to adjust for $\boldsymbol{\gamma}^G$ to make our procedure control type-I error rate. To adjust for it, we derive an asymptotic stochastic bias bound for γ^G_j such that under the null hypothesis, with probability tending to one,

$$\left|\gamma^G_j\right| \precsim \Gamma^G_j + \sqrt{\text{Var}\left[Z^G_j | \boldsymbol{X}\right]} \cdot \mathcal{O}_p\left(1\right). \tag{3.8}$$

Then, under the null hypothesis, with probability tending to one, $|z_j^G| \precsim |Z_j^G| + \Gamma_j^G$, which allows us to asymptotically control type-I error rate.

A necessary condition for the initial estimator $\hat{\boldsymbol{\beta}}^{(init)}$ is its ℓ_1 estimation accuracy, i.e., $\|\hat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}\|_1$. To complete our testing framework, we use the fact that under suitable conditions and with proper tuning parameter λ, described in Theorem 3.2.2, the ℓ_1 estimation error of the lasso, defined in (1.2), is asymptotically controlled (see, e.g., Bickel et al., 2009, van de Geer and Bühlmann, 2009). We thus use the lasso as the initial estimator for the Grace test in this chapter, i.e., $\hat{\boldsymbol{\beta}}^{(init)} = \hat{\boldsymbol{\beta}}_\lambda$. Theorem 3.2.2 then constructs a Γ_j^G that satisfies Condition (3.8). First, we present sufficient conditions for our proposal.

(**A0**) $(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L})$ is invertible.

(**A1**) $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{X} \sim_{i.i.d.} \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ for $i = 1, ..., n$ and $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma_{\boldsymbol{\epsilon}}^2 \boldsymbol{I})$.

(**A2**) Let $\mathcal{A} \equiv \{j : \beta_j \neq 0\}$ be the active set of $\boldsymbol{\beta}$ with cardinality $q \equiv |\mathcal{A}|$. Then

$$q = o\left(\left(\frac{n}{\log(p)}\right)^{\xi}\right)$$

for some $0 < \xi < 1/2$.

(**E**) The $(\mathcal{A}, 3, \boldsymbol{\Sigma})$-compatibility condition is met: for any $\boldsymbol{a} \in \mathbb{R}^p$ such that $\|\boldsymbol{a}_{\setminus\mathcal{A}}\|_1 \leq 3\|\boldsymbol{a}_{\mathcal{A}}\|_1$,

$$q\frac{\boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{a}}{\|\boldsymbol{a}_{\mathcal{A}}\|_1^2} \geq \phi^2 \tag{3.9}$$

with $\liminf_{n\to\infty} \phi^2 > 0$.

(**O**) The Grace tuning parameter, $\lambda_{\boldsymbol{L}}$, and the penalty weight matrix, $\boldsymbol{L}$, are such that

$$\frac{\left[\lambda_{\boldsymbol{L}} \left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1} \boldsymbol{L}\right]_{(j,j)}}{\sqrt{n\sigma_{\boldsymbol{\epsilon}}^2 \left[\left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1} \boldsymbol{G} \left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1}\right]_{(j,j)}}} = \mathcal{O}_p\left(\left(\frac{n}{\log(p)}\right)^{\frac{1}{2}-\xi}\right). \tag{3.10}$$

As discussed in Section 3.2.1, (**A0**) is required for uniqueness of the Grace estimator, and is shown to hold with probability tending to one under Gaussian design (Bai, 1999). (**A2**) is a standard assumption, which requires the number of relevant variables to not grow too fast, so that the signal is not overly diluted among those relevant variables. With $p = o(\exp(n))$, q can still grow to infinity as the sample size increases. The $(\mathcal{A}, 3, \boldsymbol{\Sigma})$-compatibility condition (Bühlmann and van de Geer, 2011) in (**E**) is closely related to the restricted eigenvalue assumption introduced in Bickel et al. (2009); see van de Geer and Bühlmann (2009) for the relationship between similar conditions. (**O**) is an optional condition, which can be relaxed at the cost of potential loss of power with finite samples; see Remark 3.2.3.

Theorem 3.2.2. *Suppose conditions (**A0**)-(**A2**), (**E**) and (**O**) are satisfied, and let* $\hat{\boldsymbol{\beta}}^{(init)} = \hat{\boldsymbol{\beta}}_\lambda$ *with tuning parameter* $\lambda \asymp \sqrt{\log(p)/n}$*, where* $\boldsymbol{\beta}_\lambda$ *is the lasso estimate defined in* (1.2)*. Let*

$$\Gamma_j^G \equiv \lambda_{\boldsymbol{L}} \left\| \left[(n\boldsymbol{G} + \lambda_{\boldsymbol{L}} \boldsymbol{L})^{-1} \boldsymbol{L} \right]_{(j,\setminus j)} \right\|_\infty \left(\frac{\log(p)}{n} \right)^{\frac{1}{2}-\xi}. \tag{3.11}$$

Then under the null hypothesis $H_{0,j} : \beta_j = 0$*, for any* $\alpha > 0$ *we have*

$$\lim_{n\to\infty} \Pr\left[\left| z_j^G \right| > \alpha \right] \leq \lim_{n\to\infty} \Pr\left[\left| Z_j^G \right| + \Gamma_j^G > \alpha \right]. \tag{3.12}$$

Remark 3.2.3. *If we instead define the stochastic error bound to be*

$$\lambda_{\boldsymbol{L}} \left\| \left[(n\boldsymbol{G} + \lambda_{\boldsymbol{L}} \boldsymbol{L})^{-1} \boldsymbol{L} \right]_j \right\|_\infty \left(\frac{\log(p)}{n} \right)^{\frac{1}{2}-\xi},$$

*we can relax (**O**) and still control the asymptotic type-I error rate; see the proof of Theorem 3.2.2 in Appendix B.2. However, as* $\lambda_{\boldsymbol{L}}/n \to \infty$*,* $(n\boldsymbol{G} + \lambda_{\boldsymbol{L}} \boldsymbol{L})^{-1} \lambda_{\boldsymbol{L}} \boldsymbol{L}$ *converges to a diagonal matrix, in which case* $\|[(n\boldsymbol{G} + \lambda_{\boldsymbol{L}} \boldsymbol{L})^{-1} \boldsymbol{L}]_j\|_\infty \gg \|[(n\boldsymbol{G} + \lambda_{\boldsymbol{L}} \boldsymbol{L})^{-1} \boldsymbol{L}]_{(j,\setminus j)}\|_\infty$*. This looser stochastic bound may result in lower power in finite samples.*

Theorem 3.2.2 is proven in Appendix B.2. It shows that regardless of the choice of $\boldsymbol{L}$, type-I error rate of the Grace test is asymptotically controlled. The stochastic bound Γ_j^G relies on the unknown sparsity parameter ξ. Following Bühlmann (2013), we suggest a small

value of ξ, and use $\xi = 0.05$ in the simulation experiments in Section 3.5 and real data example in Section 3.6.

Using (3.12), we can test $H_{0,j}$ using the asymptotically conservative two-sided p-value

$$p_j^G = 2\left(1 - \Phi_{\mathcal{N}}\left[\frac{\left(\left|\hat{\boldsymbol{z}}_j^G\right| - \Gamma_j^G\right)_+}{\sqrt{\mathrm{Var}\left[Z_j^G|\boldsymbol{X}\right]}}\right]\right). \tag{3.13}$$

Calculating p-values requires estimating $\sigma_{\boldsymbol{\epsilon}}^2$ and choosing a suitable tuning parameter $\lambda_{\boldsymbol{L}}$. We can estimate $\sigma_{\boldsymbol{\epsilon}}^2$ using any consistent estimator, such as the scaled lasso (Sun and Zhang, 2012). In the simulation experiments and real data example, we choose $\lambda_{\boldsymbol{L}}$ using 10-fold CV.

Note that, when simultaneously testing multiple hypotheses: $H_{0,J} : \beta_j = 0 : \forall j \in J \subseteq \{1, ..., p\}$ versus $H_{a,J} : \beta_j \neq 0 : \exists j \in J$, we may wish to control the false discovery rate (FDR). Because variables in the data could be correlated, test statistics on multiple variables may show arbitrary dependence structure. We thus suggest controlling FDR using the procedure of Benjamini and Yekutieli (2001). Alternatively, we can control FWER using, e.g., the method of Holm (1979).

3.3 Power of the Grace Test

In this section, we investigate power properties of the Grace test. Our first result describes sufficient conditions for detection of nonzero coefficients.

Theorem 3.3.1. *Assume conditions (***A0***)-(***A2***), (***E***) and (***O***) are met. If for some* $\lambda_{\boldsymbol{L}} > 0$, $0 < \alpha < 1$, $0 < \psi < 1$, *conditional on* $\boldsymbol{X}$, *we have*

$$|\beta_j| > 2\Gamma_j^G + \left(\Phi_{\mathcal{N}}^{-1}\left[1 - \frac{\alpha}{2}\right] + \Phi_{\mathcal{N}}^{-1}\left[1 - \frac{\psi}{2}\right]\right)\sqrt{\mathrm{Var}\left[Z_j^G|\boldsymbol{X}\right]}, \tag{3.14}$$

where $\Phi_{\mathcal{N}}^{-1}[\cdot]$ *is the quantile function of standard normal distribution, then using the same tuning parameter* $\lambda_{\boldsymbol{L}}$ *in the Grace test,*

$$\lim_{n\to\infty} \Pr\left[p_j^G \leq \alpha|\boldsymbol{X}\right] \geq 1 - \psi.$$

Theorem 3.3.1 is proven in Appendix B.3. Having established the sufficient conditions for detection of non-null hypotheses in Theorem 3.3.1, we next turn to comparing the power of the Grace test with its competitors: ridge test with small tuning parameters and no bias correction, and GraceI test, which is the Grace test with identity penalty weight matrix $\boldsymbol{I}$. The ridge test may be considered as a variant of the test proposed in Bühlmann (2013) without the adjustment of the projection bias – because we assume the design matrix is random, we incur no projection bias in the estimation procedure.

As indicated in Lemma 3.2.1, the estimation bias of the Grace procedure depends on the informativeness of the penalty weight matrix $\boldsymbol{L}$. When $\boldsymbol{L}$ is informative, we are able to increase the size of the tuning parameter, which shrinks the estimation variance without inducing a large estimation bias. Thus, with an informative $\boldsymbol{L}$, we are able to obtain a better prediction performance, as shown empirically in Li and Li (2008), Slawski et al. (2010), Li and Li (2010). In such setting, the larger value of the tuning parameter, e.g., as chosen by CV, also results in improved testing power, as discussed next.

Theorem 3.3.2 compares the power of the Grace test to its competitors in a simple setting of $p = 2$ predictors, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. In particular, this result identifies sufficient conditions under which the Grace test has asymptotically superior power. It also gives conditions for the GraceI test to have higher power than the ridge test. The setting of $p = 2$ predictors is considered mainly for ease of calculations, as in this case, we can directly derive closed form expressions of the corresponding test statistics. Similar results are expected to hold for $p > 2$ predictors, but require additional derivations and notations.

Assume $\boldsymbol{y} = \boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2 + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I})$, and $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ are scaled. Denote

$$\boldsymbol{L} \equiv \begin{bmatrix} 1 & l \\ l & 1 \end{bmatrix}, \qquad \boldsymbol{G} \equiv \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Theorem 3.3.2 considers the power for testing the null hypothesis $H_{0,j} : \beta_1 = 0$, in settings where $\beta_1 \neq 0$, without any constraints on β_2.

Theorem 3.3.2. *Suppose conditions (**A0**)-(**A2**), (**E**) and (**O**) are met. Let* $p_j^G(\lambda_{\boldsymbol{L}})$, $p_j^{GI}(\lambda_{\boldsymbol{I}})$

and $p_j^R(1)$ be the Grace, GraceI and ridge p-values, respectively, with tuning parameters $\lambda_{\boldsymbol{L}}$ for Grace, $\lambda_{\boldsymbol{I}}$ for GraceI, and 1 for ridge. Define

$$\Upsilon\left[h,l,\rho,|b|\right] \equiv \frac{\left(\left(h/n+1\right)^2-\left(\rho+lh/n\right)^2\right)\cdot|b|-\left(\log(p)/n\right)^{1/2-\xi}\cdot\left|\left(l-\rho\right)h/n\right|}{\sqrt{\left(1+2h/n\right)\left(1-\rho^2\right)+\left(h/n\right)^2\left(1+l^2-2l\rho\right)}}. \tag{3.15}$$

Then, conditional on the design matrix $\boldsymbol{X}$, under the alternative hypothesis $\beta_1 = b \neq 0$, the following statements hold with probability tending to one, as $n \to \infty$.

a) If $\lim_{n\to\infty} \Upsilon\left[\lambda_{\boldsymbol{L}},l,\rho,|b|\right] \geq \lim_{n\to\infty} \Upsilon_{p,n}\left[\lambda_{\boldsymbol{I}},0,\rho,|b|\right]$, *then* $\lim_{n\to\infty} p_1^G(\lambda_{\boldsymbol{L}})/p_1^{GI}(\lambda_{\boldsymbol{I}}) \leq 1$.

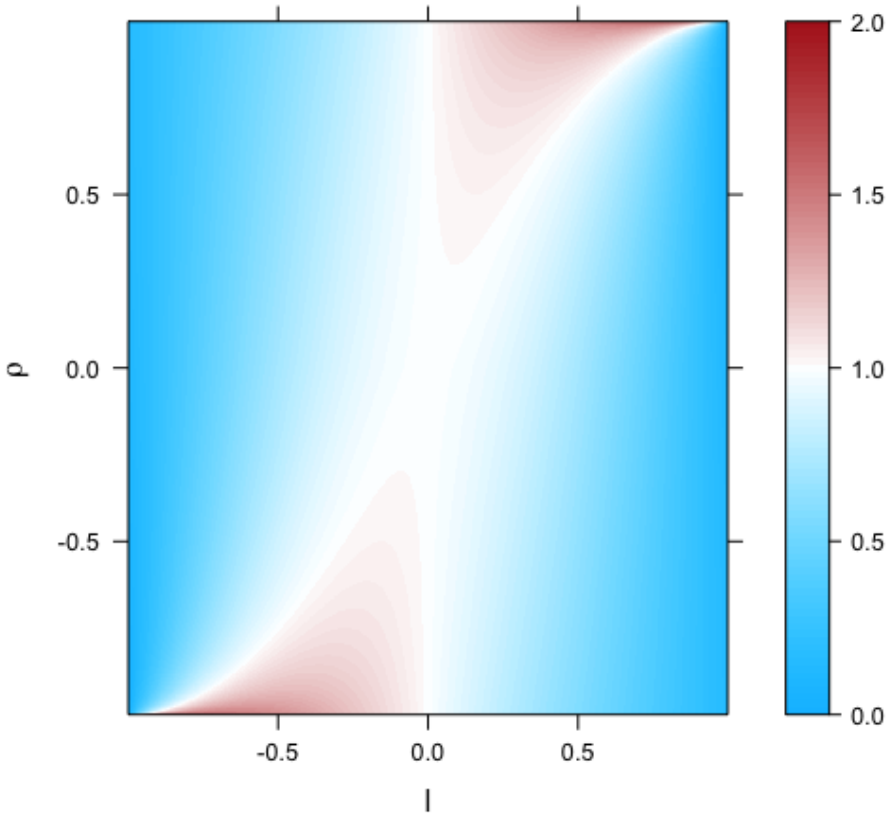
b) If $\lim_{n\to\infty} \Upsilon\left[\lambda_{\boldsymbol{L}},l,\rho,|b|\right] \geq \sqrt{1-\rho^2}|b|$, *then* $\lim_{n\to\infty} p_1^G(\lambda_{\boldsymbol{L}})/p_1^R(1) \leq 1$.

c) If $\lim_{n\to\infty} \Upsilon\left[\lambda_{\boldsymbol{I}},0,\rho,|b|\right] \geq \sqrt{1-\rho^2}|b|$, *then* $\lim_{n\to\infty} p_1^{GI}(\lambda_{\boldsymbol{I}})/p_1^R(1) \leq 1$.

Theorem 3.3.2 is proven in Appendix B.4. Observe that, as $\lambda_{\boldsymbol{L}}/n$ and $\lambda_{\boldsymbol{I}}/n$ diverge to infinity, both $\Upsilon[\lambda_{\boldsymbol{L}},l,\rho,|\beta_1|]$ and $\Upsilon[\lambda_{\boldsymbol{I}},0,\rho,|\beta_1|]$ approach infinity, and are greater than $\sqrt{1-\rho^2}|\beta_1|$. This implies that, on one hand, as shown in b) and c), for $\lambda_{\boldsymbol{L}}$ and $\lambda_{\boldsymbol{I}}$ sufficiently large, both the Grace and GraceI tests are asymptotically more powerful than the ridge test. On the other hand, as shown in a), we can only compare the powers of the Grace and GraceI tests under some constraints on their tuning parameters. With equal tuning parameters for Grace and GraceI, $\lambda_{\boldsymbol{L}} = \lambda_{\boldsymbol{I}}$, we can show that as $\lambda_{\boldsymbol{L}}/n = \lambda_{\boldsymbol{I}}/n \to \infty$, we have $\lim_{n\to\infty} \Upsilon[\lambda_{\boldsymbol{L}},l,\rho,|\beta_1|] \geq \lim_{n\to\infty} \Upsilon[\lambda_{\boldsymbol{I}},0,\rho,|\beta_1|]$ if $(1-l^2) \geq \sqrt{(1+l^2-2l\rho)}$. In this case, the Grace test is more powerful than the GraceI test if l is between 0 and l^*, where l^* is the unique root in $[-1,1]$ of the cubic equation $l^3 - 3l + 2\rho = 0$. Figure 3.1compares the powers of the Grace and GraceI tests with equal tuning parameters $\lambda_{\boldsymbol{L}}/n = \lambda_{\boldsymbol{I}}/n = 10$ and $\beta_1 = 1$. It can be seen that, the Grace test asymptotically outperforms the GraceI test when l is close to ρ with equally large tuning parameters. However, when l is far from ρ, the GraceI test could be more powerful. This observation, and the empirical results in Section 3.5 motivate the development of the GraceR test, introduced in Section 3.4.

A similar comparison for powers of the Grace and the ridge test, with $\lambda_{\boldsymbol{L}}/n = 10$ and $\beta_1 = 1$, is provided in Figure 3.2. These results suggest that, with large Grace tuning

Figure 3.1: The ratio of $\Upsilon[\lambda_{\boldsymbol{L}}, l, \rho, |\beta_1|]$ over $\Upsilon[\lambda_{\boldsymbol{I}}, 0, \rho, |\beta_1|]$ for different l and ρ with $\lambda_{\boldsymbol{L}}/n = \lambda_{\boldsymbol{I}}/n = 10$, $(\log(p)/n)^{1/2-\xi} = 0.25$ and $\beta_1 = 1$.

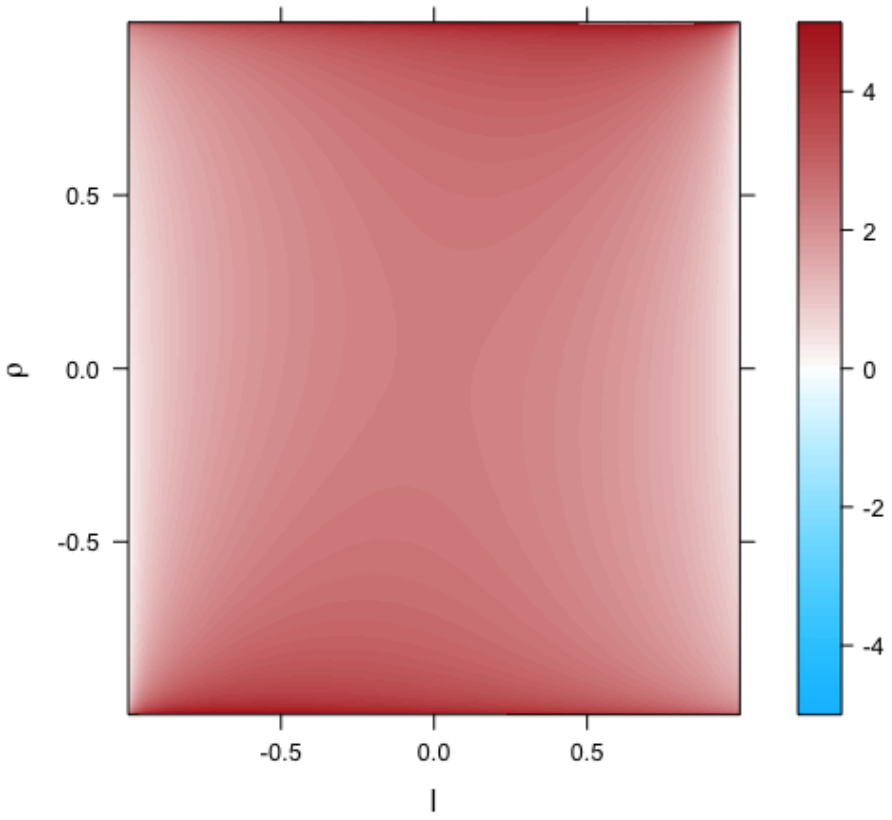


parameters, Grace substantially outperforms the ridge test in almost all scenarios. The result for the Grace and ridge comparison is similar with $\lambda_{\boldsymbol{L}}/n = 1$.

3.4 The Grace-Ridge (GraceR) Test

As discussed in Section 3.2, an informative $\boldsymbol{L}$ results in reduced bias of the Grace procedure, by choosing a larger tuning parameter $\lambda_{\boldsymbol{L}}$. The result in Theorem 3.3.2 goes beyond just the bias of the Grace procedure. It shows that for certain choices of $\boldsymbol{L}$, i.e., when l is close to the true correlation parameter ρ, the Grace test can have asymptotically superior power. This

Figure 3.2: The log-ratio of $\Upsilon[\lambda_{\boldsymbol{L}}, l, \rho, |\beta_1|]$ over $\sqrt{1-\rho^2}$ for different l and ρ with $\lambda_{\boldsymbol{L}}/n = 10$, $(\log(p)/n)^{1/2-\xi} = 0.25$ and $\beta_1 = 1$.



additional insight is obtained by accounting for, not just the bias of the Grace procedure, but also its variance, when investigating the power.

However, in practice, there is no guarantee that existing network information truly corresponds to similarities among coefficients, or is complete and accurate. To address this issue, we introduce the Grace-ridge (GraceR) test. The estimator used in GraceR incorporates two Grace-type penalties induced by $\boldsymbol{L}$ and $\boldsymbol{I}$:

$$\tilde{\boldsymbol{\beta}}^{\boldsymbol{L},\boldsymbol{I}}_{\lambda_{\boldsymbol{L}},\lambda_{\boldsymbol{I}}} = \underset{\boldsymbol{b}\in\mathbb{R}^p}{\arg\min} \left\{ \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}\|_2^2 + \lambda_{\boldsymbol{L}}\boldsymbol{b}^\top\boldsymbol{L}\boldsymbol{b} + \lambda_{\boldsymbol{I}}\boldsymbol{b}^\top\boldsymbol{b} \right\} = \left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L} + \lambda_{\boldsymbol{I}}\boldsymbol{I}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y}. \quad (3.16)$$

Using data-adaptive choices of tuning parameters $\lambda_{\boldsymbol{L}}$ and $\lambda_{\boldsymbol{I}}$, we expect this test to be as powerful as the Grace test if $\boldsymbol{L}$ is informative, and as powerful as the GraceI test, otherwise.

Another advantage of the GraceR over the Grace test is improved bias-variance tradeoff. If $\boldsymbol{L}$ is (almost) singular, the variance of the Grace test statistic, which depends on the eigenvalues of $(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L})$, could be large even for reasonably large $\lambda_{\boldsymbol{L}}$. Thus, even though our discussion in Section 3.2.1 shows that $(n\boldsymbol{G}+\lambda_{\boldsymbol{L}}\boldsymbol{L})$ is invertible with high probability, with finite samples, its smallest eigenvalue could be very small, if not zero. If $\boldsymbol{L}$ is informative, $\boldsymbol{L}\boldsymbol{\beta}$ and hence the bias in (3.3) are small. Thus, the rank-deficiency of $(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L})$ can be alleviated by choosing a large value of $\lambda_{\boldsymbol{L}}$. However, if $\boldsymbol{L}\boldsymbol{\beta}$ is non-negligible, choosing a large value of $\lambda_{\boldsymbol{L}}$ may result in a large bias, even larger than the ridge estimate, to the extent which may offset the benefit from the variance reduction. The finite sample type-I error rate of the Grace test may thus be controlled poorly. By incorporating an additional ℓ_2 penalty, we can better control the eigenvalues and achieve a better bias-variance trade-off.

The GraceR optimization problem leads to the following test statistic:

$$\boldsymbol{z}^{GR} = \tilde{\boldsymbol{\beta}}^{\boldsymbol{L},\boldsymbol{I}}_{\lambda_{\boldsymbol{L}},\lambda_{\boldsymbol{I}}} + \left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L} + \lambda_{\boldsymbol{I}}\boldsymbol{I}\right)^{-1}\left(\lambda_{\boldsymbol{L}}\boldsymbol{L} + \lambda_{\boldsymbol{I}}\boldsymbol{I}\right)\hat{\boldsymbol{\beta}}^{(init)}. \quad (3.17)$$

Similar to Section 3.2.2, we can write

$$z_j^{GR} = \beta_j + Z_j^{GR} + \gamma_j^{GR}, \qquad j = 1, ..., p, \quad (3.18)$$

where

$$Z_j^{GR}|\boldsymbol{X} \sim \mathcal{N}\left(0, n\sigma_{\boldsymbol{\epsilon}}^2 \left[\left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L} + \lambda_{\boldsymbol{I}}\boldsymbol{I}\right)^{-1} \boldsymbol{G} \left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L} + \lambda_{\boldsymbol{I}}\boldsymbol{I}\right)^{-1}\right]_{(j,j)}\right), \tag{3.19}$$

$$\boldsymbol{\gamma}^{GR} \equiv \left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L} + \lambda_{\boldsymbol{I}}\boldsymbol{I}\right)^{-1} \left(\lambda_{\boldsymbol{L}}\boldsymbol{L} + \lambda_{\boldsymbol{I}}\boldsymbol{I}\right) \left(\hat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}\right). \tag{3.20}$$

Similar to the Grace test in Section 3.2.2, we choose $\hat{\boldsymbol{\beta}}^{(init)}$ to be an initial lasso estimator, and derive an asymptotic stochastic bound for γ_j^{GR} such that $|\gamma_j^{GR}| \precsim \Gamma_j^{GR}$. Equation (3.13) is again used to obtain two-sided p-values for $H_{0,j}$. Theorems 3.4.1 and 3.4.2 parallel the previous results for the Grace test, and establish GraceR's asymptotic control of type-I error rate, and conditions for detection of non-null hypotheses. Proofs of these results are similar to Theorems 3.2.2 and 3.3.1, and are hence omitted. We first state an alternative to (**O**). This assumption can be justified using an argument similar to that for (**O**), and can also be relaxed with the cost of reduced power for the GraceR test.

- (**O'**): The GraceR tuning parameters, $\lambda_{\boldsymbol{L}}$ and $\lambda_{\boldsymbol{I}}$, and the penalty weight matrix, $\boldsymbol{L}$, are such that

$$\frac{\left[\left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L} + \lambda_{\boldsymbol{I}}\boldsymbol{I}\right)^{-1} \left(\lambda_{\boldsymbol{L}}\boldsymbol{L} + \lambda_{\boldsymbol{I}}\boldsymbol{I}\right)\right]_{(j,j)}}{\sqrt{n\sigma_{\boldsymbol{\epsilon}}^2 \left[\left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L} + \lambda_{\boldsymbol{I}}\boldsymbol{I}\right)^{-1} \boldsymbol{G} \left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L} + \lambda_{\boldsymbol{I}}\boldsymbol{I}\right)^{-1}\right]_{(j,j)}}} = \mathcal{O}_p\left(\left(\frac{n}{\log(p)}\right)^{\frac{1}{2}-\xi}\right). \tag{3.21}$$

Theorem 3.4.1. *Assume conditions (**A0**)-(**A2**), (**E**) and (**O'**) are met. The following Γ_j^{GR} satisfies the stochastic bound for GraceR.*

$$\Gamma_j^{GR} \equiv \left\| \left[\left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L} + \lambda_{\boldsymbol{I}}\boldsymbol{I}\right)^{-1} \left(\lambda_{\boldsymbol{L}}\boldsymbol{L} + \lambda_{\boldsymbol{I}}\boldsymbol{I}\right)\right]_{(j,\backslash j)} \right\|_\infty \left(\frac{\log(p)}{n}\right)^{\frac{1}{2}-\xi}. \tag{3.22}$$

Then, under the null hypothesis, for any $\alpha > 0$,

$$\lim_{n\to\infty} \Pr\left[\left|\boldsymbol{z}_j^{GR}\right| > \alpha\right] \leq \lim_{n\to\infty} \Pr\left[\left|Z_j^{GR}\right| + \Gamma_j^{GR} > \alpha\right]. \tag{3.23}$$

p-values for the GraceR test, p_j^{GR}, can be derived similarly as in the Grace test. We now show the sufficient condition for detection for the GraceR test.

Theorem 3.4.2. *Assume conditions (**A0**)-(**A2**), (**E**) and (**O'**) are met. If for some $\lambda_{\boldsymbol{L}} > 0$ and $\lambda_{\boldsymbol{I}} > 0$, conditional on $\boldsymbol{X}$, we have*

$$|\beta_j| > 2\Gamma_j^{GR} + \left(\Phi_{\mathcal{N}}^{-1}\left[1-\frac{\alpha}{2}\right] + \Phi_{\mathcal{N}}^{-1}\left[1-\frac{\psi}{2}\right]\right)\sqrt{\mathrm{Var}\left[Z_j^{GR}|\boldsymbol{X}\right]} \tag{3.24}$$

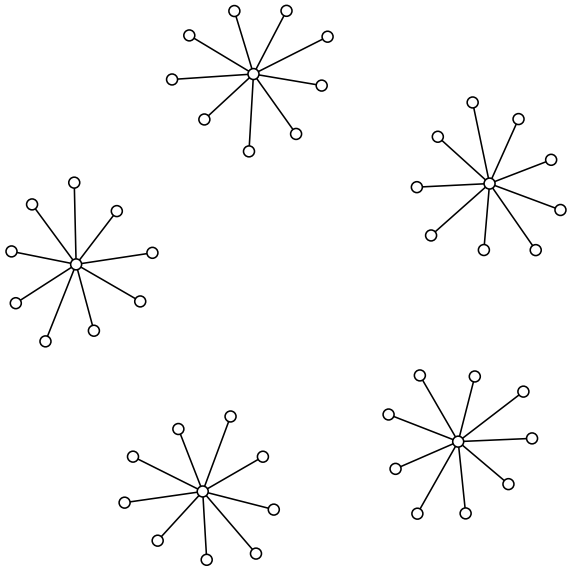
for some $0 < \alpha < 1$ and $0 < \psi < 1$. Then using the same $\lambda_{\boldsymbol{L}}$ and $\lambda_{\boldsymbol{I}}$ in the GraceR test, we get $\lim_{n\to\infty} \Pr[p_j^{GR} \leq \alpha \big| \boldsymbol{X}] \geq 1 - \psi$.

3.5 Simulation Experiments

In this section, we compare the Grace and GraceR tests with the ridge test (Bühlmann, 2013) with small tuning parameters, low-dimensional projection estimator (LDPE) for inference (Zhang and Zhang, 2014, van de Geer et al., 2014) and the GraceI test; both the ridge test and LDPE are implemented in the `hdi` R package. To this end, we consider a graph similar to Li and Li (2008), with 50 hub variables (genes), each connected to nine other satellite variables (genes). The nine satellite variables are not connected with each other, nor are variables in different hub-satellite clusters. In total the graph includes $p = 500$ variables and 450 edges. Figure 3.3 shows the graph structure used in the simulation study with five hub-satellite clusters. In the simulation study, we use 50 such hub-satellite clusters. We build the underlying true Laplacian matrix $\boldsymbol{L}^*$ according to the graph with all edge weights equal one.

To assess the effect of inaccurate or incomplete network information, we also consider variants of the Grace and GraceR tests with incorrectly specified graphs, where a number of randomly selected edges are added or removed. The number of removed or added (perturbed) edges relative to the true graph is NPE $\in$ {-225, -165, -70, -10, 0, 15, 135, 350, 600, 900, 1250, 1650, 2050, 3150}, with negative and positive numbers indicating removals and additions of edges, respectively. For example, NPE=-165 indicates 165 of the 450 edges in the true graph represented by $\boldsymbol{L}^*$ are randomly removed in the perturbed graph; denote the corresponding perturbed Laplacian matrix $\boldsymbol{L}$. This represents the case with incomplete network information. On the other hands, NPE = 350 indicates that in addition to the 450

Figure 3.3: An illustration of the hub-satellite graph structure with five hub-satellite clusters.



true edges in $\boldsymbol{L}^*$, we also randomly add 350 wrong edges to $\boldsymbol{L}$. The NPE values considered correspond to normalized spectral differences $\|\boldsymbol{L} - \boldsymbol{L}^*\|_2/\|\boldsymbol{L}^*\|_2$ equal to 0.85, 0.75, 0.50, 0.25, 0, 0.25, 0.50, 0.75, 1.00, 1.25, 1.50, 1.75, 2.00 and 2.65. Thus, the size of perturbation to the graph is roughly the same with NPE = -165 and 350. The perturbed penalty weight matrix $\boldsymbol{L}$ is then used in the Grace and GraceR tests. Since $(\boldsymbol{X}^\top\boldsymbol{X} + \lambda_{\boldsymbol{L}}\boldsymbol{L})$ may not be invertible, for Grace, we add a value of 0.01 to the diagonal entries of $\boldsymbol{L}$ to make it positive definite. No such correction is needed for GraceR and GraceI because of the presence of ℓ_2 penalty.

In each simulation replicate, we generate $n = 100$ independent samples, where for the 50 hub variables in each sample, $x_k^{hub} \sim_{i.i.d.} \mathcal{N}(0,1)$, $k = 1, ..., 50$, and for the 9 satellite variables in the k-th hub-satellite cluster, $x_l^{hub_k} \sim_{i.i.d.} N(0.9 \times x_k^{hub}, 0.9)$, $l = 1, ..., 9$, $k = 1, ..., 50$. This is equivalent to simulating $\boldsymbol{X} \sim_{i.i.d.} \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = (\boldsymbol{L}^* + 0.11 \times \boldsymbol{I})^{-1}$, i.e., $\boldsymbol{L}^*$ corresponds to the partial covariance structure of the variables.

We consider a sparse model in which variables in the first hub-satellite cluster are equally

associated with the outcome, and those in the other 49 clusters are not. Specifically, we let

$$\boldsymbol{\beta} \equiv \frac{1}{\sqrt{10}}(\underbrace{1,...,1}_{10},\underbrace{0,...,0}_{p-10})^\top.$$

We then simulate $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma_\epsilon^2 \boldsymbol{I})$, and consider $\sigma_{\boldsymbol{\epsilon}} \in \{9.5, 6.3, 4.8\}$ to produce expected $R^2 = 1 - \sigma_{\boldsymbol{\epsilon}}^2/\text{Var}[\boldsymbol{y}] \in \{0.1, 0.2, 0.3\}$.

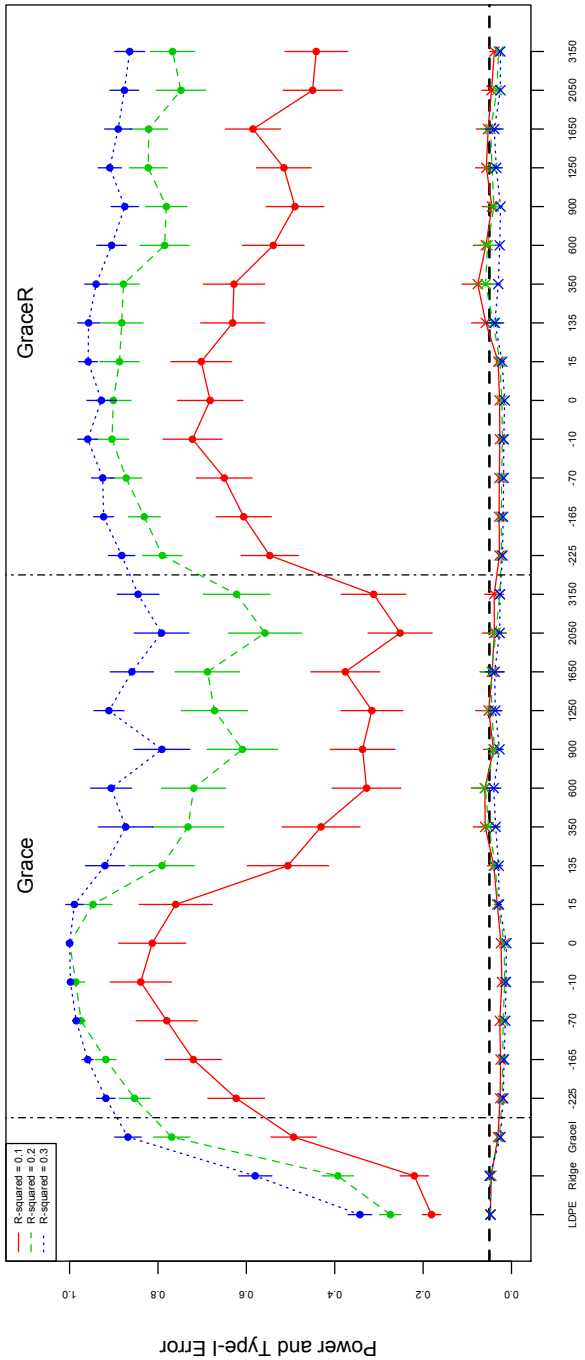
Throughout the simulation iterations, $\boldsymbol{L}^*$ and $\boldsymbol{\beta}$ are kept fixed, and $\boldsymbol{L}$, $\boldsymbol{X}$ and $\boldsymbol{\epsilon}$ are randomly generated in each repetition. We set the sparsity parameter $\xi = 0.05$, and lasso tuning parameter $\lambda = 4\sigma_{\boldsymbol{\epsilon}}\sqrt{3\log(p)/n}$, where $\sigma_{\boldsymbol{\epsilon}}$ is estimated using the scaled lasso (Sun and Zhang, 2012). As suggested in Bühlmann (2013), the tuning parameter for the ridge test is set to 1. Tuning parameters for LDPE, Grace, GraceR and GraceI are chosen by 10-fold CV. We use two-sided significance level $\alpha = 0.05$ and calculate the average and standard error of powers from ten non-zero coefficients and type-I error rates of each test from 490 zero coefficients. Figure 3.4 summarizes the mean power and type-I error rate of tests across $B = 100$ simulated data sets, along with the corresponding 95% confidence intervals.

Comparing the power of the tests, it can be seen that the Grace test with correct choices of $\boldsymbol{L}$ (NPE = 0) results in highest power. The performance of the Grace test, however, deteriorates as $\boldsymbol{L}$ becomes less accurate. The performance of the GraceR test is, on the other hand, more stable. It is close to the Grace test when the observed $\boldsymbol{L}$ is close to the truth, and is roughly as good as the GraceI test when $\boldsymbol{L}$ is significantly inaccurate. As expected, our testing procedures asymptotically control type-I error rate, in that observed type-I error rates are not significantly different from $\alpha = 0.05$.

3.6 Analysis of TCGA Prostate Cancer Data

We examine the Grace and GraceR tests on a prostate adenocarcinoma dataset from TCGA collected from prostate tumor biopsies. After removing samples with missing measurements, we obtain a dataset with $n = 321$ samples. For each sample, the prostate-specific antigen (PSA) level and the RNA sequences of 4739 genes are available. Genetic network information for these genes is obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG)

Figure 3.4: Comparison of powers and type-I error rates of different testing methods, along with their 95% confidence bands. Filled circles (•) corresponds to powers, whereas crosses (×) are type-I error rates. Numbers on x-axis for Grace and GraceR tests refer to the number of perturbed edges (NPE) in the network used for testing.



(Ogata et al., 1999, Kanehisa and Goto, 2000), resulting in a dataset with $p = 3450$ genes and $|\mathcal{E}| = 38541$ edges.

We center the outcome and center and scale the variables. For the Grace and GraceR tests, we set the sparsity parameter $\xi = 0.05$ and lasso tuning parameter $\lambda = 4\sigma_{\boldsymbol{\epsilon}}\sqrt{3\log(p)/n}$, where $\sigma_{\boldsymbol{\epsilon}}$ is estimated using the scaled lasso (Sun and Zhang, 2012). We control FDR at level 0.05 using the method of Benjamini and Yekutieli (2001).

To increase the chance of selecting "hub" genes, we use the normalized Laplacian matrix $\boldsymbol{L}^{(norm)} = \mathbf{D}^{-1/2}\boldsymbol{L}\mathbf{D}^{-1/2}$, where $\mathbf{D}$ is the diagonal degree matrix for the KEGG network with edge weights set to one. The Grace penalty induced by the normalized Laplacian matrix encourages smoothness of coefficient estimates based on the degrees of respective nodes, $\boldsymbol{b}^\top \boldsymbol{L}^{(norm)}\boldsymbol{b} = \sum_{(j,k)\in\mathcal{E}}(b_j/\sqrt{d_j} - b_k/\sqrt{d_k})^2 W_{(j,k)}$ (Li and Li, 2008). We add 0.001 to the diagonal entries of $\boldsymbol{L}^{(norm)}$ to induce positive definitiveness in the Grace test.

As shown in Figure 3.5(a), the Grace test with tuning parameter selected by 10-fold CV identifies 54 genes that are associated with PSA level. They consist of 42 histone genes, 11 histone deacetylase (HDAC) genes and the paired box gene 8 (PAX8). Histone and HDAC genes are densely connected in the KEGG network. With the network smoothing penalty, the Grace regression coefficients of histone and HDAC genes are all positive with a similar magnitude. Previous research indicates that histone genes are associated with the occurrence, clinical outcomes and recurrence of prostate cancer (Seligson et al., 2005, Ke et al., 2009). The pathological role of HDAC genes on the onset and progression of prostate cancer have also been previously studied (Halkidou et al., 2004, Chen et al., 2007, Abbas and Gupta, 2008).

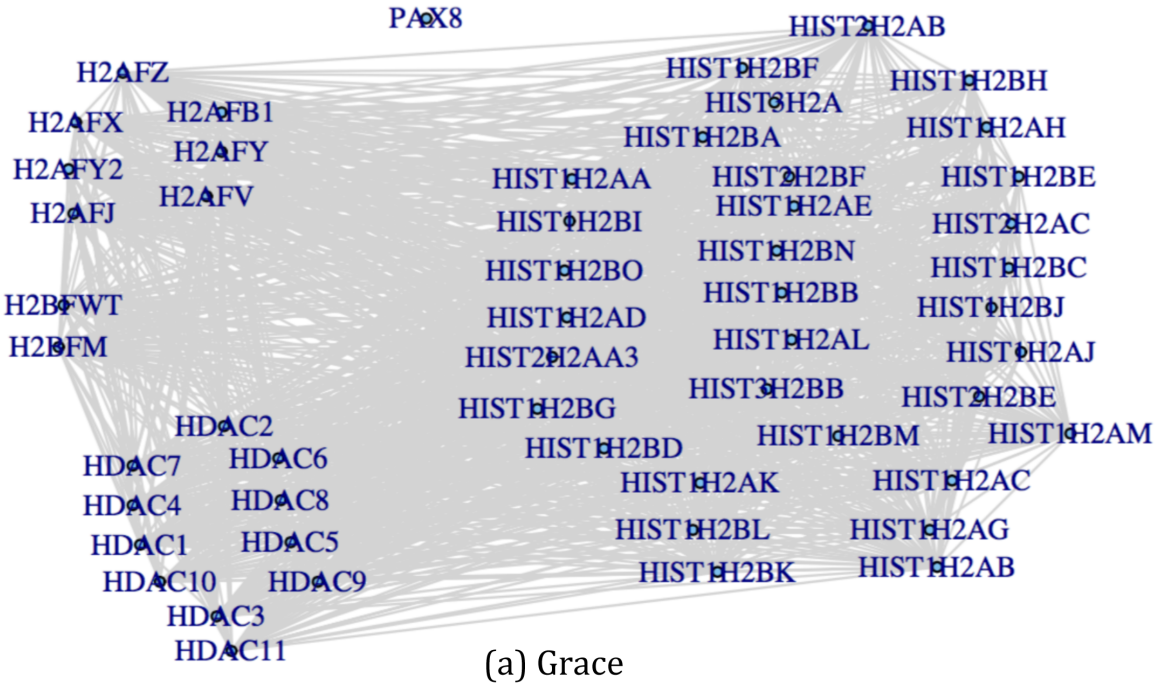
Figure 3.5(b) shows the result for the GraceR test. GraceR identifies five histone genes, which are also identified by the Grace test. In addition, GraceR identifies 11 genes that are not identified by Grace. Prior work has identified nine of those 11 genes to be associated with PSA level or the severity and stage of cancer. Specifically, prior work shows that the expression of ribonucleoside-diphosphate reductase subunit M2 (RRM2) is associated with higher Gleason scores, which correlate with the severity of prostate cancer (Huang

et al., 2014). Protein arginine methyltransferase 1 (PRMT1) may also have an effect on the proliferation of prostate cancer cells (Yu et al., 2009). Activation of olfactory receptors (OR) prevents proliferation of prostate cancer cells (Neuhaus et al., 2009). Interferon-γ (IFNG) plays a role in the differentiation of human prostate basal-epithelial cells (Untergasser et al., 2005). IFNG is connected to the interleukin receptor 22 α1 (IL22RA1), the role of which related to prostate cancer is unknown. However, several earlier studies point out the associations between prostate cancer and several other interleukin receptors in the Janus kinase and signal transducer and activator of transcription (JAK-STAT) activating family, including IL 6, 8, 11, 13 and 17 genes (Culig et al., 2005, Inoue et al., 2000, Campbell et al., 2001, Maini et al., 1997, Zhang et al., 2012). Cell-division cycle genes (CDC) may also be associated with various cancers. The association between collagen type 2 α1 (COL2A1) and prostate cancer is also not known, but other collagen genes, including type 1 $\alpha 2\beta 1$, type 4 α5 and α6, have been shown to be associated with prostate cancer progression (Hall et al., 2008, Dehan et al., 1997). Although the association between phosphate cytidylyltransferase 1 choline-α (PCYT1A) and prostate cancer or PSA level is not known, Vaezi et al. (2014) show that PCYT1A is a prognostic factor in survival for patients with lung and head and neck squamous cell carcinomas.

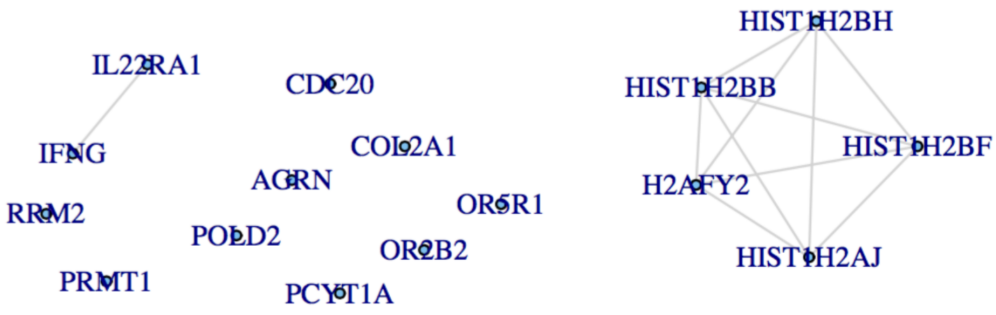
As a comparison, the GraceI test with 10-fold CV identifies 16 disconnected genes, 11 of which are also identified by the GraceR test. Ridge test (Bühlmann, 2013) with ridge tuning parameter 1 identifies four disconnected genes, which are also identified by the GraceR test. LDPE with tuning parameters chosen by 10-fold CV identifies ten disconnected genes. Seven of these genes are identified by GraceR and two by Grace.

We also investigate the stability of the Grace test to different tuning parameters and network mis-specifications. Figure 3.6 shows the number of significant genes identified by the Grace test in the TCGA data against various values of $\lambda_{\boldsymbol{L}}$. The result indicates that the number of genes found by the Grace test is relatively stable for a range of tuning parameters around the CV choice. On the other hand, very few genes are identified when the tuning parameter is too small or too large. This is because, with small tuning parameters, the

Figure 3.5: Results of analysis of TCGA prostate cancer data using the a) *Grace* and b) *GraceR* tests after controlling for FDR at 0.05 level. In each case, genes found to be significantly associated with PSA level are shown, along with their interactions based on information from KEGG.

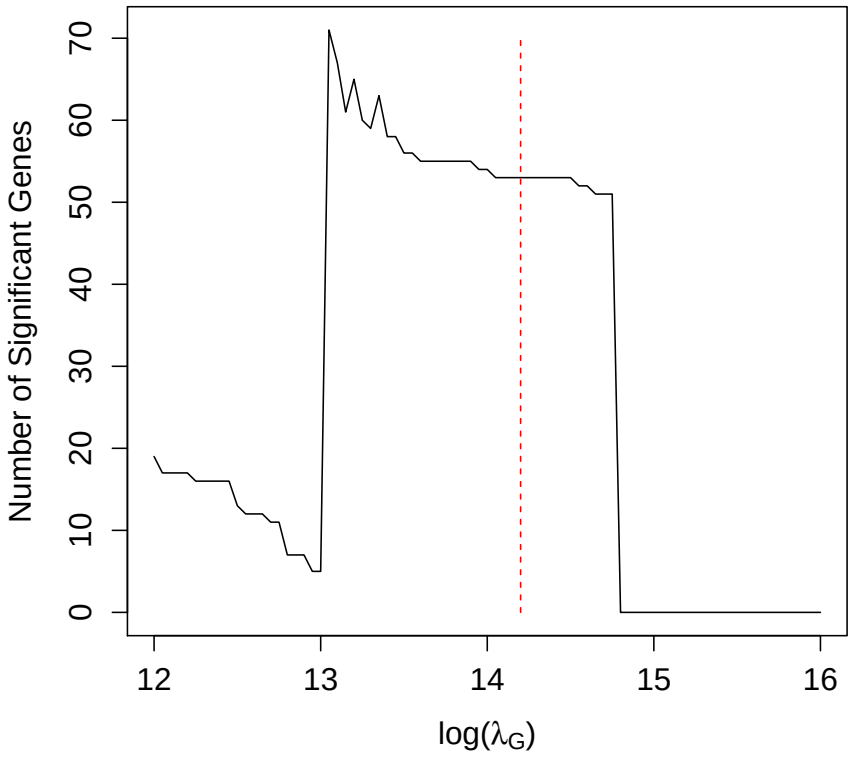


(a) Grace



(b) GraceR

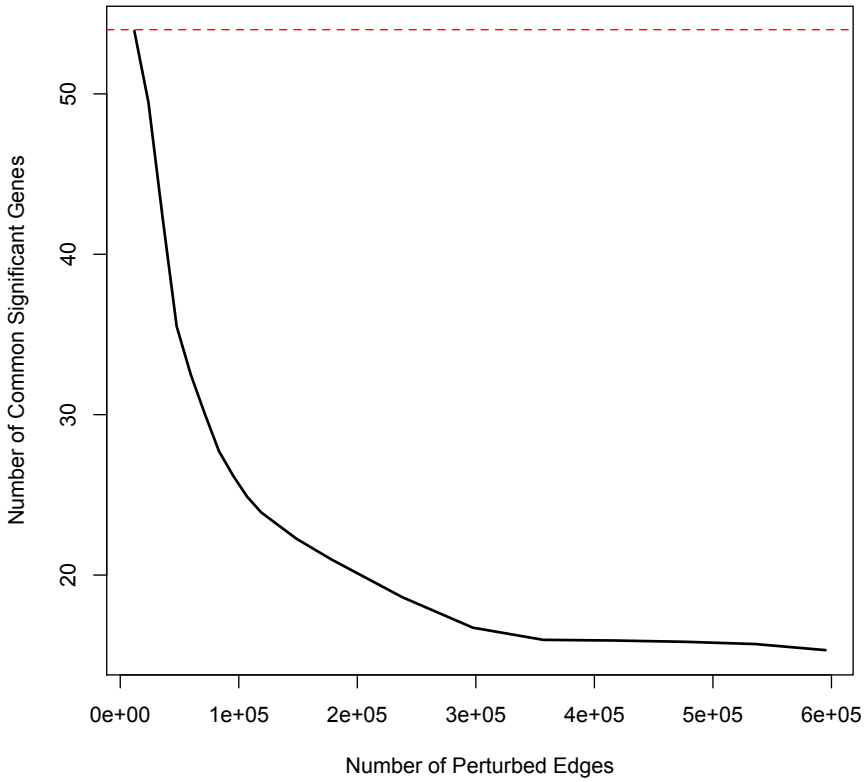
Figure 3.6: Number of genes identified by the Grace test in the TCGA data against the tuning parameter of the Grace test, $\lambda_{\boldsymbol{L}}$. The red dashed line corresponds to the choice made by 10-fold CV ($\lambda_{\boldsymbol{L}} = \exp(14.2)$).



variance is large and thus no gene is statistically significant. On the other hand, with large tuning parameters, the stochastic bound Γ_j^G dominates z_j^G. Note that the above result of power does not contradict Theorem 3.3.2, which shows that the *asymptotic* power of the Grace test improves as we use larger $\lambda_{\boldsymbol{L}}$. A vital condition for Theorem 3.3.2 to hold is $\|\hat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}\|_1 = o_p(1)$.

To examine whether the result of the Grace test on the TCGA data is sensitive to the KEGG network structure, we randomly change the connectivity of m node pairs in the KEGG network and form the new perturbed network $\mathcal{G}^{(perturbed)}$, $|\mathcal{E}\Delta\mathcal{E}^{(perturbed)}| = m$, where

Figure 3.7: Number of genes that are found significant with both the KEGG network and the perturbed network against the number of perturbed edges. The red dashed line represents the number of genes identified by the Grace test with the KEGG network.



Δ is the symmetric difference operator between two sets. In other words, for m randomly selected node pairs (a_i, b_i), $i = 1, ..., m$, if there is an edge (a_i, b_i) in the KEGG network, we remove it in the perturbed network; otherwise, we add an edge in the perturbed network. In our examination, m ranges from $10,000$ to $600,000$. Note that there are 38,541 edges in the original KEGG network. We counted the number of genes that are significant using both networks. The result shown in Figure 3.7 is an average of 50 independent replications.

Finally, we compare the prediction performance by Grace, GraceR, GraceI and lasso with tuning parameters chosen by 10-fold CV, as well as ridge with tuning parameter equal to 1.

Table 3.1: Prediction performance of the Grace, GraceR, GraceI, ridge ($\lambda_{\boldsymbol{I}} = 1$) and lasso. The performance metric is the sum of 10-fold CV prediction error rate (CVER).

	Grace	GraceR	GraceI	Ridge	Lasso
CVER	3473	3411	3418	3917	3546

The result is shown in Table 3.1. GraceR produces the smallest CV prediction error, followed closely by GraceI and Grace. This result may indicate the KEGG network information is in fact informative in prediction.

3.7 Discussion

In this chapter, we proposed the Grace and GraceR tests that incorporate external graphical information regarding the similarity between variables. Such external information is presented in the form of a penalty weight matrix $\boldsymbol{L}$, which is considered to be the (normalized) graph Laplacian matrix in this chapter. However, any positive semi-definite matrix can be used as $\boldsymbol{L}$. The proposed inference framework thus allows researchers in different fields to incorporate relevant external information through $\boldsymbol{L}$. For example, we can use various distance and kernel metrics that measure the (dis)similarity between species in phylogenetic studies. We can also use the adaptive graph Laplacian matrix (Li and Li, 2010) so that coefficients of negatively correlated variables are penalized to have the opposite signs. Regardless of the choice of $\boldsymbol{L}$, our proposed procedures asymptotically control type-I error rate; the power of the Grace test, however, depends on the informativeness of $\boldsymbol{L}$. The power of the GraceR test is on the other hand less dependent on the choice of $\boldsymbol{L}$.

In this chapter, we proposed to use the lasso as the initial estimator. Another potential choice for the initial estimator is the Dantzig selector (Candes and Tao, 2007). Koltchinskii (2009), Bickel et al. (2009) showed that with slightly different conditions than presented in this chapter, Dantzig selector was also able to produce coefficient estimates that achieved the required estimation accuracy.

The Grace test introduced in this chapter is not scale invariant. That is, the Grace test with the same tuning parameter could produce different p-values with data $(\boldsymbol{y}, \boldsymbol{X})$ and $(k\boldsymbol{y}, \boldsymbol{X})$, where $k \neq 1$ is a constant. This is clear as the test statistic z_j^G depends on the outcome $\boldsymbol{y}$ whereas the stochastic bound Γ_j^G does not. To make the Grace and GraceR tests scale invariant, we can simply choose the tuning parameter for our lasso initial estimator to be $\lambda = C\sigma_{\boldsymbol{\epsilon}}\sqrt{\log(p)/n}$ with a constant $C > 2\sqrt{2}$. Sun and Zhang (2012) showed that the lasso was scale invariant in this case. We would also need to use scaled invariant stochastic bounds $\Gamma_j^{G'} \equiv \sigma_{\boldsymbol{\epsilon}}\Gamma_j^G$ and $\Gamma_j^{GR'} \equiv \sigma_{\boldsymbol{\epsilon}}\Gamma_j^{GR}$ in our Grace and GraceR tests. Note that multiplying any constant in Γ_j^G and Γ_j^{GR} does not change our asymptotic control of type-I error rate.

In this chapter, CV is used to choose tuning parameters of the Grace and GraceR tests. However, CV does not directly maximize the power of these tests. Boonstra et al. (2015) presented several common methods for tuning parameter selection for ridge regression. Selection of tuning parameters for optimal testing performance can be a fruitful direction of future research. Another useful extension of the proposed framework is its adaptation to GLMs.

The Grace and GraceR tests are implemented in the R package **Grace**, available on the Comprehensive R Archive Network (CRAN).

Chapter 4

DIFFERENTIAL CONNECTIVITY ANALYSIS: TESTING DIFFERENCES IN STRUCTURES OF HIGH-DIMENSIONAL NETWORKS

4.1 Introduction

Changes in biological networks, such as brain connectivity and gene regulatory networks, have been associated with the onset and progression of complex diseases (see, e.g., Bassett and Bullmore, 2009, Barabási et al., 2011). Locating differentially connected nodes or edges in the network of diseased individuals compared to healthy subjects—referred to as *differential network biology* (Ideker and Krogan, 2012)—can help researchers delineate underlying disease mechanism. Such *network-based biomarkers* can also serve as effective diagnostic tools and guide new therapies. In this chapter, we propose a novel hypothesis testing framework for identifying differentially connected nodes or edges in two networks.

Consider two networks $\mathcal{G}^{\mathrm{I}} \equiv (\mathcal{V}, \mathcal{E}^{\mathrm{I}})$ and $\mathcal{G}^{\mathrm{II}} \equiv (\mathcal{V}, \mathcal{E}^{\mathrm{II}})$, with edge sets $\mathcal{E}^{\mathrm{I}}$ and $\mathcal{E}^{\mathrm{II}}$ and common nodes $\mathcal{V}$. For example, in a brain connectivity network, $\mathcal{V}$ denotes a set of regions of interest (ROI), and edges among ROIs may correspond to functional or structural connectives. While one may be interested in testing *global* differences between $\mathcal{G}^{\mathrm{I}}$ and $\mathcal{G}^{\mathrm{II}}$, scientists are often interested in identifying differentially connected nodes or edges. For $m \in \{\mathrm{I}, \mathrm{II}\}$, let

$$ne_j^m \equiv \{k \neq j : (j,k) \in \mathcal{E}^m\}, \tag{4.1}$$

be the neighborhood of node j in network G^m and denote by

$$d_{jk}^m \equiv \mathbb{1}\left[(j,k) \in \mathcal{E}^m\right], \tag{4.2}$$

the *dyad* of node-pair $(j,k), j \neq k$, where $\mathbb{1}[\cdot]$ denotes the indicator function. Patterns of differential connectivity in the two networks can then be identified by examining the null hypotheses

$$H^*_{0,j} : ne^{\mathrm{I}}_j = ne^{\mathrm{II}}_j \qquad \text{and} \qquad H^*_{0,jk} : d^{\mathrm{I}}_{jk} = d^{\mathrm{II}}_{jk}. \tag{4.3}$$

The first null hypothesis, $H^*_{0,j}$, concerns *differential connectivity of nodes*: Under $H^*_{0,j}$, node j is connected to the same nodes in both networks. On the other hand, the second null hypothesis, $H^*_{0,jk}$, concerns *differential connectivity of edges*: Under this null, node j and k are either connected in both networks, i.e., $d^{\mathrm{I}}_{jk} = d^{\mathrm{II}}_{jk} = 1$, or not connected in either network, i.e., $d^{\mathrm{I}}_{jk} = d^{\mathrm{II}}_{jk} = 0$. The set of hypotheses $H^*_{0,j}$ and $H^*_{0,jk}$ form a hierarchy, in the sense that $ne^{\mathrm{I}}_j = ne^{\mathrm{II}}_j$ implies $d^{\mathrm{I}}_{jk} = d^{\mathrm{II}}_{jk}$ for all $k \neq j$.

In most applications, the edge sets $\mathcal{E}^{\mathrm{I}}$ and $\mathcal{E}^{\mathrm{II}}$ are estimated from data, based on similarities/dependencies between variables. In particular, Gaussian graphical models (GGMs) are commonly used to estimate biological networks (e.g., Krumsiek et al., 2011). Let $\boldsymbol{X}^{\mathrm{I}}$ and $\boldsymbol{X}^{\mathrm{II}}$ be two Gaussian datasets of size $n^{\mathrm{I}} \times p$ and $n^{\mathrm{II}} \times p$ containing measurements of the same set of variables $\mathcal{V}$ (with $p = |\mathcal{V}|$) in populations I and II, respectively. The data may be high-dimensional, i.e., $p > \max\{n^{\mathrm{I}}, n^{\mathrm{II}}\}$. In GGMs, the edge set $\mathcal{E}^m$ for $m \in \{\mathrm{I}, \mathrm{II}\}$ is defined based on the population precision matrices of variables in $\boldsymbol{X}^m$, which we denote as $\boldsymbol{\Omega}^m$. More specifically, $(j,k) \in \mathcal{E}^m$ if and only if $\Omega^m_{(j,k)} \neq 0$.

To identify differential connectivities in two networks, we may naïvely eyeball the differences in two GGMs estimated using single network estimation methods (e.g., Meinshausen and Bühlmann, 2006, Yuan and Lin, 2007, Friedman et al., 2008), or joint estimation methods (e.g., Guo et al., 2011, Danaher et al., 2014, Zhao et al., 2014, Peterson et al., 2015, Saegusa and Shojaie, 2016); see Drton and Maathuis (2017) for a review of GGM estimation methodologies. However, existing estimation approaches do not provide measures of uncertainty, e.g., p-values, and are thus of limited utility in drawing formal scientific conclusions.

A number of recent approaches provide confidence intervals and/or p-values for high-dimensional precision matrices. In particular, Ren et al. (2015), Janková and van de Geer

(2015, 2017), Xia and Li (2017) provide inference for entries of single precision matrix, which can be used to test $\Omega^m_{(j,k)} = 0$ for $m \in \{\mathrm{I}, \mathrm{II}\}$. On the other hand, Xia et al. (2015), Belilovsky et al. (2016) propose to examine the difference in *values* of two precision matrices. Given that in GGMs $\mathcal{E}^{\mathrm{I}}$ and $\mathcal{E}^{\mathrm{II}}$ are defined based on $\mathbf{\Omega}^{\mathrm{I}}$ and $\mathbf{\Omega}^{\mathrm{II}}$, these methods seem the natural choice for identifying differential connectivity in high-dimensional networks. However, as we illustrate in Section 4.2, they are in fact unable to test $H^*_{0,j} : ne^{\mathrm{I}}_j = ne^{\mathrm{II}}_j$ or $H^*_{0,jk} : d^{\mathrm{I}}_{jk} = d^{\mathrm{II}}_{jk}$, and may lead to misleading conclusions. New procedures are therefore needed to identify differentially connected nodes or edges.

To address the limitations of existing procedures, we propose a new framework, called *differential connectivity analysis* (DCA), for testing $H^*_{0,j}$ and $H^*_{0,jk}$. More specifically, given that in GGMs $(j,k) \in \mathcal{E}^m$ if and only if $\Omega^m_{(j,k)} \neq 0$, we recast the hypotheses $H^*_{0,j}$ and $H^*_{0,jk}$ in (4.3) as the following equivalent hypotheses

$$H_{0,j} : \mathrm{supp}\left(\mathbf{\Omega}^{\mathrm{I}}_j\right) = \mathrm{supp}\left(\mathbf{\Omega}^{\mathrm{II}}_j\right) \quad \text{and} \quad H_{0,jk} : \mathbb{1}\left[\Omega^{\mathrm{I}}_{jk} \neq 0\right] = \mathbb{1}\left[\Omega^{\mathrm{II}}_{jk} \neq 0\right], \tag{4.4}$$

where $\mathbf{\Omega}_j$ denotes the j-th column of matrix $\mathbf{\Omega}$. In other words, our proposal directly examines the *support* of two precision matrices, rather than their *values*. While inference based on *qualitative* hypotheses $H_{0,j}$ and $H_{0,jk}$ is more challenging, these hypotheses avoid false positives due to the changes in the values of inverse covariances, which, as shown in Section 4.2, impede the application of existing *quantitative* inference procedures.

The rest of the chapter is organized as follows. In Section 4.2, we discuss the challenges of obtaining valid inference procedures for $H^*_{0,j}$ and $H^*_{0,jk}$. Our proposed framework is presented in Section 4.3. In Section 4.4, we investigate the use of lasso neighborhood selection in the proposed framework. Simulation studies and real-data examples are presented in Sections 4.5 and 4.6, respectively. We end with a discussion of future research directions in Section 4.7. Technical proofs and additional details are collected in Appendix C.

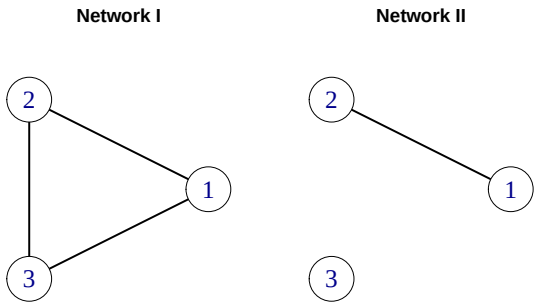
4.2 Challenges of Obtaining Valid Inference for $H^*_{0,j}$ and $H^*_{0,jk}$

In this section, we illustrate why existing hypothesis testing methods based on *values* of precision matrices are unable to test $H^*_{0,j}$ or $H^*_{0,jk}$.

The first class of inference procedures that are based on precision matrix values focuses on a single matrix (Ren et al., 2015, Janková and van de Geer, 2015, 2017, Xia and Li, 2017). These methods examine the null hypothesis $\Omega_{(j,k)} = 0$ for $j \neq k$, and hence could control the probability of falsely detecting an inexistent edge. However, they could not control the false positive rates of $H^*_{0,j}$ or $H^*_{0,jk}$. This is because $H^*_{0,j}$ and $H^*_{0,jk}$ concern the coexistence of edges in two networks. Thus, the false positive rate of $H^*_{0,j}$ and $H^*_{0,jk}$ not only depends on the probability of falsely detecting an inexistent edge, but also depends on the probability of correctly detecting an existent edge. While single network hypothesis testing methods control the former probability, they do not control the latter one, and may thus result in false positives for $H^*_{0,j}$ and $H^*_{0,jk}$.

The second class of inference procedures that are based on values of precision matrices examines the equality of entires of the two matrices. For example, Xia et al. (2015) tests whether $\Omega^{\mathrm{I}}_{(j,k)} = \Omega^{\mathrm{II}}_{(j,k)}$, whereas Belilovsky et al. (2016) tests $\Omega^{\mathrm{I}}_{(j,k)}/\Omega^{\mathrm{I}}_{(j,j)} = \Omega^{\mathrm{II}}_{(j,k)}/\Omega^{\mathrm{II}}_{(j,j)}$; the latter is equivalent to testing $\beta^{\mathrm{I},j}_k = \beta^{\mathrm{II},j}_k$, where for $m \in \{\mathrm{I}, \mathrm{II}\}$, $\beta^{m,j}_k = \Omega^m_{(j,k)}/\Omega^m_{(j,j)}$ are coefficients of the kth variable in the regression of $\boldsymbol{X}_j$ on other variables, which is used, e.g., in neighborhood selection (Meinshausen and Bühlmann, 2006); see (4.6) in Section 4.3.2 for more details. While these procedures examine the equality of entries of two precision matrices, they do not reveal whether the GGM structures are the same. More importantly, the primary limitation of of these methods is that examining differences in magnitudes of $\boldsymbol{\Omega}^{\mathrm{I}}$ and $\boldsymbol{\Omega}^{\mathrm{II}}$ may lead to misleading conclusions. To see this issue, consider the following toy example with three Gaussian variables: Suppose in population I, variable 1 causally affects variables 2 and 3, and variable 2 causally affects variable 3. Suppose, in addition, that in population II the effect of variable 1 on variable 2 remains intact, whereas variables 1 or 2 no longer affects variable 3 due to, e.g., a biological anomaly in variable 3. The undirected

Figure 4.1: Conditional dependency structures of variables in populations I and II.



networks corresponding to the two GGMs are shown in Figure 4.1.

Suppose, without loss of generality, that the precision matrix of variables in population I is

$$\boldsymbol{\Omega}^{\mathrm{I}} = \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}.$$

Further, suppose that $\boldsymbol{X}_1^{\mathrm{II}}$ has the same distribution as $\boldsymbol{X}_1^{\mathrm{I}}$, i.e., $\boldsymbol{X}_1^{\mathrm{II}} =_d \boldsymbol{X}_1^{\mathrm{I}}$. The unchanged (causal) relationship of variables 1 and 2 leads to $\boldsymbol{X}_2^{\mathrm{II}} =_d \boldsymbol{X}_2^{\mathrm{I}}$. On the other hand, $\boldsymbol{X}_3^{\mathrm{II}}$ is independent of $\boldsymbol{X}_1^{\mathrm{II}}$ and $\boldsymbol{X}_2^{\mathrm{II}}$, i.e., $\boldsymbol{X}_3^{\mathrm{II}} \perp\!\!\!\perp \boldsymbol{X}_{\{1,2\}}^{\mathrm{II}}$. Since $\boldsymbol{X}^{\mathrm{I}}$ is normally distributed, we have $\boldsymbol{X}^{\mathrm{I}} \sim_{i.i.d.} \mathcal{N}_3(\mathbf{0}, \boldsymbol{\Sigma}^{\mathrm{I}})$, where

$$\boldsymbol{\Sigma}^{\mathrm{I}} = \left(\boldsymbol{\Omega}^{\mathrm{I}}\right)^{-1} = \begin{bmatrix} 1.5 & -0.5 & -0.5 \\ -0.5 & 1.5 & -0.5 \\ -0.5 & -0.5 & 1.5 \end{bmatrix}.$$

In population II, we have $\boldsymbol{X}_1^{\mathrm{II}} =_d \boldsymbol{X}_1^{\mathrm{I}}$, $\boldsymbol{X}_2^{\mathrm{II}} =_d \boldsymbol{X}_2^{\mathrm{I}}$ and $\boldsymbol{X}_3^{\mathrm{II}} \perp\!\!\!\perp \boldsymbol{X}_{\{1,2\}}^{\mathrm{II}}$. Thus, $\boldsymbol{X}^{\mathrm{II}} \sim_{i.i.d.} \mathcal{N}_3(\mathbf{0}, \boldsymbol{\Sigma}^{\mathrm{II}})$, where

$$\boldsymbol{\Sigma}^{\mathrm{II}} = \begin{bmatrix} 1.5 & -0.5 & 0 \\ -0.5 & 1.5 & 0 \\ 0 & 0 & \mathrm{Var}\left[\boldsymbol{X}_3^{\mathrm{II}}\right] \end{bmatrix},$$

which implies that

$$\boldsymbol{\Omega}^{\mathrm{II}} = \left(\boldsymbol{\Sigma}^{\mathrm{II}}\right)^{-1} = \begin{bmatrix} 0.75 & 0.25 & 0 \\ 0.25 & 0.75 & 0 \\ 0 & 0 & 1/\mathrm{Var}\left[\boldsymbol{X}_3^{\mathrm{II}}\right]. \end{bmatrix}.$$

Assuming, for simplicity, that $\mathrm{Var}[\boldsymbol{X}_3^{\mathrm{II}}] = 1$,

$$\boldsymbol{\Omega}^{\mathrm{II}} = \begin{bmatrix} 0.75 & 0.25 & 0 \\ 0.25 & 0.75 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

In this example, the relationship between variables 1 and 2 is the same in both populations. In particular, the dependence relationship between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ is the same, as indicated in Figure 4.1. However, $\Omega^{\mathrm{I}}_{(1,2)} \neq \Omega^{\mathrm{II}}_{(1,2)}$. Thus, the procedures of Xia et al. (2015) would falsely detect $(1, 2)$ as a differentially connected edge. Moreover this discrepancy is not due to scaling of the entries of the precision matrices. In fact, it can also be seen that $\beta_2^{\mathrm{I},1} \neq \beta_2^{\mathrm{II},1}$ and $\beta_1^{\mathrm{I},2} \neq \beta_1^{\mathrm{II},2}$, where, as before, $\beta_k^{m,j} = \Omega^m_{(j,k)}/\Omega^m_{(j,j)}$. Thus, a similar false conclusion would also be reached using methods based on regression coefficients, such as Belilovsky et al. (2016).

To summarize, existing approaches are unable to test $H^*_{0,j} : ne_j^{\mathrm{I}} = ne_j^{\mathrm{II}}$ and $H^*_{0,jk} : d_{jk}^{\mathrm{I}} = d_{jk}^{\mathrm{II}}$: Estimation approaches do not provide measures of uncertainty; single network hypothesis testing approaches are not guaranteed to control the false positive rate of $H^*_{0,j}$ or $H^*_{0,jk}$, and inference methods that examine the equivalence of two precision matrices could lead to misleading conclusions.

4.3 Differential Connectivity Analysis

4.3.1 Summary of the Proposed Framework

We present in this subsection a high-level summary of the proposed differential connectivity analysis (DCA) framework. Details are presented in the following subsections.

Consider a node $j \in \mathcal{V}$. Then, any other node $k \neq j$ must belong to one of three categories, which are depicted in Figure 4.2a:

i) k is a common neighbor of j in both networks, i.e., $k \in ne_j^{\mathrm{I}} \cap ne_j^{\mathrm{II}} \equiv ne_j^0$;

ii) k is a neighbor of j in one and only one of the two networks, i.e., $k \in ne_j^{\mathrm{I}} \triangle ne_j^{\mathrm{II}}$, where "$\triangle$" is the symmetric difference operator for two sets;

iii) k is not a neighbor of j in either of the two networks, i.e., $k \notin ne_j^{\mathrm{I}} \cup ne_j^{\mathrm{II}}$.

Clearly, $ne_j^{\mathrm{I}} = ne_j^{\mathrm{II}}$ implies $ne_j^{\mathrm{I}} \triangle ne_j^{\mathrm{II}} = \emptyset$. If, to the contrary, there exists a node k such that $k \in ne_j^{\mathrm{I}} \triangle ne_j^{\mathrm{II}}$, then j is differentially connected, and (j, k) is a differentially connected node-pair, i.e., $d_{jk}^{\mathrm{I}} \neq d_{jk}^{\mathrm{II}}$.

Under $H_{0,jk}^* : d_{jk}^{\mathrm{I}} = d_{jk}^{\mathrm{II}}$, node k cannot belong to the second category, and under $H_{0,j}^* : ne_j^{\mathrm{I}} = ne_j^{\mathrm{II}}$, no node should belong to the second category. Thus, to test $H_{0,j}^*$ and $H_{0,jk}^*$, we propose to examine whether there exists any node k such that $k \in ne_j^{\mathrm{I}} \triangle ne_j^{\mathrm{II}} \equiv (ne_j^{\mathrm{I}} \cup ne_j^{\mathrm{II}}) \backslash (ne_j^{\mathrm{I}} \cap ne_j^{\mathrm{II}})$. Specifically, for any $k \neq j$ such that $k \notin ne_j^{\mathrm{I}} \cap ne_j^{\mathrm{II}}$, we check whether $k \in ne_j^{\mathrm{I}} \cup ne_j^{\mathrm{II}}$, i.e., whether k is a neighbor of j in either one of the networks.

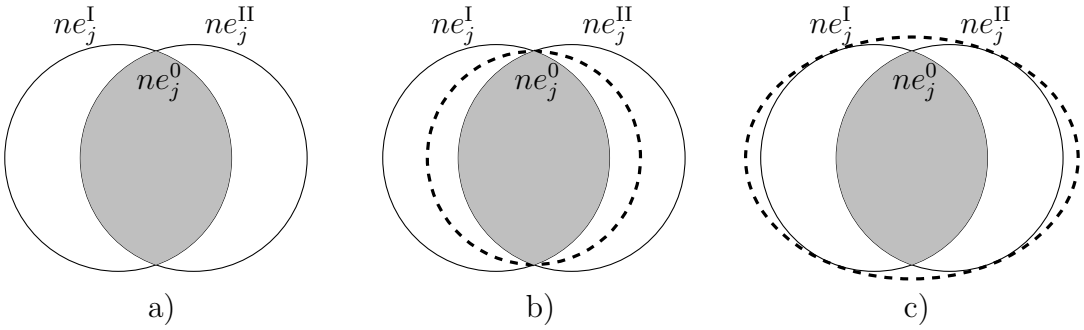
In practice, we do not observe ne_j^0, and need to estimate it. Our inference framework thus consists of two steps:

1. *Estimation*: We estimate the common neighbors of each node j in the two networks, $ne_j^0 \equiv ne_j^{\mathrm{I}} \cap ne_j^{\mathrm{II}}$; this estimate is denoted $\hat{ne}_j^0$.

2. *Inference*: We test whether there exists any $k \notin \hat{ne}_j^0$ such that $k \in ne_j^{\mathrm{I}} \cup ne_j^{\mathrm{II}}$. In Section 4.3.3, we show that for $k \notin \hat{ne}_j^0$ testing $k \notin ne_j^{\mathrm{I}} \cup ne_j^{\mathrm{II}}$ is equivalent to testing $\boldsymbol{X}_k^m \perp\!\!\!\perp \boldsymbol{X}_j^m \mid \boldsymbol{X}_{\backslash\{j,k\}}^m$ for $m \in \{\mathrm{I}, \mathrm{II}\}$. We use properties of GGMs to test this conditional independence.

Details of the above two steps are described in the next two subsections, where it becomes clear that the procedure naturally extends to testing differential connectivity in more than

Figure 4.2: Illustration of the common neighborhood $ne_j^0 = ne_j^{\mathrm{I}} \cap ne_j^{\mathrm{II}}$ of node j in two networks $\mathcal{E}^{\mathrm{I}}$ and $\mathcal{E}^{\mathrm{II}}$: In all figures, ne_j^0 is shaded in gray, and its estimate, $\hat{ne}_j^0$, is shown in dashed ovals; the unshaded parts of ne_j^{I} and ne_j^{II} correspond to $ne_j^{\mathrm{I}} \triangle ne_j^{\mathrm{II}} = \emptyset$. In (b), $\hat{ne}_j^0$ satisfies the *coverage property* of Section 4.3.2 and allows differential connectivity to be estimated; in (c), $\hat{ne}_j^0 \supseteq ne_j^{\mathrm{I}} \cup ne_j^{\mathrm{II}}$ and differential connectivity of j cannot be detected, as illustrated in Section 4.4.1



two networks. From the discussion in the following subsection it will also become clear that the estimated common neighborhood $\hat{ne}_j^0$ plays an important role in the validity and power of the proposed framework. In Section 4.3.2, we show that in order for $\hat{ne}_j^0$ to be useful in the inference step, it needs to satisfy $\Pr[\hat{ne}_j^0 \supseteq ne_j^0] \to 1$ (Figure 4.2b). On the other hand, as the cardinality of $\hat{ne}_j^0$ exceeds that of ne_j^0, the power of the proposed framework deteriorates. In fact, if $\hat{ne}_j^0 \supseteq ne_j^{\mathrm{I}} \cup ne_j^{\mathrm{II}}$ (Figure 4.2c), the differential connectivity of node j cannot be detected. In the following subsection, we also discuss how the randomness in estimating $\hat{ne}_j^0$ may affect the results of the inference step and how valid inferences can be obtained.

4.3.2 Estimating Common Neighbors

Given any $j \in \mathcal{V}$, the first step of the proposed DCA framework involves obtaining an estimate $\hat{ne}_j^0$ of ne_j^0. Our proposed framework does not require $\hat{ne}_j^0$ to be a consistent estimate of ne_j^0, which usually requires stringent conditions. Instead, we observe that under the null hypothesis $H_{0,j} : ne_j^{\mathrm{I}} = ne_j^{\mathrm{II}}$, we have $ne_j^{\mathrm{I}} = ne_j^{\mathrm{II}} = ne_j^0$, which indicates that if $\hat{ne}_j^0 \supseteq ne_j^0$ then $\hat{ne}_j^0 \supseteq ne_j^{\mathrm{I}} \cup ne_j^{\mathrm{II}}$. In other words, if $\hat{ne}_j^0 \supseteq ne_j^0$, then the null hypothesis $H_{0,j} : ne_j^{\mathrm{I}} = ne_j^{\mathrm{II}}$

can be examined by testing whether there exists any $k \notin \hat{ne}_j^0$ such that $k \in ne_j^{\mathrm{I}} \cup ne_j^{\mathrm{II}}$. In that case, we also know that (k, j) is a differentially connected edge. Based on the above observation, in the proposed framework we require that

$$\Pr[\mathcal{C}] \equiv \Pr[\hat{ne}_j^0 \supseteq ne_j^0] \to 1. \tag{4.5}$$

We call (4.5) the *coverage property* of estimated common neighbors (see Figure 4.2b).

To estimate the common neighborhood ne_j^0, we use the fact that if $\boldsymbol{X}^m, m \in \{\mathrm{I}, \mathrm{II}\}$ is Gaussian, then

$$\boldsymbol{X}_j^m = \boldsymbol{X}_{\setminus j}^m \boldsymbol{\beta}^{m,j} + \boldsymbol{\epsilon}^{m,j}, \tag{4.6}$$

where $\boldsymbol{\beta}^{m,j}$ is a $(p-1)$-vector of coefficients and $\boldsymbol{\epsilon}^{m,j}$ is an n^m-vector of random Gaussian errors. Moreover, by multivariate Gaussian theory, $\beta_k^{m,j} = \Omega_{(j,k)}^m / \Omega_{(j,j)}^m$. Thus, $\beta_k^{m,j} \neq 0$ if and only if $\Omega_{(j,k)}^m \neq 0$, which, as discussed before, is equivalent to $k \in ne_j^m$. Therefore, the common neighbors of node j in the two populations is

$$ne_j^0 \equiv ne_j^{\mathrm{I}} \cap ne_j^{\mathrm{II}} = \left\{ k : \beta_k^{\mathrm{I},j} \neq 0 \;\&\; \beta_k^{\mathrm{II},j} \neq 0 \right\}. \tag{4.7}$$

Based on (4.7), an estimate of ne_j^0 may be obtained from the estimated supports of $\boldsymbol{\beta}^{\mathrm{I},j}$ and $\boldsymbol{\beta}^{\mathrm{II},j}$. Various procedures can be used to estimate $\boldsymbol{\beta}^{\mathrm{I},j}$ and $\boldsymbol{\beta}^{\mathrm{II},j}$ and, in turn, ne_j^0. We present a lasso-based estimator as an example in Section 4.4. Proposition 4.4.1 shows that under appropriate conditions, the lasso-based estimate satisfies the coverage property (4.5), and is thus valid for the estimation step of the proposed framework.

4.3.3 Testing Differential Connectivity

Recall, from our discussion in the previous section, that the estimated joint neighborhood $\hat{ne}_j^0$ needs to satisfy the coverage property $\Pr[\mathcal{C}] \to 1$, where $\mathcal{C}$ is defined in (4.5). Moreover, under the event $\mathcal{C}$, if there exists any $k \notin \hat{ne}_j^0$ such that $k \in ne_j^{\mathrm{I}} \cup ne_j^{\mathrm{II}}$, then with probability tending to one, $ne_j^{\mathrm{I}} \neq ne_j^{\mathrm{II}}$. In GGMs, $k \in ne_j^{\mathrm{I}} \cup ne_j^{\mathrm{II}}$ if and only if $\boldsymbol{X}_k^{\mathrm{I}} \not\perp \boldsymbol{X}_j^{\mathrm{I}} \,|\, \boldsymbol{X}_{\setminus\{k,j\}}^{\mathrm{I}}$

or $\boldsymbol{X}_k^{\mathrm{II}} \not\perp\!\!\!\perp \boldsymbol{X}_j^{\mathrm{II}} \,|\, \boldsymbol{X}_{\setminus\{k,j\}}^{\mathrm{II}}$. Thus, to determine whether there exists any $k \notin \hat{ne}_j^0$ such that $k \in ne_j^{\mathrm{I}} \cup ne_j^{\mathrm{II}}$, we test the following hypotheses

$$H_{0,j}^m : \boldsymbol{X}_k^m \perp\!\!\!\perp \boldsymbol{X}_j^m \,\big|\, \boldsymbol{X}_{\setminus\{k,j\}}^m, \forall k \notin \hat{ne}_j^0 \cup \{j\}. \tag{4.8}$$

for $m \in \{\mathrm{I}, \mathrm{II}\}$. As mentioned in Section 4.3.2, these hypotheses are equivalent to $H_{0,j}^m : \boldsymbol{\beta}_{\setminus \hat{ne}_j^0}^{m,j} = \mathbf{0}$ for $m \in \{\mathrm{I}, \mathrm{II}\}$ based on the linear representations in (4.6). Because $H_{0,j}^* : ne_j^{\mathrm{I}} = ne_j^{\mathrm{II}}$ is rejected if $H_{0,j}^{\mathrm{I}}$ or $H_{0,j}^{\mathrm{II}}$ is rejected, according to the Šidák correction, to control type-I error rate of $H_{0,j}$ at level $\alpha > 0$, we control type-I error rate of $H_{0,j}^{\mathrm{I}}$ and of $H_{0,j}^{\mathrm{II}}$ at level $1 - \sqrt{1-\alpha}$. Note that if $\hat{ne}_j^0 \cup \{j\} = \mathcal{V}$, we do not reject $H_{0,j}$.

We can use a similar strategy to test $H_{0,jk}^* : d_{jk}^{\mathrm{I}} = d_{jk}^{\mathrm{II}}$ for $k \neq j$, where the dyad d_{jk} for a node-pair (j,k) is defined in (4.2). Here, we test whether $k \in ne_j^{\mathrm{I}} \triangle ne_j^{\mathrm{II}} = (ne_j^{\mathrm{I}} \cup ne_j^{\mathrm{II}}) \setminus ne_j^0$. If $k \in \hat{ne}_j^0$, we estimate k to be a neighbor of j in both networks, and do not reject the null hypothesis $H_{0,jk}$. Otherwise, if $k \notin \hat{ne}_j^0$, we use a procedure similar to that used for testing $H_{0,j}^* : ne_j^{\mathrm{I}} = ne_j^{\mathrm{II}}$. Specifically, to test $H_{0,jk}^* : d_{jk}^{\mathrm{I}} = d_{jk}^{\mathrm{II}}$, we test $H_{0,jk}^m : \boldsymbol{X}_k^m \perp\!\!\!\perp \boldsymbol{X}_j^m \,|\, \boldsymbol{X}_{\setminus\{k,j\}}^m$ for $m \in \{\mathrm{I}, \mathrm{II}\}$. As before, $H_{0,jk}$ is rejected at level α if $H_{0,jk}^{\mathrm{I}}$ or $H_{0,jk}^{\mathrm{II}}$ is rejected at level $1 - \sqrt{1-\alpha}$. By (4.6), these hypotheses are equivalent to $H_{0,jk}^m : \beta_k^{m,j} = 0$ for $m \in \{\mathrm{I}, \mathrm{II}\}$.

Even when $\hat{ne}_j^0$ satisfies the coverage property, the hypotheses $H_{0,j}^{\mathrm{I}}$ and $H_{0,j}^{\mathrm{II}}$ depend on the data through $\hat{ne}_j^0$, which is a random quantity. This dependence complicates hypothesis testing: we look at the same data twice, once to formulate hypotheses, and once to test the hypotheses that are formulated using the same data. Conventional statistical wisdom indicates that with such double-peeking, common hypothesis testing methods are no longer valid (see, e.g., Pötscher, 1991, Kabaila, 1998, Leeb and Pötscher, 2003, 2005, 2006a,b, 2008, Kabaila, 2009, Berk et al., 2013). One strategy to avoid double-peeking is to use sample-splitting (see, e.g., Wasserman and Roeder, 2009, Meinshausen et al., 2009), wherein the data are divided in two parts; the first part is used to estimate $\hat{ne}_j^0$ and the second to test $H_{0,j}^*$.

Sample-splitting breaks down the dependence between the data that are used to generate hypotheses and the data that are used to test the hypotheses. However, sample-splitting has

two drawback: first, using only one half of the samples to test $H^*_{0,j}$ diminishes the power; second, we also only use half of the samples to estimate $\hat{ne}^0_j$, which makes the coverage property of common neighbors (4.5) less likely to happen with finite samples. Because our procedure requires the coverage property in order to control the type-I error, in practice, sample-splitting may inflate the false positive rate. Proposition 4.4.2 in Section 4.4 offers an alternative to sample-splitting. This result shows that although $\hat{ne}^0_j$ is in general random, under appropriate conditions, the lasso-based estimate of $\hat{ne}^0_j$ discussed in Section 4.4 converges in probability to a *deterministic* set, which is not affected by the randomness of the data. Thus, asymptotically, we can treat $\hat{ne}^0_j$ as deterministic, and hence treat $H^{\mathrm{I}}_{0,j}$ and $H^{\mathrm{II}}_{0,j}$ as classical non-data-dependent hypotheses. This new result simplifies the implementation of the proposed two-step inference procedure. Our empirical analyses in Section 4.5 show that the lasso-based inference also offers advantages over its sample-splitting counterpart.

The equivalence between the qualitative tests $H^*_{0,j} : ne^{\mathrm{I}}_j = ne^{\mathrm{II}}_j$ and those based on conditional independence, $H^{\mathrm{I}}_{0,j} : \boldsymbol{\beta}^{\mathrm{I},j}_{\setminus \hat{ne}^0_j} = \mathbf{0}$ and $H^{\mathrm{II}}_{0,j} : \boldsymbol{\beta}^{\mathrm{II},j}_{\setminus \hat{ne}^0_j} = \mathbf{0}$, offers a concrete test for differential connectivity. $H^{\mathrm{I}}_{0,j}$ and $H^{\mathrm{II}}_{0,j}$ can be tested by using recent proposals for testing coefficients in high-dimensional linear regressions (e.g., Javanmard and Montanari, 2014a, Zhang and Zhang, 2014, van de Geer et al., 2014, Zhao and Shojaie, 2016, Ning and Liu, 2017). Because the above procedures examine individual regression coefficients, they are also appropriate for testing $H^{\mathrm{I}}_{0,jk} : \beta^{\mathrm{I},j}_k = 0$ and $H^{\mathrm{II}}_{0,jk} : \beta^{\mathrm{II},j}_k = 0$. To control type-I error rates of $H^{\mathrm{I}}_{0,j}$ and $H^{\mathrm{II}}_{0,j}$, we need to control the family-wise error rate (FWER) on individual regression coefficients using, e.g., the Holm procedure. Alternatively, $H^{\mathrm{I}}_{0,j}$ and $H^{\mathrm{II}}_{0,j}$ can be tested using group hypothesis testing procedures that examine a group of regression coefficients, such as the least-squares kernel machines (LSKM) test (Liu et al., 2007). Although such group hypothesis testing approaches cannot be used to examine $H^*_{0,jk}$, they often result in computational and power advantages in testing $H^*_{0,j}$, compared to hypothesis testing approaches that examine individual regression coefficients.

To summarize, our proposed framework consists of an estimation step followed by an inference step. Each of these steps can incorporate different methods. For the estimation

step, we require that for each $j \in \mathcal{V}$, the estimated common neighborhood, $\hat{ne}_j^0$, satisfies the coverage property (4.5), i.e., $\Pr[\hat{ne}_j^0 \supseteq ne_j^0] \to 1$. Moreover, we require that either $\hat{ne}_j^0$ is deterministic with high probability, or that the dependence between the estimation and inference steps is broken down by using a sample-splitting approach. For the inference step, valid high-dimensional hypothesis testing methods that examine individual regression coefficients are suitable for testing both $H_{0,jk}^* : d_{jk}^{\mathrm{I}} = d_{jk}^{\mathrm{II}}$ and $H_{0,j}^* : ne_j^{\mathrm{I}} = ne_j^{\mathrm{II}}$. The latter hypothesis, $H_{0,j}^*$, can also be examined using group hypothesis testing approaches. We thus arrive at to the following result.

Theorem 4.3.1. *Suppose the procedure used in the estimation step of differential connectivity analysis (DCA) satisfies the following conditions for each $j \in \mathcal{V}$:*

1. *The estimated common neighborhood of node j in the estimation step, $\hat{ne}_j^0$, satisfies the coverage property, i.e., $\lim_{n\to\infty} \Pr[\hat{ne}_j^0 \supseteq ne_j^0] = 1$;*

2. *Either the estimated common neighborhood $\hat{ne}_j^0$ is deterministic with probability tending to one, or the data used to test hypotheses $H_{0,j}^m$ ($H_{0,jk}^m$) for $m \in \{\mathrm{I}, \mathrm{II}\}$ are independent of the data used to estimate $\hat{ne}_j^0$.*

Then if, for $m \in \{\mathrm{I}, \mathrm{II}\}$, the inference procedure for testing $H_{0,j}^m : \boldsymbol{\beta}_{\setminus \hat{ne}_j^0}^{m,j} = \mathbf{0}$ ($H_{0,jk}^m : \beta_k^{m,j} = 0$) is asymptotically valid, DCA asymptotically controls the type-I error rate of $H_{0,j}^* : ne_j^{\mathrm{I}} = ne_j^{\mathrm{II}}$ ($H_{0,jk}^* : d_{jk}^{\mathrm{I}} = d_{jk}^{\mathrm{II}}$).

4.4 The Validity of Lasso for DCA

A convenient procedure for estimating ne_j^0 is the *lasso neighborhood selection* (Meinshausen and Bühlmann, 2006). Specifically, we define $\hat{ne}_j^0 = \hat{ne}_j^{\mathrm{I}} \cap \hat{ne}_j^{\mathrm{II}} = \mathrm{supp}(\hat{\boldsymbol{\beta}}^{\mathrm{I},j}) \cap \mathrm{supp}(\hat{\boldsymbol{\beta}}^{\mathrm{II},j})$, where for $m \in \{\mathrm{I}, \mathrm{II}\}$,

$$\hat{\boldsymbol{\beta}}^{m,j} \equiv \underset{\boldsymbol{b}\in\mathbb{R}^{p-1}}{\arg\min} \left\{ \frac{1}{2n} \left\| \boldsymbol{X}_j^m - \boldsymbol{X}_{\setminus j}^m \boldsymbol{b} \right\|_2^2 + \lambda_j^m \left\| \boldsymbol{b} \right\|_1 \right\}. \tag{4.9}$$

In this section, we present two propositions regarding lasso neighborhood selection, which show that under appropriate conditions, the estimate $\hat{ne}_j^0$ satisfies the coverage property $\Pr[\hat{ne}_j^0 \supseteq ne_j^0] \to 1$, and is deterministic with high probability. These results imply that lasso neighborhood selection satisfies the requirements of Theorem 4.3.1. Moreover, no sample-splitting is necessary to draw valid inference from lasso-based estimates. We end this section with a brief discussion of the power of DCA when lasso is used in the estimation step.

To establish that lasso-based estimates of neighborhoods are deterministic with high probability, in Proposition 4.4.2 we establish a relationship between the lasso neighborhood selection estimator (4.9) and its noiseless (and hence deterministic) counterpart

$$\check{\boldsymbol{\beta}}^{m,j} \equiv \underset{\boldsymbol{b}\in\mathbb{R}^{p-1}}{\arg\min} \left\{ \mathrm{E}\left[\frac{1}{2n} \left\| \boldsymbol{X}_j^m - \boldsymbol{X}_{\setminus j}^m \boldsymbol{b} \right\|_2^2 \right] + \lambda_j^m \left\| \boldsymbol{b} \right\|_1 \right\}. \tag{4.10}$$

To the best of our knowledge, the fact that the lasso support is asymptotically deterministic has not been previously established for regressions with random design $\boldsymbol{X}$.

The following are sufficient conditions for our propositions.

(**A1**) For $m \in \{\mathrm{I}, \mathrm{II}\}$, rows of the data $\boldsymbol{X}^m$ are independent and identically distributed Gaussian random vectors: $\boldsymbol{X}^m \sim_{i.i.d.} \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma}^m)$, where, without loss of generality, we assume $\mathrm{diag}(\boldsymbol{\Sigma}^m) = \boldsymbol{1}$. Further, the minimum and maximum eigenvalues of $\boldsymbol{\Sigma}^m$ satisfy

$$\liminf_{n\to\infty} \phi^2_{\min}\left[\boldsymbol{\Sigma}^m\right] > 0 \quad \text{and} \quad \limsup_{n\to\infty} \phi^2_{\max}\left[\boldsymbol{\Sigma}^m\right] < \infty.$$

(**A2**) For $m \in \{\mathrm{I}, \mathrm{II}\}$ and a given variable $j \in \mathcal{V}$, the sample size n^m, dimension p, lasso neighborhood selection tuning parameters λ_j^m, number of neighbors $q_j^m \equiv |ne_j^m|$, and minimum non-zero coefficients $b_{\min}^{m,j} \equiv \min\{|\beta_k^{m,j}| : \beta_k^{m,j} \neq 0\}$, where $\boldsymbol{\beta}^{m,j}$is defined in (4.6), satisfy

$$\lambda_j^m q_j^m \to l^m < \infty \quad , \quad \sqrt{\frac{\log(p)}{n^m}} \frac{1}{\lambda_j^m} \to 0 \quad \text{and} \quad \frac{\lambda_j^m \sqrt{q_j^m}}{b_{\min}^{m,j}} \to 0.$$

(**A3**) For $m \in \{\mathrm{I}, \mathrm{II}\}$ and a given variable $j \in \mathcal{V}$, define the sub-gradient $\check{\boldsymbol{\tau}}^{m,j}$ based on the stationary condition (Karush, 1939, Kuhn and Tucker, 1951) of (4.10)

$$\check{\boldsymbol{\tau}}^{m,j} = \frac{1}{n\lambda_j^m} \mathrm{E}\left[\boldsymbol{X}_{\setminus j}^{m\top} \left(\boldsymbol{X}_j^m - \boldsymbol{X}_{\setminus j}^m \check{\boldsymbol{\beta}}^{m,j} \right) \right]. \tag{4.11}$$

We assume $\check{\boldsymbol{\tau}}^{m,j}$ satisfies $\limsup_{n\to\infty} \left\| \check{\boldsymbol{\tau}}^{m,j}_{\setminus \check{ne}^m_j} \right\|_\infty \leq 1 - \delta^m$ such as

$$\sqrt{\frac{\log(p)}{n^m}} \frac{1}{\lambda^m_j \delta^m} \to 0,$$

and

$$\frac{1}{\lambda^m_j} \sqrt{\frac{\log(p)}{n^m}} \left(\min_{k \in \check{ne}^m_j \setminus ne^m_j} \left| \left[\boldsymbol{\Sigma}_{(\check{ne}^m_j, \check{ne}^m_j)} \right]^{-1} \check{\boldsymbol{\tau}}^{m,j}_{\check{ne}^m_j} \right|_k \right)^{-1} \to 0.$$

Condition (**A1**) characterizes the data distribution. Combining the first two requirements of (**A2**), for $m \in \{\mathrm{I}, \mathrm{II}\}$, we get $q^m_j = o(\sqrt{n^m / \log(p)})$, as commonly required in the lasso literature. The third constraint in (**A2**) is the $\boldsymbol{\beta}$-min condition, which prevents the signal from being too weak to be detected; this condition may be relaxed to allow the presence of some weak signal variables. In addition, (**A2**) requires the tuning parameters λ^m_j to approach zero at a slower rate than $\sqrt{\log(p)/n^m}$, which is the minimum tuning parameter rate for prediction consistency of lasso with Gaussian data (see, e.g., Bickel et al., 2009). Since $\sqrt{\log(p)/n^m}/\lambda^m_j \to 0$ by (**A2**), condition (**A3**) requires that the tuning parameter λ does not converge to any transition points too fast, where some entries of $\check{\boldsymbol{\beta}}^{m,j}$ change from zero to nonzero, or vice versa. (**A3**) also requires that $\min_{k \in \check{ne}^m_j \setminus ne^m_j} |[\boldsymbol{\Sigma}_{(\check{ne}^m_j, \check{ne}^m_j)}]^{-1} \check{\boldsymbol{\tau}}^{m,j}_{\check{ne}^m_j}|_k$ does not converge to zero too fast.

We now present Propositions 4.4.1 and 4.4.2. As mentioned in Section 4.3.1, Proposition 4.4.1 implies that lasso neighborhood selection is a valid method for estimation in our framework, and Proposition 4.4.2 relieves us from using sample-splitting to circumvent double-peeking by our procedure.

Proposition 4.4.1. *Suppose conditions (**A1**) and (**A2**) for variable $j \in \mathcal{V}$ hold. Then $\hat{ne}^0_j$ estimated using lasso neighborhood selection satisfies*

$$\lim_{n\to\infty} \Pr\left[\hat{ne}^0_j \supseteq ne^0_j\right] = 1. \tag{4.12}$$

Proposition 4.4.2. *Suppose conditions (**A1**) – (**A3**) for variable $j \in \mathcal{V}$ hold. Then $\hat{ne}_j^0$ estimated using lasso neighborhood selection satisfies*

$$\lim_{n\to\infty} \Pr\left[\hat{ne}_j^0 = \check{ne}_j^0\right] = 1, \tag{4.13}$$

where $\check{ne}_j^0 \equiv \text{supp}(\check{\boldsymbol{\beta}}^{\text{I},j}) \cap \text{supp}(\check{\boldsymbol{\beta}}^{\text{II},j})$, and $\check{\boldsymbol{\beta}}^{\text{I},j}$ and $\check{\boldsymbol{\beta}}^{\text{II},j}$ are defined in (4.10).

The proofs of Propositions 4.4.1 and 4.4.2 are presented in Appendices C.1 and C.2, respectively. The result in Proposition 4.4.2 should not be confused with the variable selection consistency of lasso, proved in, e.g., Meinshausen and Bühlmann (2006). In particular, Meinshausen and Bühlmann (2006) proved that under the stringent neighborhood stability condition, the selected neighborhoods converge in probability to the true neighborhoods, i.e., $\lim_{n\to\infty} \Pr[\hat{ne}_j^0 = ne_j^0] = 1$. On the other hand, under considerably milder conditions, Proposition 4.4.2 indicates that although the selected neighborhoods may not converge to the true neighborhoods, they still converge to deterministic sets, $\check{ne}_j^0$. This result allows us to ignore the randomness in $\hat{ne}_j^0$, asymptotically.

4.4.1 Power of DCA when using lasso in the estimation step

In Section 4.3.2, we argued that the estimated common neighborhood $\hat{ne}_j^0$ needs to satisfy the coverage property, i.e., $\Pr[\hat{ne}_j^0 \supseteq ne_j^0] \to 1$. In this section, we briefly discuss how the cardinality of $\hat{ne}_j^0$ affects the power of the proposed differential connectivity test. We also discuss a sufficient condition, where, using lasso in the estimation step, the power of DCA could approach one asymptotically for detecting differential connectivity.

As mentioned in Section 4.3.3, to examine $H_{0,j}^* : ne_j^{\text{I}} \neq ne_j^{\text{II}}$ and $H_{0,jk}^* : d_{jk}^{\text{I}} \neq d_{jk}^{\text{I}}$, in the second step of the proposed procedure, we test whether variable j is conditionally independent of variables that are not in the estimated common neighborhood. In the case where $ne_j^{\text{I}} \neq ne_j^{\text{II}}$, if the estimated neighborhood of variable j is too large, such that $\hat{ne}_j^0 \supseteq ne_j^{\text{I}} \cup ne_j^{\text{II}}$ (see Figure 4.2c), then for any $k \notin \hat{ne}_j^0 \cup \{j\}$, $\boldsymbol{X}_j^m \perp\!\!\!\perp \boldsymbol{X}_k^m \mid \boldsymbol{X}_{\setminus\{j,k\}}^m$ for $m \in \{\text{I}, \text{II}\}$. In this case, we will not be able to identify differential connectivity of node j. Thus, even

though the validity of DCA requires that $\hat{ne}_j^0$ achieves the coverage property, $\hat{ne}_j^0$ should not be exceedingly larger than ne_j^0.

To examine the power of DCA with the lasso-based estimate of $\hat{ne}_j^0$, suppose $|ne_j^{\mathrm{II}}| = o(|ne_j^{\mathrm{I}}|)$. Belloni and Chernozhukov (2013) show that, under mild conditions, $|ne_j^m| \asymp_p |\hat{ne}_j^m| = |\mathrm{supp}(\hat{\boldsymbol{\beta}}^{m,j})|$, i.e., with high probability, the size of the estimated neighborhood of j is of the same order as the size of its true neighborhood. Therefore, $|\hat{ne}_j^0| \leq |\hat{ne}_j^{\mathrm{II}}| \asymp_p |ne_j^{\mathrm{II}}| = o(|ne_j^{\mathrm{I}}|)$, i.e., with high probability, $|ne_j^{\mathrm{I}}| \gg |\hat{ne}_j^0|$ so that $\hat{ne}_j^0 \not\supseteq ne_j^{\mathrm{I}}$. Similarly, if $|ne_j^{\mathrm{I}}| = o(|ne_j^{\mathrm{II}}|)$, then with high probability $\hat{ne}_j^0 \not\supseteq ne_j^{\mathrm{II}}$. Thus, if $|ne_j^{\mathrm{I}}|$ and $|ne_j^{\mathrm{II}}|$ are not of the same order, then, with high probability, $\hat{ne}_j^0 \not\supseteq ne_j^{\mathrm{I}} \triangle ne_j^{\mathrm{II}}$, and there exists $k \notin \hat{ne}_j^0 \cup \{j\}$ such that $\boldsymbol{X}_j^m \not\perp \boldsymbol{X}_k^m \,|\, \boldsymbol{X}_{\setminus\{j,k\}}^m$ for $m \in \{\mathrm{I}, \mathrm{II}\}$. In this case, with any conditional testing method that achieves asymptotic power one, DCA is asymptotically guaranteed to detect the differential connectivity of j.

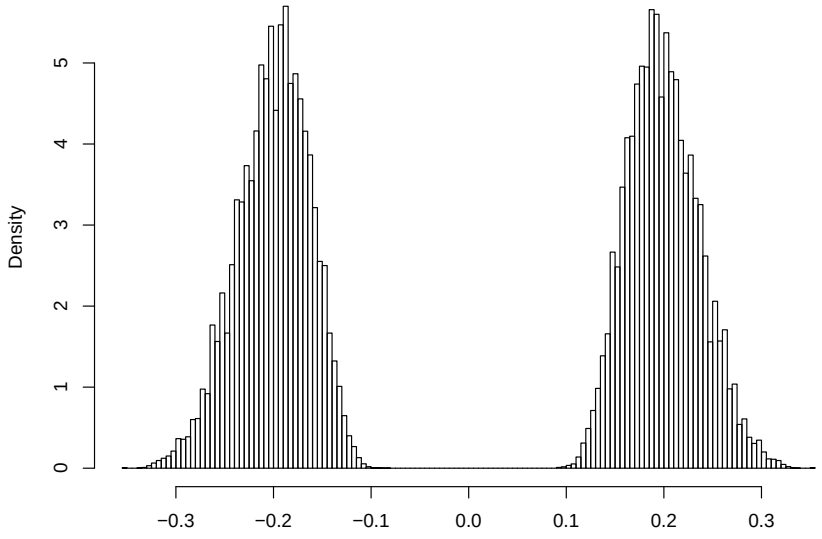
While the conditions presented in the above special case are sufficient and not necessary, the scenario sheds light on the power properties of DCA. We defer to future research a more thorough assessment of power properties of DCA.

4.5 Simulation Studies

In this section, we present results of a simulation study that evaluates the power and type-I error rate of the DCA framework using various choices of procedures in the estimation and inference steps.

In this simulation, we generate $\mathcal{E}^{\mathrm{I}}$ from a power-law degree distribution with power parameter 5, $|\mathcal{V}| \equiv p = 200$ and $|\mathcal{E}^{\mathrm{I}}| = p(p-1)/100$; this corresponds to an edge density of 0.02 in graph G^{I}. Power-law degree distributions are able to produce graphs with hubs, which are expected in real-world networks (Newman, 2003). To simulate $\mathcal{E}^{\mathrm{II}}$, among the 100 most connected nodes in G^{I}, we randomly select 20 nodes, remove all the edges that are connected to them, and then randomly add edges to graph G^{II} so that $|\mathcal{E}^{\mathrm{II}}| = |\mathcal{E}^{\mathrm{I}}|$. To simulate $\boldsymbol{\Omega}^{\mathrm{I}}$, for

Figure 4.3: Distribution of non-zero partial correlations in simulated $\mathbf{\Omega}^{\mathrm{I}}$ and $\mathbf{\Omega}^{\mathrm{II}}$.



$j \neq k$, we let

$$\Omega^{\mathrm{I}}_{(j,k)} = \begin{cases} 0 & (j,k) \notin \mathcal{E}^{\mathrm{I}} \\ 0.5 & (j,k) \in \mathcal{E}^{\mathrm{I}}, \text{ with 50\% probability} \\ -0.5 & (j,k) \in \mathcal{E}^{\mathrm{I}}, \text{ with 50\% probability} \end{cases}.$$

To simulate $\mathbf{\Omega}^{\mathrm{II}}$, for $j \neq k$, we let

$$\Omega^{\mathrm{II}}_{(j,k)} = \begin{cases} \Omega^{\mathrm{I}}_{(j,k)} & (j,k) \in \mathcal{E}^{\mathrm{I}} \cap \mathcal{E}^{\mathrm{II}} \\ 0 & (j,k) \notin \mathcal{E}^{\mathrm{II}} \\ 0.5 & (j,k) \in \mathcal{E}^{\mathrm{II}} \backslash \mathcal{E}^{\mathrm{I}}, \text{ with 50\% probability} \\ -0.5 & (j,k) \in \mathcal{E}^{\mathrm{II}} \backslash \mathcal{E}^{\mathrm{I}}, \text{ with 50\% probability} \end{cases}.$$

Finally, for $m \in \{\mathrm{I}, \mathrm{II}\}$, we let $\Omega^m_{(j,j)} = \sum_{k \neq j} |\Omega^m_{(j,k)}| + u^m$ for $j = 1, \ldots, p$, where u^m is chosen such that $\phi^2_{\min}[\mathbf{\Omega}^m] = 0.1$, where $\phi^2_{\min}[\mathbf{\Omega}^m]$ is the smallest eigenvalues of $\mathbf{\Omega}^m$. Figure 4.3 shows the distribution of non-zero partial correlations in $\mathbf{\Omega}^{\mathrm{I}}$ and $\mathbf{\Omega}^{\mathrm{II}}$. From $\mathbf{\Omega}^{\mathrm{I}}$ and $\mathbf{\Omega}^{\mathrm{II}}$, we generate $\boldsymbol{X}^{\mathrm{I}} \sim_{i.i.d.} \mathcal{N}_p(\mathbf{0}, (\mathbf{\Omega}^{\mathrm{I}})^{-1})$ and $\boldsymbol{X}^{\mathrm{II}} \sim_{i.i.d.} \mathcal{N}_p(\mathbf{0}, (\mathbf{\Omega}^{\mathrm{II}})^{-1})$, where $n^{\mathrm{I}} = n^{\mathrm{II}} = n \in \{100, 200, 400, 800\}$.

To estimate common neighbors of each node $j \in \mathcal{V}$, we use lasso neighborhood selection, with tuning parameters chosen by 10-fold cross-validation (CV). We either use sample-splitting to address the issue of double-peeking, with half of samples used to estimate $\hat{ne}_j^0$ and the other half to test $H_{0,j}$, or use a naïve approach, where the whole dataset is used to estimate $\hat{ne}_j^0$ and to test $H_{0,j}^* : ne_j^{\mathrm{I}} = ne_j^{\mathrm{II}}$; the latter approach is justified by Proposition 4.4.2. To examine $H_{0,j}$ for each $j = 1, \ldots, p$, we consider LSKM (Liu et al., 2007) and the GraceI test (Zhao and Shojaie, 2016), which are implemented in the `SKAT` and `Grace` R packages, respectively. As a result, we compare in total 4 approaches: {naïve lasso neighborhood selection, sample-splitting lasso neighborhood selection}×{LSKM, GraceI}.

Figure 4.4 shows average type-I error rates of falsely rejecting $H_{0,j}^* : ne_j^{\mathrm{I}} = ne_j^{\mathrm{II}}$, as well as average powers of various DCA variants based on $R = 100$ repetitions. Denote by $z_{j,r}$ the decision function based on the GraceI test or LSKM: $z_j^r = 1$ if we reject hypothesis $H_{0,j}^* : ne_j^{\mathrm{I}} = ne_j^{\mathrm{II}}$ in the r-th repetition, and $z_j^r = 0$ otherwise. The average type-I error rate is defined as

$$\mathrm{T1ER} = \left(\sum_{r=1}^{R} \left| \left\{ j \in \mathcal{V} : ne_j^{\mathrm{I},r} = ne_j^{\mathrm{II},r} \right\} \right| \right)^{-1} \sum_{r=1}^{R} \left\{ \sum_{j \in \mathcal{V} : ne_j^{\mathrm{I},r} = ne_j^{\mathrm{II},r}} z_j^r \right\}, \tag{4.14}$$

i.e., the proportion of null hypotheses in R repetitions that we falsely reject $H_{0,j}^*$. For $k \in \{1, 3, 5, 10\}$, the average power of rejecting $H_{0,j}^* : ne_j^{\mathrm{I}} = ne_j^{\mathrm{II}}$ when ne_j^{I} and ne_j^{II} differ by at least k members is defined as

$$\mathrm{Pk} = \left(\sum_{r=1}^{R} \left| \left\{ j \in \mathcal{V} : \left| ne_j^{\mathrm{I},r} \triangle ne_j^{\mathrm{II},r} \right| \geq k \right\} \right| \right)^{-1} \sum_{r=1}^{R} \left\{ \sum_{j \in \mathcal{V} : \left| ne_j^{\mathrm{I},r} \triangle ne_j^{\mathrm{II},r} \right| \geq k} z_j^r \right\}, \tag{4.15}$$

where "$\triangle$" is the symmetric difference operator of two sets.

The simulation reveals several interesting patterns. First, naïve procedures which use the same dataset to estimate $\hat{ne}_j^0$ and test $H_{0,j}^*$ tend to have better statistical power than their sample-splitting counterparts. This is understandable, as sample-splitting only uses half of the data for hypothesis testing. On the other hand, naïve procedures also have better control of type-I error rate than sample-splitting procedures. This is because the event $\hat{ne}_j^0 \supseteq ne_j^0$,

Figure 4.4: The average type-I error rate of falsely rejecting $H^*_{0,j} : ne^{\mathrm{I}}_j = ne^{\mathrm{II}}_j$, as well as the average powers of rejecting $H^*_{0,j} : ne^{\mathrm{I}}_j = ne^{\mathrm{II}}_j$ when ne^{I}_j and ne^{II}_j differ in at least one (P1), three (P3), five (P5) or ten (P10) elements. The axis for type-I error rate is on the left of each panel, whereas the axis for power is on the right.

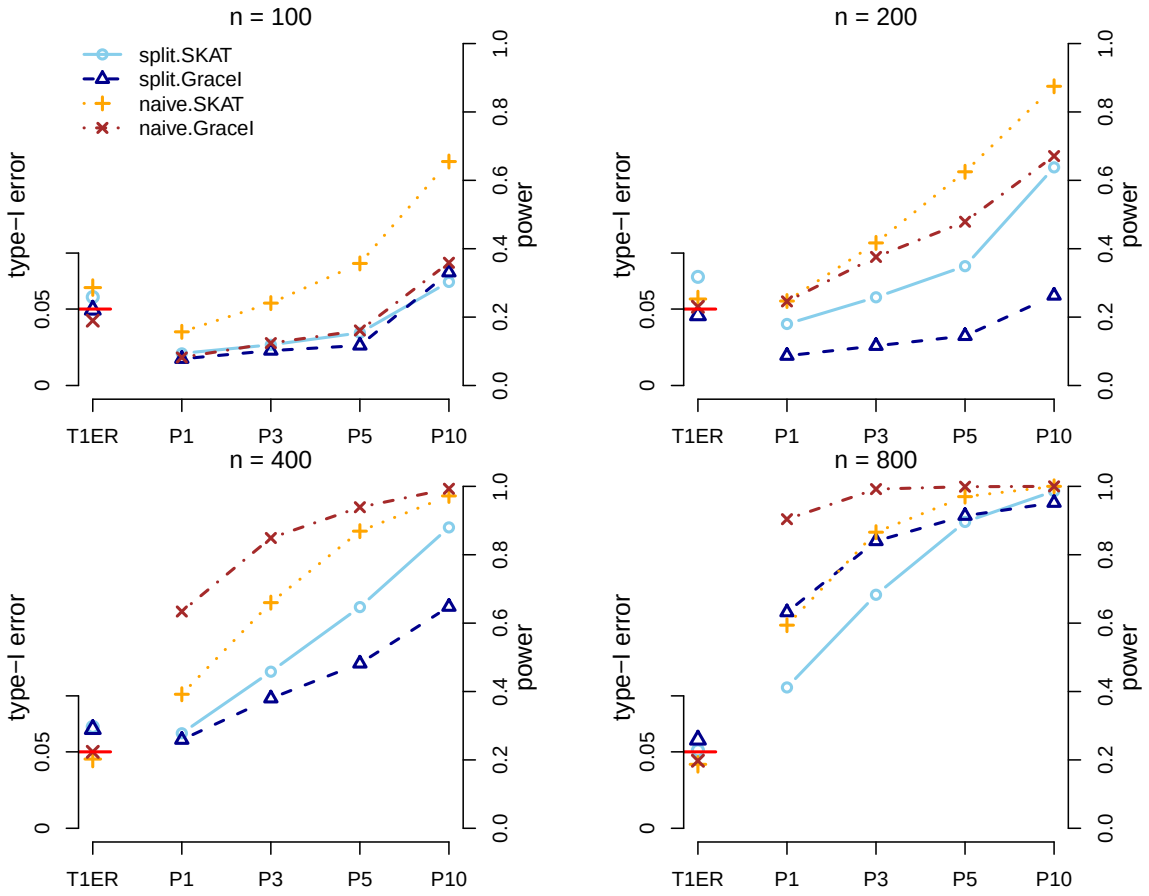


Table 4.1: The average type-I error rate (T1ER) of falsely rejecting $H_{0,j} : ne_j^{\mathrm{I}} = ne_j^{\mathrm{II}}$, as well as the average powers of rejecting $H_{0,j} : ne_j^{\mathrm{I}} = ne_j^{\mathrm{II}}$ when ne_j^{I} and ne_j^{II} differ in at least one (P1), three (P3), five (P5) or ten (P10) elements.

		Naïve					Sample-Splitting				
	n	T1ER	P1	P3	P5	P10	T1ER	P1	P3	P5	P10
LSKM	100	0.082	0.157	0.241	0.357	0.655	0.067	0.094	0.119	0.154	0.303
	200	0.064	0.247	0.417	0.625	0.875	0.101	0.180	0.258	0.349	0.638
	400	0.041	0.392	0.660	0.869	0.972	0.088	0.278	0.458	0.647	0.880
	800	0.035	0.594	0.866	0.970	1.000	0.052	0.412	0.683	0.896	0.986
GraceI	100	0.036	0.082	0.124	0.161	0.359	0.049	0.078	0.102	0.117	0.331
	200	0.053	0.246	0.376	0.479	0.671	0.042	0.087	0.116	0.145	0.263
	400	0.050	0.634	0.849	0.939	0.993	0.084	0.259	0.380	0.482	0.648
	800	0.039	0.904	0.992	0.999	1.000	0.067	0.632	0.840	0.914	0.952

which is crucial for controlling the type-I error rate and is guaranteed to happen with high probability asymptotically, is less likely to happen with the smaller samples available for the sample-splitting estimator. In addition, we can see that LSKM has better power than the GraceI test for smaller sample sizes (LSKM also has a slightly worse control on the type-I error rate than GraceI). But as sample size increases, the power of GraceI eventually surpasses LSKM. Finally, as expected, the probability of rejecting $H_{0,j}^* : ne_j^{\mathrm{I}} = ne_j^{\mathrm{II}}$ is higher when ne_j^{I} and ne_j^{II} differ by more elements. Detailed values of estimated powers and type-I errors are shown in Table 4.1.

4.6 Real Data Examples

4.6.1 Application to Brain Imaging Data

Mild, uncomplicated traumatic brain injury (TBI), or concussion, can occur from a variety of head injury exposures. In youth, sports and recreational activities comprise a predominate

number of these exposures with uncomplicated mild comprising the vast majority of TBIs. By definition, these are diagnostically ambiguous injuries with no radiographic findings on conventional CT or MRI. While some children recover just fine, a subset remain symptomatic for a sustained period of time. This group—often referred to as the 'miserable minority'—make up the majority of the patient population in concussion clinics. Newer imaging methods are needed to provide more sensitive diagnostic and prognostic screening tools that elucidate the underlying pathophysiological changes in these concussed youth whose symptoms do not resolve. To this end, a collaborative team evaluated 10-14 year olds following a sports or recreational concussion who remained symptomatic at 3-4 weeks post-injury and a group of age and gender matched controls with no history of head injury, psychological health diagnoses, or learning disabilities. Advanced neuroimaging was collected on each participant which included collecting diffusion tensor imaging (DTI). DTI has been shown to be sensitive to more subtle changes in white matter that have been reported to strongly correlate with axonal injury pathology Bennett et al. (2012), Mac Donald et al. (2007) and relate to outcome in other concussion groups Bazarian et al. (2012), Cubon et al. (2011), Gajawelli et al. (2013).

MRI scans were completed on a 3T Phillips Achieva with a 32-channel head coil. Each imaging session lasted 40 minutes and included a 1mm isotropic MPRAGE (5:13), 1mm isotropic 3D T2-weighted image (5:22), 1mm isotropic 3D T2-Star (3:41), 2D FLAIR collected at an in-plane resolution of 1 x 1mm with a slice thickness of 4mm, no gaps (2:56), 3mm isotropic BOLD image for resting-state fMRI (6:59), and a 32 direction 2mm isotropic diffusion sequence acquired with reverse polarity (A-P, P-A), b=1000 sec/mm2, and 6 non-diffusion weighted images for diffusion tensor imaging (DTI) analysis (each 6:39). The DTI data was then post-processed by the collaborative group by the following methods before the data was then utilized for the current analysis. Briefly, the first portion of the pipeline uses TORTOISE - Tolerably Obsessive Registration and Tensor Optimization Indolent Software Ensemble (Pierpaoli et al., 2010). For reverse polarity data, each DWI acquisition both A-P and P-A is initially run through DiffPrep (Oishi et al., 2009, Zhang et al., 2010) in TORTOISE for susceptibility distortion correction, motion correction, eddy current correction,

and registration to 3D high resolution structural image. For EPI distortion correction, the diffusion images were registered to the 1mm isotropic T2 image using non-linear b-splines. Eddy current and motion distortion were corrected using standard affine transformations followed by re-orientation of the b-matrix for the rotational aspect of the rigid body motion. Following DiffPrep, the output images from both the A-P and P-A DWI acquisitions were then sent through Diffeomorphic Registration for Blip-Up Blip-Down Diffusion Imaging (DR-BUDDI, Irfanoglu et al., 2015) in TORTOISE for further EPI distortion and eddy current correction that can be completed with diffusion data that has been collected with reverse polarity. This step combines the reverse polarity imaging data creating a single, cleaned, DWI data set that is then sent through DiffCalc (Pierpaoli et al., 2010, Koay et al., 2006, 2009, Basser et al., 1994, Mangin et al., 2002, Chang et al., 2005, 2012, Rohde et al., 2005) in TORTOISE. This step completes the tensor estimation using the robust estimation of tensors by outlier rejection (RESTORE)10 approach. Following tensor estimation, a variety of DTI metrics can be derived. For this study, we specifically focused on fractional anisotropy (FA) as our main metric.

Following this post-processing in TORTOISE, 3D image stacks for MD and FA were introduced into DTIstudio (Zhang et al., 2010, Oishi et al., 2009) for segmentation of the DTI atlas (Mori et al., 2010) on to each participants DTI data set in 'patient space' through the Diffeomap program in DTIstudio using both linear and non-linear transformations. This is a semi-automated process that allows for the extraction of DTI metrics within each 3D-atlas-based region of interest providing a comprehensive sampling throughout the entire brain into 189 regions including ventricular space. For this study, selection of regions was limited to regions of white matter as the main hypothesis regarding DTI was that there would be reductions in white matter integrity observed with FA related to brain injury. This reduced the number of regions used for further analysis down to 78. To select only white matter, FA images were then threshold at 0.2 or greater, and this final 3D segmentation was then applied to all other co-registered DTI metrics and the data within each DTI metric for the 78 regions of interest was extracted in ROIeditor for further analysis.

Upon preprocessing the data, we obtained data on $p = 78$ brain regions from $n^{\mathrm{I}} = 27$ healthy controls and $n^{\mathrm{II}} = 25$ TBI patients. To assess whether brain connectivity patterns of TBI patients differs from that of healthy controls, we used the DCA framework with the GraceI test after naïve lasso neighborhood selection. We chose the lasso tuning parameter using10-fold CV and controlled the FWER of falsely rejecting $H^*_{0,j} : ne^{\mathrm{I}}_j = ne^{\mathrm{II}}_j$ for any $j = 1, \ldots, p$ at level 0.1 using the Holm procedure. The resulting brain connectivity networks are shown in Figures 4.5–4.7, where differentially connected and common edges are drawn in different colors. It can be seen that a number of connections differ between TBI patients and healthy controls. The DTI data used in this study provide characterize microstructural changes in brain regions and not their functions. The assumption of multivariate normality may also not be realistic in this application. Finally, the study is based on small samples. Nonetheless, the results suggest orchestrated structural changes in TBI patients that may help form new hypotheses.

4.6.2 Application to Cancer Genetics Data

Based on the expression of estrogen receptor (ER), breast cancer can be classified into ER positive (ER+) and ER negative (ER-) categories. ER+ breast cancer has a larger number of estrogen receptors, and has better survival than ER- breast cancer (see, e.g., Carey et al., 2006). To investigate the difference in genetic pathways between ER+ and ER- breast cancer patients, we obtain gene expression data from TCGA. The data contain the expression level of $p = 358$ genes for $n^{\mathrm{I}} = 117$ ER- and $n^{\mathrm{II}} = 407$ ER+ breast cancer patients.

Similar to the example with brain imaging dataset, we use the GraceI test after naïve lasso neighborhood selection to examine the difference in genetic pathways between ER+ and ER- breast cancer patients. Lasso neighborhood selection tuning parameter is selected using 10-fold CV. FWER is controlled at level 0.1 using the Holm procedure. Edges that are differentially connected are shown in Figure 4.8.

Figure 4.5: Left view of common and differentially connected edges between concussed youth athletes and matched controls. Red and blue nodes are brain regions on the left and right cerebral cortices, respectively, whereas pink and turquoise nodes are other regions in the left and right brains. Gray edges are estimated common brain connections based on lasso neighborhood selection; blue edges are connections that are found in healthy controls but not in TBI patients; red edges are connections that are found in TBI patients but not in healthy controls.

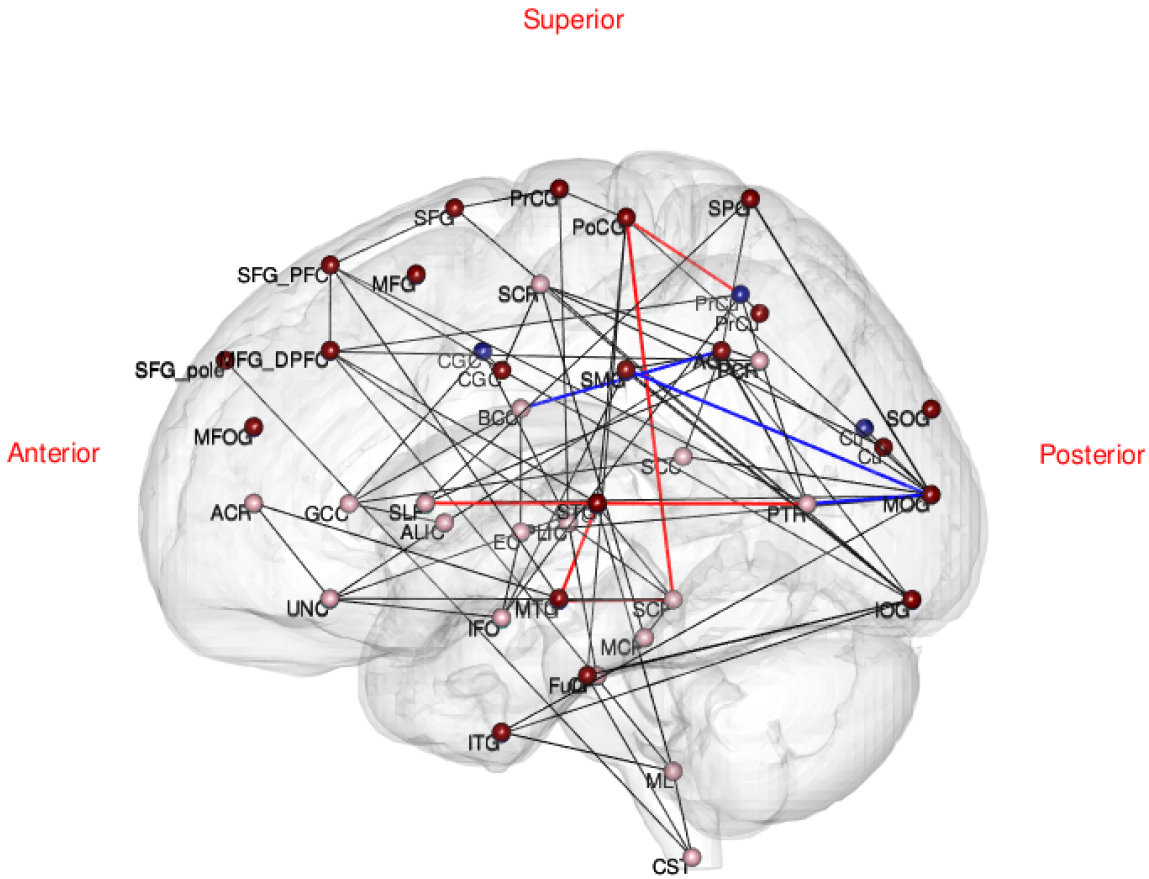


Figure 4.6: Posterior view of common and differentially connected edges between concussed youth athletes and matched controls. See the caption of Figure 4.5 for details of the figure.

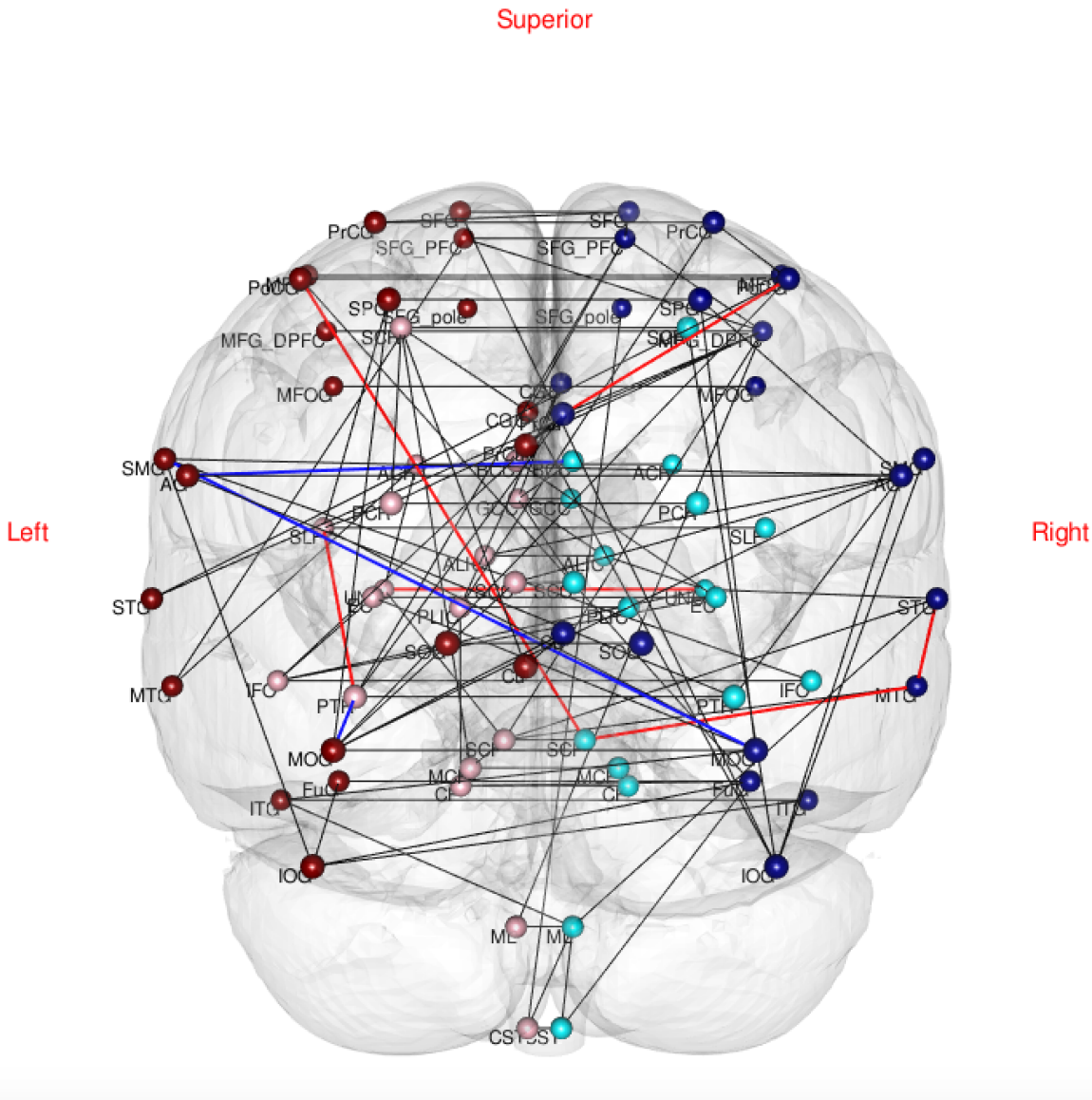


Figure 4.7: Posterior view of common and differentially connected edges between concussed youth athletes and matched controls. See the caption of Figure 4.5 for details of the figure.

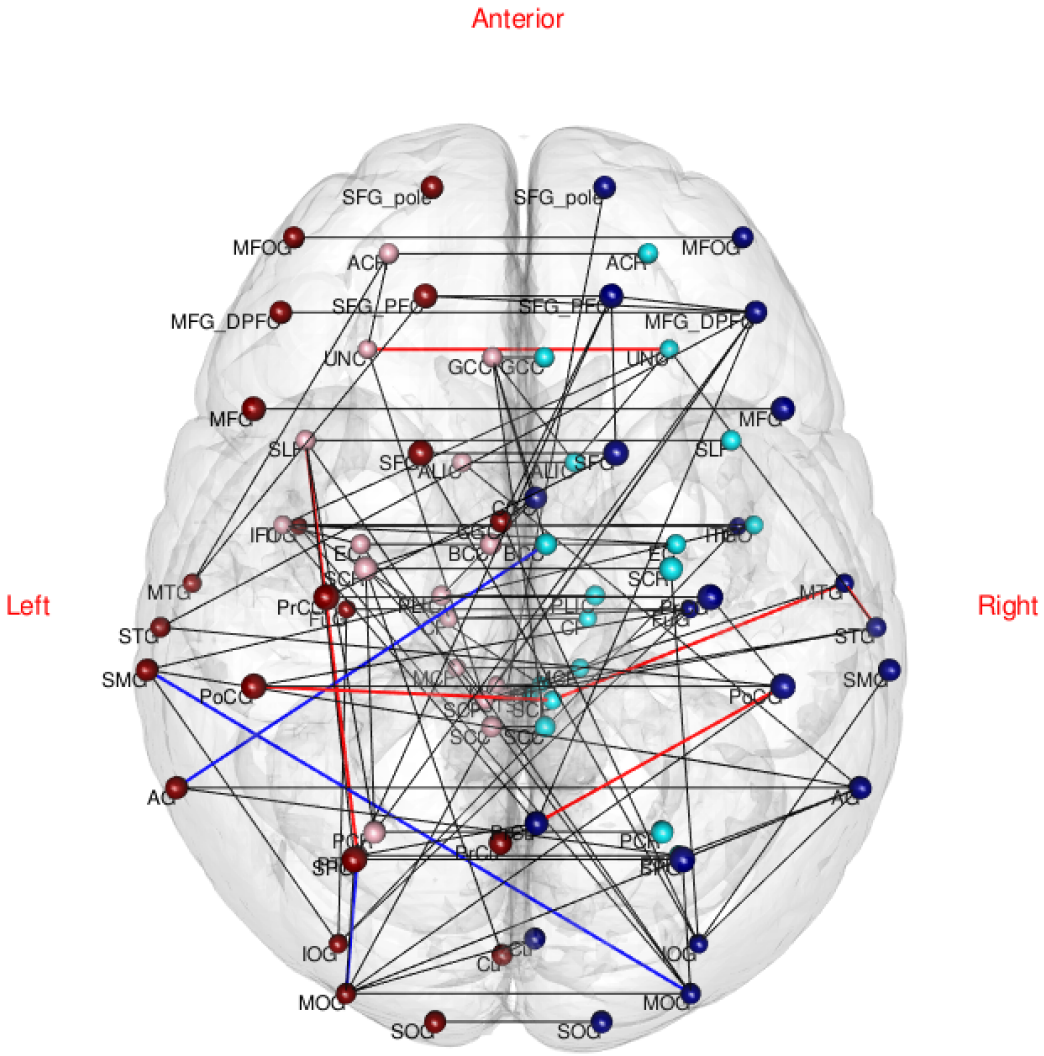
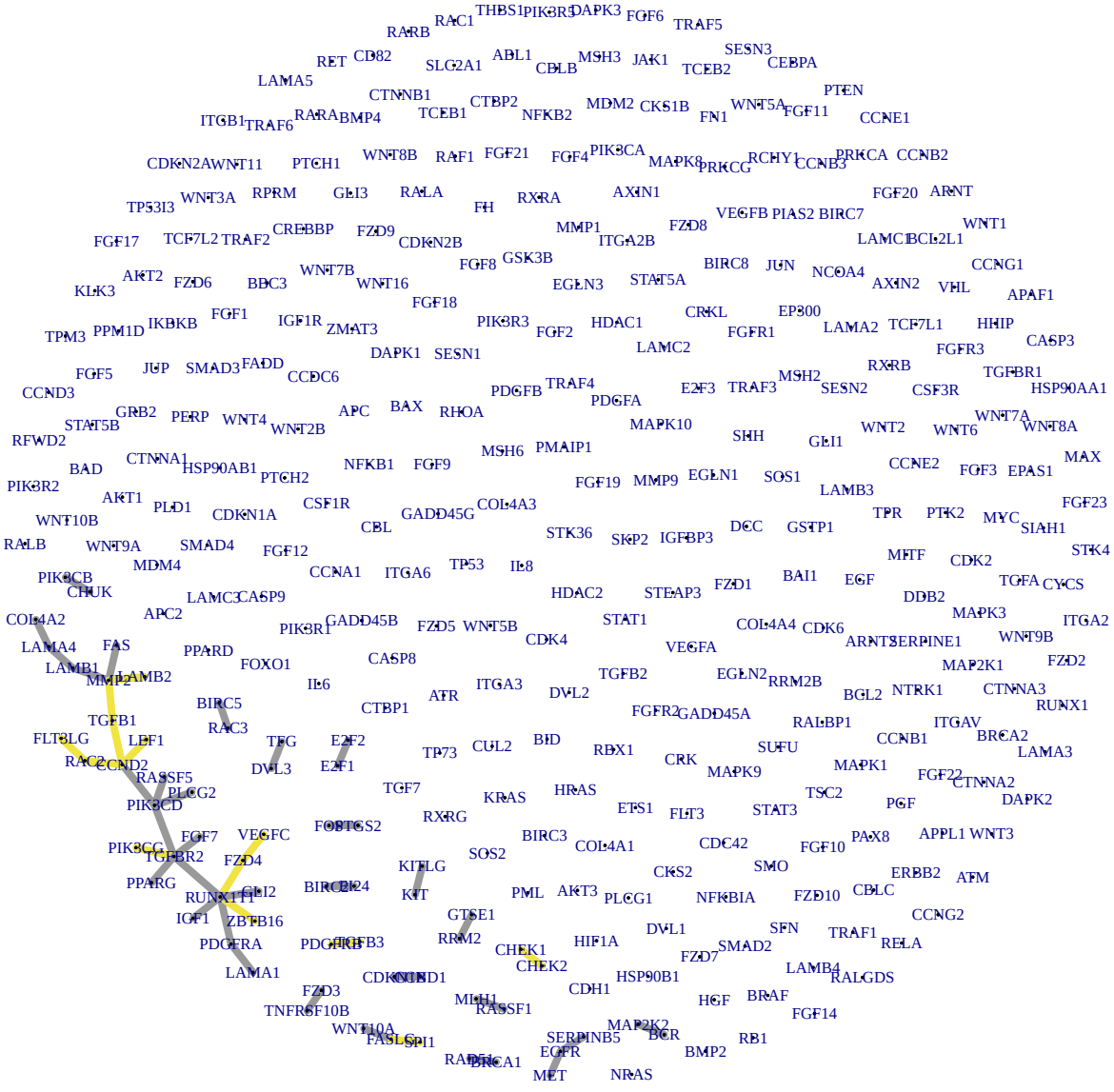


Figure 4.8: Differentially connected edges between ER- and ER+ breast cancer patients. Yellow edges are genetic interactions that are found in ER- but not ER+ breast cancer patients by the GraceI test; gray edges are genetic interactions that are found in ER+ but not ER- breast cancer patients by the GraceI test.



4.7 Discussion

In this chapter, we highlighted challenges of identifying differential connectivity in high-dimensional networks using existing approaches, and proposed a new inference framework, called *differential connectivity analysis* (DCA), for identifying differences in two graphical models. The DCA framework can incorporate various estimation and hypothesis testing methods, and can be easily extended to test for differential connectivity in multiple networks. Below are extensions and improvements to the DCA framework that can be fruitful topics of future research.

Investigating the choice of tuning parameters: In the simulation studies and real data example, for simplicity, we chose our lasso neighborhood selection tuning parameters using cross validation (CV). However, to make the selected set $\hat{ne}_j^0$ deterministic with high probability, the tuning parameter should be (asymptotically) independent of the data. It would be of interesting to developing non-data-adaptive methods for choosing tuning parameters that satisfy condisiton (**A2**) and (**A3**).

Investigating properties of other estimation methods: In Section 4.4 we presented two propositions that support the use of lasso neighborhood selection in DCA without sample-splitting. These properties are likely not exclusive to the lasso. Future work may aim at discovering similar properties for other estimation methods, which may require milder conditions. In addition, future work may aim at exploring the possibility of incorporating joint estimation methods of multiple networks into DCA.

Easing the computational complexity: To examine differential connectivity of all p nodes, our proposal requires p node-wise regressions, which can be computationally expensive, especially in high dimensions. Our current implementation in the R-package `netdiff` handles moderate size problems, as shown in the study of changes in regulatory networks of $p = 358$ cancer related genes across breast cancer subtypes in Section 4.6. However, the development of estimation and inference procedures with improved computational complexity can significantly facilitate the application of the proposed framework.

Exploring selective inference procedures: In this chapter, we proposed two methods for estimation: sample-splitting, which breaks down the dependence of estimation and hypothesis testing, and naïve inference, which utilizes the fact that the estimated support of lasso is deterministic with high probability. Another option for handling this problem is the use of recent advances in conditional inference procedures (see, e.g., Lee et al., 2016, Tibshirani et al., 2016). These procedures provide higher power than sample-splitting alternatives, and rely on fewer assumptions than the naïve approach. However, they are only appropriate to test conditional independence corresponding to the small set of selected variables, whereas in this chapter, we are interested in testing conditional independence corresponding to the potentially large set of *unselected* variables, corresponding to nodes that are not in the estiamted common neighborhood of each node. Exploring whether conditional inference procedures can be adapted to fit into our framework could be very valuable.

Chapter 5

DISCUSSION

5.1 Summary

In this dissertation, we developed three hypothesis testing procedures for linear and graphical models.

In Chapter 2, motivated by the simulation results of Leeb et al. (2015), we described a simple and naïve two-step hypothesis testing procedure for linear models. Conventional statistical wisdom indicates that such a naïve procedure could not control type-I error rate due to the issue of "double-peeking". However, by proving a long-overseen fact that with appropriate assumptions, the set of variables selected by the lasso is deterministic with high probability, we established the asymptotic validity of this naïve two-step procedure. Based on this theoretical property of the lasso, we also proposed the naïve score test, which, unlike existing post-selection inference proposals, could examine hypotheses associated with all regression coefficients, not just those selected by lasso. Finally, we established the connections between these two naïve procedures and several debiased lasso tests, and thus bridged the gap between the classes of post-selection inference and debiased lasso tests.

In Chapter 3, we proposed the Grace test for linear regression. The Grace test incorporates external graphical information about the relationship among variables into hypothesis testing. Such information can be derived from, e.g., gene regulatory networks and phylogenetic trees. We showed, both theoretically and through simulation studies, that when the external information is informative, the power of the Grace test could be substantially higher than existing proposals that ignore such information. Because the external graphical information could potentially be inaccurate or uninformative, we further proposed the GraceR test, which adaptively chooses the amount of external information to be incorporated based

on the data. We showed through simulation studies that even when the incorporated external information is wildly inaccurate, the GraceR test continues to provide a viable alternative to existing approaches.

In Chapter 4, we proposed a two-step hypothesis testing framework to examine whether the edge sets of two GGMs are the same. Unlike earlier proposals, which focus on testing the equality of two population covariance or inverse covariance matrices, our proposal examines the support of two inverse covariance matrices. As illustrated in the chapter, such a treatment would lead to more interpretable results. We showed theoretically and numerically that our proposed framework could correctly control type-I error rate.

5.2 Unsolved Issues

In this section, we highlight some of the issues that were not addressed in this dissertation. These can be viable directions of future research.

- *The choice of tuning parameter is an open question.* In Chapter 2, we require that the lasso tuning parameter satisfies $\lambda \succ \sqrt{\log(p)/n}$. It also should not depend on the outcome $\boldsymbol{y}$; otherwise, the set of variables selected by the lasso is not deterministic due to the randomness of lasso tuning parameter. In Chapter 3, we require the Grace and GraceR tuning parameters to not depend on the outcome $\boldsymbol{y}$; the randomness in the tuning parameter complicates the distribution of Grace and GraceR test statistics. Methods to select tuning parameters that achieve these requirements still need to be explored.

- *The degree of stringency of Conditions (***T***) in Chapter 2 and (***A3***) in Chapter 4 is unknown.* To the best of our knowledge, these two highly related conditions have not appeared in the existing literature. Thus, it would be beneficial to investigate how stringent these two conditions are. It is also worthwhile to explore through extensive real-world-data-guided simulation studies whether the set of variables selected by the

lasso is deterministic with very large p and n. The result of such numerical studies could be used to infer the likelihood of satisfaction of assumptions in real-world settings.

- *Efficient computer software programs are lacking.* High-dimensionality imposes higher requirements on the computational efficiency and memory efficiency of computer programs. Most of simulations presented in this dissertation with moderate n and p require several weeks to finish. In real-world settings with large n and p, researchers may need to spend several weeks or even months to obtain the analysis result with existing software, which likely will prevent researchers from adopting these modern statistical methodologies. Thus, more work needs to be done to develop more efficient computer software programs for the analysis of high-dimensional data.

5.3 Future Research

The research in this dissertation can be extended in several possible directions. For example, the naïve hypothesis testing procedures presented in Chapter 2 and Grace and GraceR tests presented in Chapter 3 all use penalized least squares for estimation. Adapting these methodologies to penalized likelihood-based M-estimators, e.g., penalized logistic and Poisson regressions, and penalized survival models could be a fruitful area of research. Also, as shown in the numerical experiments in Dezeure et al. (2015), existing high-dimensional hypothesis testing methods only control type-I error rate in certain settings. It is thus important to explore more extensively the settings under which existing methods fail, so that researchers could obtain a better understanding of the stringency of assumptions, and develop more robust and assumption-free high-dimensional hypothesis testing procedures. Finally, all the methodologies presented in this dissertation require independent data. However, most of spatial, temporal or clustered data do not meet such independence assumption. Therefore, it is also important to consider hypothesis testing methods with high-dimensional correlated observations.

In the following, we outline a hypothesis testing framework for temporally correlated

data by focusing on vector autoregression (VAR) models for time series. Times series VAR models are commonly used in econometrics (Sims, 1980, Stock and Watson, 2001, Bernanke et al., 2005), functional genomics (Shojaie and Michailidis, 2010a, Michailidis and d'Alché Buc, 2013) and FMRI brain imaging studies (Friston, 2009, Smith, 2012, Seth et al., 2013).

Let $\{\boldsymbol{x}^t\}$, $t = 1, \ldots, T$ be a p-dimensional stationary time series, which may, for instance, represent the brain activity level at p brain ROIs at time t. We may assume that $\boldsymbol{x}^t \in \mathbb{R}^p$ is dependent on the brain activity levels of the past d time points. A VAR(d) model with serially uncorrelated Gaussian errors may be used to describe this mechanism, i.e.,

$$\boldsymbol{x}^t = \boldsymbol{A}^1\boldsymbol{x}^{t-1} + \boldsymbol{A}^2\boldsymbol{x}^{t-2} + \cdots + \boldsymbol{A}^d\boldsymbol{x}^{t-d} + \boldsymbol{\epsilon}^t, \qquad t = d+1, \ldots, T, \tag{5.1}$$

where $\boldsymbol{A}^1, \ldots, \boldsymbol{A}^d \in \mathbb{R}^{p\times p}$ are transition matrices, and $\boldsymbol{\epsilon}^t \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$ is a vector of Gaussian random error with population covariance $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$.

To infer the functional association between brain ROIs, transition matrices $\boldsymbol{A}^1, \ldots, \boldsymbol{A}^d$ need to be estimated. Because there are dp^2 unknown parameters, VAR estimation is usually a high-dimensional problem. Penalties are often imposed on least-squares criterion or log-likelihood to impose sparsity (see, e.g., Qiu et al., 2015, Fan et al., 2015, Melnyk and Banerjee, 2016, Davis et al., 2016, Guo et al., 2016). However, despite recent advances in estimation methodologies, the problem of hypothesis testing with VAR models is still relatively unexplored.

Motivated by the recent development in debiased tests for independent data, we propose here a debiased test for time series VAR problem. We write the VAR(d) model as:

$$\underbrace{\begin{bmatrix} (\boldsymbol{x}^T)^\top \\ \vdots \\ (\boldsymbol{x}^{d+1})^\top \end{bmatrix}}_{\boldsymbol{Y}} = \underbrace{\begin{bmatrix} (\boldsymbol{x}^{T-1})^\top & \ldots & (\boldsymbol{x}^{T-d})^\top \\ \vdots & \ddots & \vdots \\ (\boldsymbol{x}^{d})^\top & \ldots & (\boldsymbol{x}^{1})^\top \end{bmatrix}}_{\boldsymbol{X}} \underbrace{\begin{bmatrix} (\boldsymbol{A}^1)^\top \\ \vdots \\ (\boldsymbol{A}^d)^\top \end{bmatrix}}_{\boldsymbol{B}} + \underbrace{\begin{bmatrix} (\boldsymbol{\epsilon}^T)^\top \\ \vdots \\ (\boldsymbol{\epsilon}^{d+1})^\top \end{bmatrix}}_{\boldsymbol{E}}. \tag{5.2}$$

After vectorization of (5.2), we can conveniently write the VAR(d) problem in the linear

form,

$$\begin{aligned} \boldsymbol{y} \equiv \text{vec}(\boldsymbol{Y}) &= \text{vec}(\boldsymbol{X}\boldsymbol{B}) + \text{vec}(\boldsymbol{E}) \\ &= (\boldsymbol{I} \otimes \boldsymbol{X})\text{vec}(\boldsymbol{B}) + \text{vec}(\boldsymbol{E}) \\ &= \boldsymbol{Z}\boldsymbol{\beta} + \text{vec}(\boldsymbol{E}), \end{aligned} \tag{5.3}$$

where $\boldsymbol{Z} \equiv \boldsymbol{I} \otimes \boldsymbol{X} \in \mathbb{R}^{(T-d)p \times dp^2}$and $\boldsymbol{\beta} \equiv \text{vec}(\boldsymbol{B}) \in \mathbb{R}^{dp^2}$. Thus, entries in $\boldsymbol{\beta}$ correspond to entries in the transition matrices $\boldsymbol{A}^1, \ldots, \boldsymbol{A}^d$, and making inference on $\boldsymbol{A}^1, \ldots, \boldsymbol{A}^d$ is equivalent to making inference on $\boldsymbol{\beta}$.

Motivated by debiased tests, we propose to estimate $\boldsymbol{\beta}$ by the linear estimator

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{M}\boldsymbol{Z}^\top \boldsymbol{Y}, \tag{5.4}$$

where $\boldsymbol{M}$ is some matrix derived from $\boldsymbol{Z}$. With low-dimensional data, setting $\boldsymbol{M} = (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}$ results in the unbiased least squares estimator. With high-dimensional data, $\boldsymbol{Z}^\top \boldsymbol{Z}$ is not invertible, and $\boldsymbol{M}$ may be obtained in various ways. For example, in LDPE (Zhang and Zhang, 2014, van de Geer et al., 2014), $\boldsymbol{M}$ is obtained through node-wise lasso linear regression, whereas in SSLasso (Javanmard and Montanari, 2014a), $\boldsymbol{M}$ is obtained through ℓ_1-penalized linear programming. We discuss the requirements on $\boldsymbol{M}$ for VAR inference later.

With Gaussian error and conditional on $\boldsymbol{Z}$, $\tilde{\boldsymbol{\beta}}$ follows a normal distribution, given by

$$\tilde{\boldsymbol{\beta}} | \boldsymbol{Z} \sim \mathcal{N}_{dp^2}\left(\boldsymbol{M}\boldsymbol{Z}^\top \boldsymbol{Z}\boldsymbol{\beta}, \boldsymbol{M}\boldsymbol{Z}^\top \text{Var}\left(\text{vec}\left(\boldsymbol{E}\right)\right) \boldsymbol{Z}\boldsymbol{M}^\top\right). \tag{5.5}$$

Since

$$\boldsymbol{E} \equiv \begin{bmatrix} (\boldsymbol{\epsilon}^T)^\top \\ \vdots \\ (\boldsymbol{\epsilon}^{d+1})^\top \end{bmatrix},$$

where $\boldsymbol{\epsilon}^T \sim \cdots \sim \boldsymbol{\epsilon}^{d+1} \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$ with no serial correlation, $\text{Var}(\text{vec}(\boldsymbol{E})) = \boldsymbol{I} \otimes \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$, and hence,

$$\tilde{\boldsymbol{\beta}} | \boldsymbol{Z} \sim \mathcal{N}_{dp^2}\left(\boldsymbol{M}\boldsymbol{Z}^\top \boldsymbol{Z}\boldsymbol{\beta}, \boldsymbol{M}\boldsymbol{Z}^\top \left(\boldsymbol{I} \otimes \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\right) \boldsymbol{Z}\boldsymbol{M}^\top\right).$$

To compensate the bias of $\tilde{\boldsymbol{\beta}}$, we add a one-step adjustment, $(\boldsymbol{I}-\boldsymbol{M}\boldsymbol{Z}^\top\boldsymbol{Z})\hat{\boldsymbol{\beta}}$, to $\tilde{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is some "initial" estimator of $\boldsymbol{\beta}$ whose estimation accuracy is theoretically guaranteed. We define the test statistic

$$\boldsymbol{z} = \tilde{\boldsymbol{\beta}} + \left(\boldsymbol{I} - \boldsymbol{M}\boldsymbol{Z}^\top\boldsymbol{Z}\right)\hat{\boldsymbol{\beta}}. \tag{5.6}$$

Combining (5.5) with (5.6), we obtain

$$\sqrt{m}\left(\boldsymbol{z}-\boldsymbol{\beta}\right)|\boldsymbol{Z} \sim \mathcal{N}_{dp^2}\left(\boldsymbol{0}, m\boldsymbol{M}\boldsymbol{Z}^\top\left(\boldsymbol{I}\otimes\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\right)\boldsymbol{Z}\boldsymbol{M}^\top\right) + \sqrt{m}\left(\boldsymbol{M}\boldsymbol{Z}^\top\boldsymbol{Z}-\boldsymbol{I}\right)\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right). \tag{5.7}$$

In this formulation, m is the smallest number such that the minimum eigenvalue of the covariance matrix $m\boldsymbol{M}\boldsymbol{Z}^\top\left(\boldsymbol{I}\otimes\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\right)\boldsymbol{Z}\boldsymbol{M}^\top$ is bounded away from zero with high probability.

The second term in (5.7), $\sqrt{m}(\boldsymbol{M}\boldsymbol{Z}^\top\boldsymbol{Z}-\boldsymbol{I})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})$, characterizes the bias of the test statistic $\boldsymbol{z}$. Since the distribution in the first term is non-degenerate, we want

$$\left\|\sqrt{m}\left(\boldsymbol{M}\boldsymbol{Z}^\top\boldsymbol{Z}-\boldsymbol{I}\right)\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right\|_\infty = \mathcal{O}_p(1) \tag{5.8}$$

to make

$$\sqrt{m}\left(\boldsymbol{z}-\boldsymbol{\beta}\right)|\boldsymbol{Z} \to_d \mathcal{N}_{dp^2}\left(\boldsymbol{0}, m\boldsymbol{M}\boldsymbol{Z}^\top\left(\boldsymbol{I}_n\otimes\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\right)\boldsymbol{Z}\boldsymbol{M}^\top\right), \tag{5.9}$$

which could then be used to derive p-values and confidence intervals.

Equation (5.8) characterizes the requirement on $\boldsymbol{M}$ and $\hat{\boldsymbol{\beta}}$. The estimation accuracy of $\hat{\boldsymbol{\beta}}$ in the case of the lasso is established in Propositions 4.1 and 4.3 of Basu and Michailidis (2015), which show that the lasso in the time series setting achieves the same rate of estimation as in the setting with independent observations, up to a multiplicative factor characterizing the dependance of the data. Some theoretical properties on $\boldsymbol{M}$ generated by the graphical lasso are developed in Chen et al. (2013), whereas theoretical properties of a Dantzig-selector-based estimate of $\boldsymbol{M}$ are explored in Chen et al. (2016). However, it is still unclear whether any of these estimate of $\boldsymbol{M}$ together with a lasso estimate of $\hat{\boldsymbol{\beta}}$ achieve (5.8).

We now present a method to use node-wise lasso regression to construct $\boldsymbol{M}$, and derive its bound of $\|\sqrt{m}\left(\boldsymbol{M}\boldsymbol{Z}^\top\boldsymbol{Z}-\boldsymbol{I})(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\right)\|_\infty$. Since

$$\left\|\sqrt{m}\left(\boldsymbol{M}\boldsymbol{Z}^\top\boldsymbol{Z}-\boldsymbol{I}\right)\left(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\right)\right\|_\infty \leq \sqrt{m}\left\|\boldsymbol{M}\boldsymbol{Z}^\top\boldsymbol{Z}-\boldsymbol{I}\right\|_{\max}\left\|\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\right\|_1,$$

where $\|\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\|_1$ is derived in Propositions 4.1 and 4.3 of Basu and Michailidis (2015), we here focus on the bound of $\|\boldsymbol{M}\boldsymbol{Z}^\top\boldsymbol{Z}-\boldsymbol{I}\|_{\max}$.

Let

$$\boldsymbol{M}=\frac{1}{(T-d)p}\cdot\begin{bmatrix}\frac{1}{\hat{w}_1^2} & \frac{\hat{\gamma}_2^1}{\hat{w}_1^2} & \cdots & \frac{\hat{\gamma}_{dp^2}^1}{\hat{w}_1^2}\\ \frac{\hat{\gamma}_1^2}{\hat{w}_2^2} & \frac{1}{\hat{w}_2^2} & \cdots & \frac{\hat{\gamma}_{dp^2}^2}{\hat{w}_2^2}\\ \vdots & \vdots & \ddots & \vdots\\ \frac{\hat{\gamma}_1^{dp^2}}{\hat{w}_{dp^2}^2} & \frac{\hat{\gamma}_2^{dp^2}}{\hat{w}_{dp^2}^2} & \cdots & \frac{1}{\hat{w}_{dp^2}^2}\end{bmatrix}, \tag{5.10}$$

where for $j\in\{1,\ldots,dp^2\}$,

$$\hat{\boldsymbol{\gamma}}^j=\underset{\boldsymbol{g}\in\mathbb{R}^{dp^2-1}}{\arg\min}\left\{\frac{1}{2(T-d)p}\left\|\boldsymbol{Z}_j-\boldsymbol{Z}_{\setminus j}\boldsymbol{g}\right\|_2^2+\lambda_j\|\boldsymbol{g}\|_1\right\} \tag{5.11}$$

is the lasso node-wise regression coefficient estimate with tuning parameter λ_j and

$$\hat{w}_j^2=\frac{1}{(T-d)p}\left\|\boldsymbol{Z}_j-\boldsymbol{Z}_{\setminus j}\hat{\boldsymbol{\gamma}}^j\right\|_2^2+\lambda_j\left\|\hat{\boldsymbol{\gamma}}^j\right\|_1 \tag{5.12}$$

The stationary condition (Karush, 1939, Kuhn and Tucker, 1951) of (5.11) implies that

$$\boldsymbol{Z}_{\setminus j}^\top\left(\boldsymbol{Z}_j-\boldsymbol{Z}_{\setminus j}\hat{\boldsymbol{\gamma}}^j\right)=(T-d)p\lambda_j\hat{\boldsymbol{\tau}}^j, \tag{5.13}$$

where $\hat{\boldsymbol{\tau}}^j$ is the sub-gradient of the objective function, $\hat{\tau}_k^j=\text{sign}(\hat{\gamma}_k^j)$ if $\hat{\gamma}_k^j\neq 0$ and $\hat{\tau}_k^j\in[-1,1]$ if $\hat{\gamma}_k^j=0$. Dividing both sides of (5.13) by $\hat{w}_j^2$, we obtain

$$\boldsymbol{Z}_{\setminus j}^\top\left(\frac{1}{\hat{w}_j^2}\boldsymbol{Z}_j-\frac{1}{\hat{w}_j^2}\boldsymbol{Z}_{\setminus j}\hat{\boldsymbol{\gamma}}^j\right)=(T-d)p\frac{\lambda_j}{\hat{w}_j^2}\hat{\boldsymbol{\tau}}^j, \tag{5.14}$$

where

$$\frac{1}{\hat{w}_j^2}\boldsymbol{Z}_j-\frac{1}{\hat{w}_j^2}\boldsymbol{Z}_{\setminus j}\hat{\boldsymbol{\gamma}}^j=(T-d)p\boldsymbol{Z}\boldsymbol{M}_{(j,\cdot)}^\top.$$

Thus, based on (5.14),

$$\boldsymbol{Z}_{\setminus j}^{\top}\boldsymbol{Z}\boldsymbol{M}_{(j,\cdot)}^{\top} = \frac{\lambda_j}{\hat{w}_j^2}\hat{\boldsymbol{\tau}}^j, \tag{5.15}$$

which implies that

$$\left\|\boldsymbol{M}_{(j,\cdot)}\boldsymbol{Z}^{\top}\boldsymbol{Z}_{\setminus j}\right\|_{\infty} \leq \frac{\lambda_j}{\hat{w}_j^2}, \tag{5.16}$$

because $\|\hat{\boldsymbol{\tau}}^j\|_{\infty} \leq 1$.

On the other hand, expanding (5.12), we obtain

$$\hat{w}_j^2 = \frac{1}{(T-d)p}\boldsymbol{Z}_j^{\top}\left(\boldsymbol{Z}_j - \boldsymbol{Z}_{\setminus j}\hat{\boldsymbol{\gamma}}^j\right) - \frac{1}{(T-d)p}\hat{\boldsymbol{\gamma}}^{j\top}\boldsymbol{Z}_{\setminus j}^{\top}\left(\boldsymbol{Z}_j - \boldsymbol{Z}_{\setminus j}\hat{\boldsymbol{\gamma}}^j\right) + \lambda_j\left\|\hat{\boldsymbol{\gamma}}^j\right\|_1. \tag{5.17}$$

Note that

$$\left\|\hat{\boldsymbol{\gamma}}^j\right\|_1 = \sum_{k=1}^{dp^2-1}\left|\hat{\gamma}_k^j\right| = \sum_{k:\hat{\gamma}_k^j\neq 0}\operatorname{sign}(\hat{\gamma}_k^j)\hat{\gamma}_k^j + \sum_{k:\hat{\gamma}_k^j=0} 0\cdot\hat{\tau}_k^j = \sum_{k=1}^{dp^2-1}\hat{\tau}_k^j\hat{\gamma}_k^j = \hat{\boldsymbol{\gamma}}^{j\top}\hat{\boldsymbol{\tau}}^j.$$

Thus

$$\begin{aligned}\hat{w}_j^2 &= \frac{1}{(T-d)p}\boldsymbol{Z}_j^{\top}\left(\boldsymbol{Z}_j - \boldsymbol{Z}_{\setminus j}\hat{\boldsymbol{\gamma}}^j\right) - \frac{1}{(T-d)p}\hat{\boldsymbol{\gamma}}^{j\top}\left(\boldsymbol{Z}_{\setminus j}^{\top}\left(\boldsymbol{Z}_j - \boldsymbol{Z}_{\setminus j}\hat{\boldsymbol{\gamma}}^j\right) - (T-d)p\lambda_j\hat{\boldsymbol{\tau}}^j\right)\\ &= \frac{1}{(T-d)p}\boldsymbol{Z}_j^{\top}\left(\boldsymbol{Z}_j - \boldsymbol{Z}_{\setminus j}\hat{\boldsymbol{\gamma}}^j\right),\end{aligned} \tag{5.18}$$

where the last equality is based on (5.13). Dividing both sides of (5.18) by $\hat{w}_j^2$, we obtain

$$\boldsymbol{Z}_j^{\top}\left(\frac{1}{\hat{w}_j^2}\boldsymbol{Z}_j - \frac{1}{\hat{w}_j^2}\boldsymbol{Z}_{\setminus j}\hat{\boldsymbol{\gamma}}^j\right) = (T-d)p, \tag{5.19}$$

which implies that

$$\boldsymbol{Z}_j^{\top}\boldsymbol{Z}\boldsymbol{M}_{(j,\cdot)}^{\top} - 1 = 0. \tag{5.20}$$

Combining (5.16) with (5.20), we obtain

$$\left\|\boldsymbol{M}_{(j,\cdot)}\boldsymbol{Z}^{\top}\boldsymbol{Z} - \boldsymbol{e}^j\right\|_{\infty} \leq \frac{\lambda_j}{\hat{w}_j^2}, \tag{5.21}$$

where $\boldsymbol{e}^j$ is the row vector with the j-th entry equal to one. Thus

$$\left\|\boldsymbol{M}\boldsymbol{Z}^\top\boldsymbol{Z}-\boldsymbol{I}\right\|_{\max}\leq\max_j\left\{\frac{\lambda_j}{\hat{w}_j^2}\right\}. \tag{5.22}$$

We now propose an outline to derive an asymptotic upper bound for $\lambda_j/\hat{w}_j^2$. Specifically, we outline a procedure to obtain an asymptotic lower bound for $\hat{w}_j^2$. Let $\boldsymbol{\gamma}^j$ be the population regression parameter of $\boldsymbol{Z}_j$ on $\boldsymbol{Z}_{\setminus j}$, and let $\boldsymbol{\xi}^j$ be the random error, i.e.,

$$\boldsymbol{\xi}^j\equiv\boldsymbol{Z}_j-\boldsymbol{Z}_{\setminus j}\boldsymbol{\gamma}^j. \tag{5.23}$$

Recall that

$$\hat{w}_j^2=\frac{1}{(T-d)p}\left\|\boldsymbol{Z}_j-\boldsymbol{Z}_{\setminus j}\hat{\boldsymbol{\gamma}}^j\right\|_2^2+\lambda_j\left\|\hat{\boldsymbol{\gamma}}^j\right\|_1. \tag{5.24}$$

To derive the bound for the first term $\|\boldsymbol{Z}_j-\boldsymbol{Z}_{\setminus j}\hat{\boldsymbol{\gamma}}^j\|_2^2/(T-d)/p$, note that

$$\begin{aligned}\left\|\boldsymbol{Z}_j-\boldsymbol{Z}_{\setminus j}\hat{\boldsymbol{\gamma}}^j\right\|_2^2&=\left\|\boldsymbol{Z}_j-\boldsymbol{Z}_{\setminus j}\boldsymbol{\gamma}^j+\boldsymbol{Z}_{\setminus j}\boldsymbol{\gamma}^j-\boldsymbol{Z}_{\setminus j}\hat{\boldsymbol{\gamma}}^j\right\|_2^2\\&=\left\|\boldsymbol{Z}_j-\boldsymbol{Z}_{\setminus j}\boldsymbol{\gamma}^j\right\|_2^2+\left\|\boldsymbol{Z}_{\setminus j}\left(\hat{\boldsymbol{\gamma}}^j-\boldsymbol{\gamma}^j\right)\right\|_2^2+2\left(\boldsymbol{Z}_j-\boldsymbol{Z}_{\setminus j}\boldsymbol{\gamma}^j\right)^\top\boldsymbol{Z}_{\setminus j}\left(\hat{\boldsymbol{\gamma}}^j-\boldsymbol{\gamma}^j\right)\\&=\|\boldsymbol{\xi}\|_2^2+\left\|\boldsymbol{Z}_{\setminus j}\left(\hat{\boldsymbol{\gamma}}^j-\boldsymbol{\gamma}^j\right)\right\|_2^2+2\boldsymbol{\xi}^{j\top}\boldsymbol{Z}_{\setminus j}\left(\hat{\boldsymbol{\gamma}}^j-\boldsymbol{\gamma}^j\right).\end{aligned} \tag{5.25}$$

We would like to show

i with appropriate assumptions, $\|\boldsymbol{\xi}\|_2^2/(T-d)/p$ converges to a nonzero quantity in probability.

ii
$$\frac{1}{(T-d)p}\left\|\boldsymbol{Z}_{\setminus j}\left(\hat{\boldsymbol{\gamma}}^j-\boldsymbol{\gamma}^j\right)\right\|_2^2=\mathcal{O}_p(1).$$
Note that a similar bound has been derived in Propositions 4.1 and 4.3 in Basu and Michailidis (2015).

iii
$$\frac{1}{(T-d)p}\boldsymbol{\xi}^{j\top}\boldsymbol{Z}_{\setminus j}\left(\hat{\boldsymbol{\gamma}}^j-\boldsymbol{\gamma}^j\right)=\mathcal{O}_p(1).$$

Upper bound for $\|\hat{\boldsymbol{\gamma}}^j - \boldsymbol{\gamma}^j\|_2$ can be similarly derived as in Propositions 4.1 and 4.3 in Basu and Michailidis (2015). Therefore, we also need to show $\|\boldsymbol{\xi}^{j\top}\boldsymbol{Z}_{\setminus j}/(T-d)/p\|_\infty$ is small.

For the second term, $\lambda_j\|\hat{\boldsymbol{\gamma}}^j\|_1$, we have

$$\lambda_j \left\|\hat{\boldsymbol{\gamma}}^j\right\|_1 \leq \lambda_j \left\|\boldsymbol{\gamma}^j\right\|_1 + \lambda_j \left\|\hat{\boldsymbol{\gamma}}^j - \boldsymbol{\gamma}^j\right\|_1. \tag{5.26}$$

Since $\hat{\boldsymbol{\gamma}}^j$ is a lasso estimate of the regression coefficient, with similar assumptions as presented in Basu and Michailidis (2015), we would like to show

iv

$$\lambda_j \left\|\hat{\boldsymbol{\gamma}}^j - \boldsymbol{\gamma}^j\right\|_1 = o_p(1).$$

Thus, if we also assume $\lambda_j\|\boldsymbol{\gamma}^j\|_1 = o(1)$, then $\lambda_j\|\hat{\boldsymbol{\gamma}}^j\|_1 = o_p(1)$

If we can successfully show 1)–4), then $\hat{w}_j^2$ is asymptotically equivalent of $\|\boldsymbol{\xi}\|_2^2/(T-d)/p$, which converges in probability to a nonzero value. Thus, we have

$$\left\|\boldsymbol{M}\boldsymbol{Z}^\top\boldsymbol{Z} - \boldsymbol{I}\right\|_{\max} = \mathcal{O}_p(\lambda_j). \tag{5.27}$$

Alternatively, we can use a similar strategy as in Chapter 3 to devise valid hypothesis testing procedures. Specifically, we construct a stochastic bias bound for the distribution of $\boldsymbol{z}$, which is

$$\begin{aligned}\left\|\sqrt{m}\left(\boldsymbol{M}\boldsymbol{Z}^\top\boldsymbol{Z} - \boldsymbol{I}\right)\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)\right\|_\infty &\leq \left\|\sqrt{m}\left(\boldsymbol{M}\boldsymbol{Z}^\top\boldsymbol{Z} - \boldsymbol{I}\right)\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)\right\|_2 \\ &\leq \sqrt{m}\left\|\boldsymbol{M}\boldsymbol{Z}^\top\boldsymbol{Z} - \boldsymbol{I}\right\|_{\max}\left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|_1,\end{aligned} \tag{5.28}$$

where the bound on $\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_1$ is derived in Propositions 4.1 and 4.3 in Basu and Michailidis (2015). Following a similar procedure as Grace and GraceR tests in Chapter 3, we could perform asymptotically valid inference for $\boldsymbol{\beta}$ based on this asymptotic bias bound.

To summarize, to develop valid hypothesis testing procedure for high-dimensional VAR(d) models, we need to

1. establish the correct rate for the factor m;

2. consider various options for estimating $\boldsymbol{M}$ and explore whether they satisfy (5.8); one option is to show i–iv;

3. explore methods for consistent estimation of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$.

BIBLIOGRAPHY

Abbas, A. and Gupta, S. (2008). The role of histone deacetylases in prostate cancer. *Epigenetics*, 3(6):300–309.

Albert, R. (2005). Scale-free networks in cell biology. *Journal of Cell Science*, 118(21):4947–4957.

Bach, F. R. (2008). Bolasso: Model consistent lasso estimation through the bootstrap. In Cohen, W. W., Mccallum, A., and Roweis, S. T., editors, *Proceedings of the 25th International Conference on Machine Learning*, pages 33–40, Brookline, MA. Microtome Publishing.

Bai, Z., Jiang, D., Yao, J.-F., and Zheng, S. (2009). Corrections to LRT on large-dimensional covariance matrix by RMT. *The Annals of Statistics*, 37(6B):3822–3840.

Bai, Z. D. (1999). Methodologies in spectral analysis of large dimensional random matrices, a review. *Statistica Sinica*, 9(3):611–677.

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516.

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: A network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68.

Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113.

Basser, P. J., Mattiello, J., and Lebihan, D. (1994). Estimation of the effective self-diffusion *tensor* from the NMR spin echo. *Journal of Magnetic Resonance, Series B*, 103(3):247–254.

Bassett, D. S. and Bullmore, E. T. (2009). Human brain networks in health and disease. *Current Opinion in Neurology*, 22(4):340–347.

Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567.

Bazarian, J. J., Zhu, T., Blyth, B., Borrino, A., and Zhong, J. (2012). Subject-specific changes in brain white matter on diffusion tensor imaging after sports-related concussion. *Magnetic Resonance Imaging*, 30(2):171–180.

Belilovsky, E., Varoquaux, G., and Blaschko, M. B. (2016). Testing for differences in gaussian graphical models: Applications to brain connectivity. In Lee, D. D., Sugiyama, M., Luxberg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29, pages 595–603. Curran Associates, Inc., Red Hook, NY.

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studeis*, 81(2):608–650.

Belloni, A., Chernozhukov, V., and Wang, L. (2011). Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188.

Bennett, R. E., Mac Donald, C. L., and Brody, D. L. (2012). Diffusion tensor imaging

detects axonal injury in a mouse model of repetitive closed-skull traumatic brain injury. *Neuroscience Letters*, 513(2):160–165.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837.

Bernanke, B. S., Boivin, J., and Eliasz, P. (2005). Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach. *The Quarterly Journal of Economics*, 120(1):387–422.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.

Birke, M. and Dette, H. (2005). A note on testing the covariance matrix for large dimension. *Statistics & Probability Letters*, 74(3):281–289.

Boonstra, P. S., Mukherjee, B., and Taylor, J. M. G. (2015). A small-sample choice of the tuning parameter in ridge regression. *Statistica Sinica*, 25(3):1185–1206.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384.

Buena, F., Tsybakov, A., and Wegkamp, M. (2007). Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, pages 169–194.

Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4):1212–1242.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer-Verlag, Heidelberg, DE.

Cai, T., Liu, W., and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607.

Cai, T., Liu, W., and Xia, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association*, 108(501):265–277.

Cai, T. T., Li, H., Liu, W., and Xie, J. (2016). Joint estimation of multiple high-dimensional precision matrices. *Statistica Sinica*, 26(2):445–464.

Cai, T. T. and Ma, Z. (2013). Optimal hypothesis testing for high dimensional covariance matrices. *Bernoulli*, 19(5B):2359–2388.

Cai, T. T. and Zhang, A. (2016). Inference on high-dimensional differential correlation matrices. *Journal of Multivariate Analysis*, 143:107–126.

Campbell, C. L., Jiang, Z., Savarese, D. M. F., and Savarese, T. M. (2001). Increased expression of the interleukin-11 receptor and evidence of STAT3 activation in prostate carcinoma. *The American Journal of Pathology*, 158(1):25–32.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351.

Candès, E. J. and Plan, Y. (2009). Near-ideal model selection by ℓ_1 minimization. *The Annals of Statistics*, 37(5A):2145–2177.

Carey, L. A., Perou, C. M., Livasy, C. A., Dressler, L. G., Cowan, D., Conway, K., Karaca, G., Troester, M. A., Tse, C. K., Edmiston, S., Deming, S. L., Geradts, J., Cheang, M. C. U., Nielsen, T. O., Moorman, P. G., Earp, H. S., and Millikan, R. C. (2006). Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *JAMA*, 295(21):2492–2502.

Chang, L.-C., Jones, D. K., and Pierpaoli, C. (2005). RESTORE: Robust estimation of tensors by outlier rejection. *Magnetic Resonance in Medicine*, 53(5):1088–1095.

Chang, L.-C., Walker, L., and Pierpaoli, C. (2012). *Informed* RESTORE: A method for robust estimation of diffusion tensor from low redundancy datasets in the presence of physiological noise artifacts. *Magnetic Resonance in Medicine*, 68(5):1654–1663.

Chatterjee, A. and Lahiri, S. N. (2011). Bootstrapping Lasso estimators. *Journal of the American Statistical Association*, 106(494):608–625.

Chatterjee, A. and Lahiri, S. N. (2013). Rates of convergence of the Adaptive LASSO estimators to the Oracle distribution and higher order refinements by the bootstrap. *The Annals of Statistics*, 41(3):1232–1259.

Chen, C.-S., Wang, Y.-C., Yang, H.-C., Huang, P.-H., Kulp, S. K., Yang, C.-C., Lu, Y.-S., Matsuyama, S., Chen, C.-Y., and Chen, C.-S. (2007). Histone deacetylase inhibitors sensitize prostate cancer cells to agents that produce DNA double-strand breaks by targeting Ku70 acetylation. *Cancer Research*, 67(11):5318–5327.

Chen, S. S., Donoho, D. L., and Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61.

Chen, S. X., Zhang, L.-X., and Zhong, P.-S. (2010). Tests for high-dimensional covariance matrices. *Journal of the American Statistical Association*, 105(490):810–819.

Chen, X., Xu, M., and Wu, W. B. (2013). Covariance and precision matrix estimation for high-dimensional time series. *The Annals of Statistics*, 41(6):2994–3021.

Chen, X., Xu, M., and Wu, W. B. (2016). Regularized estimation of linear functionals of precision matrices for high-dimensional time series. *IEEE Transactions on Signal Processing*, 64(24):6459–6470.

Chung, F. R. K. (1997). *Spectral Graph Theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. American Mathematical Society, Providence, RI.

Clarke, R., Ressom, H. W., Wang, A., Xuan, J., Liu, M. C., Gehan, E. A., and Wang, Y. (2008). The properties of high-dimensional data spaces: Implications for exploring gene and protein expression data. *Nature Reviews Cancer*, 8(1):37–49.

Cox, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2):441–444.

Cubon, V. A., Putukian, M., Boyer, C., and Dettwiler, A. (2011). A diffusion tensor imaging study on the white matter skeleton in individuals with sports-related concussion. *Journal of Neurotrauma*, 28(2):189–201.

Culig, Z., Steiner, H., Bartsch, G., and Hobisch, A. (2005). Interleukin-6 regulation of prostate cancer cell growth. *Journal of Cellular Biochemistry*, 95(3):497–505.

Danaher, P., Wang, P., and Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B*, 76(2):373–397.

Davis, R. A., Zang, P., and Zheng, T. (2016). Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics*, 25(4):1077–1096.

de la Fuente, A., Bing, N., Hoeschele, I., and Mendes, P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, 20(18):3565–3574.

Dehan, P., Waltregny, D., Beschin, A., Noel, A., Castronovo, V., Tryggvason, K., De Leval, J., and Foidart, J. M. (1997). Loss of type IV collagen alpha 5 and alpha 6 chains in human invasive prostate carcinomas. *The American Journal of Pathology*, 151(4):1097–1104.

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015). High-dimensional inference: Confidence intervals, p-values and R-software `hdi`. *Statistical Science*, 30(4):533–558.

Dicker, L. H. (2014). Variance estimation in high-dimensional linear models. *Biometrika*, 101(2):269–284.

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., and West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90(1):196–212.

Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. In *AMS Conference on Mathematical Challenges of the 21st Century*.

Dossal, C. (2012). A necessary and sufficient condition for exact sparse recovery by $\ell 1$ minimization. *Comptes Rendus Mathematique*, 350(1-2):117–120.

Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, New York, NY.

Drton, M. and Maathuis, M. H. (2017). Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4:365–393.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

Erdős, P. and Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae*, 6:290–297.

Fan, F., Lu, H., and Liu, H. (2015). A direct estimation of high dimensional stationary vector autoregressions. *Journal of Machine Learning Research*, 16:3115–3150.

Fan, J., Feng, Y., and Wu, Y. (2009). Network exploration via the adaptive LASSO and SCAD penalties. *The Annals of Applied Statistics*, 3(2):521–541.

Fan, J., Guo, S., and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B*, 74(1):37–65.

Fan, J., Han, F., and Liu, H. (2014a). Challenges of Big Data analysis. *National Science Review*, 1(2):293–314.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. In Sanz-Solé, M., Soria, J., Varona, J. L., and Verdera, J., editors, *Proceedings of the International Congress of Mathematicians*, volume 3, pages 595–622, Zürich, CH. European Mathematical Society.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101–148.

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory*, 57(8):5467–5484.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961.

Fan, J., Xue, L., and Zou, H. (2014b). Strong oracle optimality of folded concave penalized estimation. *The Annals of Statistics*, 42(3):819–849.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22.

Friston, K. (2009). Causal modelling and brain connectivity in functional magnetic resonance imaging. *PLOS ONE*, 7(2):e1000033.

Fu, W. and Knight, K. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5):1356–1378.

Fu, W. J. (1998). Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416.

Fuchs, J. J. (2005). Recovery of exact sparse representations in the presence of noise. *IEEE Transactions on Information Theory*, 51(10):3601–3608.

Fukuyama, J., McMurdie, P. J., Dethlefsen, L., Relman, D. A., and Holmes, S. (2012). Comparisons of distance methods for combining covariates and abundances in microbiome studies. In *Pacific Symposium on Biocomputing*, volume 17, pages 213–224.

Gajawelli, N., Lao, Y., Apuzzo, M. L. J., Romano, R., Liu, C., Tsao, S., Hwang, D., Wilkins, B., Lepore, N., and Law, M. (2013). Neuroimaging changes in the brain in contact versus noncontact sport athletes using diffusion tensor imaging. *World Neurosurgery*, 80(6):824–828.

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy – analysis of *Affymetrix GeneChip* data at the probe level. *Bioinformatics*, 20(3):307–315.

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press, Cambridge, UK.

Gilbert, E. N. (1959). Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144.

Greenshtein, E. (2006). Best subset selection, persistence in high-dimensional statistical learning and optimization under l_1 constraint. *The Annals of Statistics*, 34(5):2367–2386.

Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15.

Guo, S., Wang, Y., and Yao, Q. (2016). High-dimensional and banded vector autoregressions. *Biometrika*, 103(4):889–903.

Haemers, W. H. (1995). Interlacing eigenvalues and graphs. *Linear Algebra and its Applications*, 226-228:593–616.

Halkidou, K., Gaughan, L., Cook, S., Leung, H. Y., Neal, D. E., and Robson, C. N. (2004). Upregulation and nuclear recruitment of HDAC1 in hormone refractory prostate cancer. *The Prostate*, 59(2):177–189.

Hall, C. L., Dubyk, C. W., Riesenberger, T. A., Shein, D., Keller, E. T., and van Golen, K. L. (2008). Type I collagen receptor ($\alpha_2\beta_1$) signaling promotes prostate cancer invasion through RhoC GTPase. *Neoplasia*, 10(8):797–803.

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, 96(4):835–845.

Hans, C. (2010). Model uncertainty and variable selection in Bayesian lasso regression. *Statistics and Computing*, 20(2):221–229.

Hart, G. T., Ramani, A. K., and Marcotte, E. M. (2006). How complete are current yeast and human protein-interaction networks? *Genome Biology*, 7(11):120.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer-Verlag New York, New York, NY.

Hoerl, A. E. and Kennard, R. W. (1970a). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12(1):69–82.

Hoerl, A. E. and Kennard, R. W. (1970b). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Horvath, S. and Dong, J. (2008). Geometric interpretation of gene co-expression network analysis. *PLOS Computational Biology*, 4(8):e1000117.

Horvath, S., Zhang, B., Carlson, M., Lu, K. V., Zhu, S., M., F. R., Laurance, M. F., Zhao, W., Qi, S., Chen, Z., Lee, Y., Scheck, A. C., Liau, L. M., Wu, H., Geschwind, D. H., Febbo, P. G., Kornblum, H. I., Cloughesy, T. F., Nelson, S. F., and Mischel, P. S. (2006). Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. *Proceedings of the National Academy of Sciences*, 103(46):17402–17407.

Hsieh, C.-J., Dhillon, I. S., Ravikumar, P. K., and Sustik, M. A. (2011). Sparse inverse covariance matrix estimation using quadratic approximation. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F. C. N., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 24, pages 2330–2338. Curran Associates, Inc., Red Hook, NY.

Huang, J., Ma, S., Li, H., and Zhang, C.-H. (2011). The sparse Laplacian shrinkage estimator for high-dimensional regression. *The Annals of Statistics*, 39(4):2021–2046.

Huang, J. Z., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98.

Huang, Y., Liu, X., Wang, Y.-H., Yeh, S.-D., Chen, C.-L., Nelson, R., Chu, P., Wilson, T., and Yen, Y. (2014). The prognostic value of ribonucleotide reductase small subunit M2 in predicting recurrence for prostate cancers. *Urologic Oncology: Seminars and Original Investigations*, 32(1):51.e9–51.e19.

Hunter, D. R. and Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37.

Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms. *The Annals of Statistics*, 33(4):1617–1642.

Ideker, T. and Krogan, N. J. (2012). Differential network biology. *Molecular Systems Biology*, 8(1):565.

Inoue, K., Slaton, J. W., Eve, B. Y., Kim, S. J., Perrotte, P., Balbay, M. D., Yano, S., Bar-Eli, M., Radinsky, R., Pettaway, C. A., and Dinney, C. P. N. (2000). Interleukin 8 expression regulates tumorigenicity and metastases in androgen-independent prostate cancer. *Clinical Cancer Research*, 6(5):2104–2119.

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D., and Lin, X. (2013). Sequence kernel association tests for the combined effect of rare and common variants. *American Journal of Human Genetics*, 92(6):841–853.

Irfanoglu, M. O., Modi, P., Nayak, A., Hutchinson, E. B., Sarlls, J., and Pierpaoli, C. (2015). *DR-BUDDI* (Diffeomorphic Registration for Blip-Up blip-Down Diffusion Imaging) method for correcting echo planar imaging distortions. *NeuroImage*, 106:284–299.

Jacob, L., Obozinski, G., and Vert, J.-P. (2009). Group lasso with overlap and graph lasso. In Danyluk, A. P., Bottou, L., and Littman, M. L., editors, *Proceedings of the 26th International Conference on Machine Learning*, pages 433–440, Brookline, MA. Microtome Publishing.

Janková, J. and van de Geer, S. (2015). Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics*, 9(1):1205–1229.

Janková, J. and van de Geer, S. (2017). Honest confidence regions and optimality in high-dimensional precision matrix estimation. *TEST*, 26(1):143–162.

Javanmard, A. and Montanari, A. (2013a). Confidence intervals and hypothesis testing for high-dimensional statistical models. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 26, pages 1187–1195. Curran Associates, Inc., Red Hook, NY.

Javanmard, A. and Montanari, A. (2013b). Nearly optimal sample size in hypothesis testing for high-dimensional regression. In *51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1427–1434.

Javanmard, A. and Montanari, A. (2014a). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909.

Javanmard, A. and Montanari, A. (2014b). Hypothesis testing in high-dimensional regression under the Gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory*, 60(10):6522–6554.

Jiang, D., Jiang, T., and Yang, F. (2012). Likelihood ratio tests for covariance matrices of high-dimensional normal distributions. *Journal of Statistical Planning and Inference*, 142(8):2241–2256.

Jin, T., Wang, Y., Li, G., Du, S., Yang, H., Geng, T., Hou, P., and Gong, Y. (2015). Analysis of difference of association between polymorphisms in the XRCC5, RPA3 and RTEL1 genes and glioma, astrocytoma and glioblastoma. *American Journal of Cancer Research*, 5(7):2294–2300.

Johnstone, I. M. and Titterington, D. M. (2009). Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4237–4253.

Kabaila, P. (1998). Valid confidence intervals in regression after variable selection. *Econometric Theory*, 14(4):463–482.

Kabaila, P. (2009). The coverage properties of confidence regions after model selection. *International Statistical Review*, 77(3):405–414.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30.

Karush, W. (1939). Minima of functions of several variables with inequalities as side constraints. Master's thesis, University of Chicago, Chicago, IL.

Ke, X.-S., Qu, Y., Rostad, K., Li, W.-C., Lin, B., Halvorsen, O. J., Haukaas, S., Jonassen, I., Petersen, K., Goldfinger, N., Rotter, V., Akslen, L. A., Oyan, A. M., and Kalland, K.-H. (2009). Genome-wide profiling of histone h3 lysine 4 and lysine 27 trimethylation reveals an epigenetic signature in prostate carcinogenesis. *PLOS ONE*, 4(3):e4687.

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. *PLOS Computational Biology*, 8(2):e1002375.

Kim, Y., Choi, H., and Oh, H.-S. (2008). Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, 103(484):1665–1673.

Koay, C. G., Chang, L.-C., Carew, J. D., Pierpaoli, C., and Basser, P. J. (2006). A unifying theoretical and algorithmic framework for least squares methods of estimation in diffusion tensor imaging. *Journal of Magnetic Resonance*, 182(1):115–125.

Koay, C. G., Özarslan, E., and Basser, P. J. (2009). A signal transformational framework for breaking the noise floor and its applications in MRI. *Journal of Magnetic Resonance*, 197(2):108–119.

Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer Series in Statistics. Springer Science+Business Media, Berlin, DE.

Koltchinskii, V. (2009). The Dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15(3):799–828.

Kong, S. W., Pu, W. T., and Park, P. J. (2006). A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics*, 22(19):2373–2380.

Krumsiek, J., Suhre, K., Illig, T., Adamski, J., and Theis, F. J. (2011). Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Systems Biology*, 5:21.

Kuhn, H. W. and Tucker, A. W. (1951). Nonlinear programming. In Neyman, J., editor, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 481–492, Berkeley, CA. University of California Press.

Kwee, L. C., Liu, D., Lin, X., Ghosh, D., and Epstein, M. P. (2007). A powerful and flexible multilocus association test for quantitative traits. *American Journal of Human Genetics*, 82(2):386–397.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics*, 37(6B):4254–4278.

Ledoit, O. and Wolf, M. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *The Annals of Statistics*, 30(4):1081–1102.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927.

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., NHLBI GO Exome Sequencing Project-ESP Lung Project Team, Christiani, D. C., Wurfel, M. M., and Lin, X. (2012a). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American Journal of Human Genetics*, 91(2):224–237.

Lee, S., Wu, M. C., and Lin, X. (2012b). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4):762–775.

Leeb, H. and Pötscher, B. M. (2003). The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econometric Theory*, 19(1):100–142.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, 21(1):21–59.

Leeb, H. and Pötscher, B. M. (2006a). Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, 34(5):2554–2591.

Leeb, H. and Pötscher, B. M. (2006b). Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk-bound results. *Econometric Theory*, 22(1):69–97.

Leeb, H. and Pötscher, B. M. (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, 24(2):338–376.

Leeb, H., Pötscher, B. M., and Ewald, K. (2015). On various confidence intervals post-model-selection. *Statistical Science*, 30(2):216–227.

Li, C. and Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182.

Li, C. and Li, H. (2010). Variable selection and regression analysis for graph-structured covariates with an application to genomics. *The Annals of Applied Statistics*, 4(3):1498–1516.

Li, J. and Chen, S. X. (2012). Two sample tests for high-dimensional covariance matrices. *The Annals of Statistics*, 40(2):908–940.

Liu, D., Lin, X., and Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics*, 63(4):1079–1088.

Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics*, 2:90–102.

Lozupone, C. and Knight, R. (2005). UniFrac: A new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, 71(12):8228–8235.

Lozupone, C. A., Hamady, M., Kelley, S. T., and Knight, R. (2007). Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology*, 73(5):1576–1585.

Mac Donald, C. L., Dikranian, K., Song, S. K., Bayly, P. V., Holtzman, D. M., and Brody, D. L. (2007). Detection of traumatic axonal injury with diffusion tensor imaging in a mouse model of traumatic brain injury. *Experimental Neurology*, 205(1):116–131.

Maini, A., Hillman, G., Haas, G. P., Wang, C. Y., Montecillo, E., Hamzavi, F., Pontes, E. J., Leland, P., Pastan, I., Debinski, W., and Puri, R. K. (1997). Interleukin-13 receptors on human prostate carcinoma cell lines represent a novel target for a chimeric protein composed of IL-13 and a mutated form of Pseudomonas exotoxin. *The Journal of Urology*, 158(3):948–953.

Mangin, J.-F., Poupon, C., Clark, C., Le Bihan, D., and Bloch, I. (2002). Distortion correction and robust tensor estimation for MR diffusion imaging. *Medical Image Analysis*, 6(3):191–198.

Mazumder, R., Friedman, J. H., and Hastie, T. (2011). *SparseNet*: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495):1125–1138.

Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B*, 72(4):417–473.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). p-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246–270.

Melnyk, I. and Banerjee, A. (2016). Estimating structured vector autoregressive models. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of the 33rd International Conference on Machine Learning*, pages 830–839, Brookline, MA. Microtome Publishing.

Michailidis, G. (2012). Statistical challenges in biological networks. *Journal of Computational and Graphical Statistics*, 21(4):840–855.

Michailidis, G. and d'Alché Buc, F. (2013). Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues. *Mathematical Biosciences*, 246(2):326–334.

Mohan, K., London, P., Fazel, M., Witten, D., and Lee, S.-I. (2014). Node-based learning of multiple Gaussian graphical models. *Journal of Machine Learning Research*, 15:445–488.

Mori, S., van Zijl, P. C. M., Oishi, K., and Faria, A. V. (2010). *MRI Atlas of Human White Matter*. Academic Press, Cambridge, MA.

Neuhaus, E. M., Zhang, W., Gelis, L., Deng, Y., Noldus, J., and Hatt, H. (2009). Activation of an olfactory receptor inhibits proliferation of prostate cancer cells. *The Journal of Biological Chemistry*, 284(24):16218–16225.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 45(2):167–256.

Ning, Y. and Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1):158–195.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27(1):29–34.

Oishi, K., Faria, A., Jiang, H., Li, X., Akhter, K., Zhang, J., Hsu, J. T., Miller, M. I., van Zijl, P. C. M., Albert, M., Lyketsos, C. G., Woods, R., Toga, A. W., Pike, G. B., Rosa-Neto, P., Evans, A., Mazziotta, J., and Mori, S. (2009). Atlas-based whole brain white matter analysis using large deformation diffeomorphic metric mapping: application to normal elderly and Alzheimer's disease participants. *NeuroImage*, 46(2):486–499.

Osborne, M. R., Presnell, B., and Turlach, B. A. (2000). On the LASSO and its dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337.

Pan, W., Xie, B., and Shen, X. (2010). Incorporating predictor network in penalized regression with application to microarray data. *Biometrics*, 66(2):474–484.

Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746.

Peterson, C., Stingo, F. C., and Vannucci, M. (2015). Bayesian inference of multiple Gaussian graphical models. *Journal of the American Statistical Association*, 110(509):159–174.

Pierpaoli, C., Walker, L., Irfanoglu, M. O., Barnett, A., Basser, P., Chang, L.-C., Koay, C., Pajevic, S., Rohde, G., Sarlls, J., and Wu, M. (2010). TORTOISE: An integrated software package for processing of diffusion MRI data. In *Joint Annual Meeting ISMRM-ESMRMB 2010*.

Pötscher, B. M. (1991). Effects of model selection on inference. *Econometric Theory*, 7(2):163–185.

Purdom, E. (2011). Analysis of a data matrix and a graph: Metagenomic data and the phylogenetic tree. *The Annals of Applied Statistics*, 5(4):2326–2358.

Qiu, H., Xu, S., Han, F., Liu, H., and Caffo, B. (2015). Robust estimation of transition matrices in high dimensional heavy-tailed vector autoregressive processes. In Blei, D. and Bach, F., editors, *Proceedings of the 32nd International Conference on Machine Learning*, pages 1843–1851, Brookline, MA. Microtome Publishing.

Randolph, T. W., Harezlak, J., and Feng, Z. (2012). Structured penalties for functional linear models – partially empirical eigenvectors for regression. *Electronic Journal of Statistics*, 6:323–353.

Raskutti, G., Wainwright, M. J., and Yu, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing $\ell 1$-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980.

Reid, S., Tibshirani, R., and Friedman, J. (2016). A study of error variance estimation in Lasso regression. *Statistica Sinica*, 26(1):35–67.

Ren, Z., Sun, T., Zhang, C.-H., and Zhou, H. H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *The Annals of Statistics*, 43(3):991–1026.

Rohde, G. K., Barnett, A. S., Basser, P. J., and Pierpaoli, C. (2005). Estimating intensity variance due to noise in registered images: Applications to diffusion tensor MRI. *NeuroImage*, 26(3):673–684.

Rosset, S. and Zhu, J. (2007). Piecewise lienar regularized solution paths. *The Annals of Statistics*, 35(3):1012–1030.

Rosset, S., Zhu, J., and Hastie, T. (2004). Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5:941–973.

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515.

Rudelson, M. and Zhou, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory*, 59(6):3434–3447.

Saegusa, T. and Shojaie, A. (2016). Joint estimation of precision matrices in heterogenous populations. *Electronic Journal of Statistics*, 10(1):1341–1392.

Schott, J. R. (2007). A test for the equality of covariance matrices when the dimension is large relative to the sample sizes. *Computational Statistics & Data Analysis*, 51(12):6535–6542.

Seligson, D. B., Horvath, S., Shi, T., Yu, H., Tze, S., Grunstein, M., and Kurdistani, S. K. (2005). Global histone modification patterns predict risk of prostate cancer recurrence. *Nature*, 435(7046):1262–1266.

Seth, A. K., Chorley, P., and Barnett, L. C. (2013). Granger causality analysis of fMRI BOLD signals is invariant to hemodynamic convolution but not downsampling. *NeuroImage*, 65:540–555.

Shah, R. D. and Samworth, R. J. (2013). Variable selection with error control: Another look at stability selection. *Journal of the Royal Statistical Society: Series B*, 75(1):55–80.

Shao, J. and Deng, X. (2012). Estimation in high-dimensional linear models with deterministic design matrices. *The Annals of Statistics*, 40(2):812–831.

Shen, X., Huang, H.-C., and Pan, W. (2012). Simultaneous supervised clustering and feature selection over a graph. *Biometrika*, 99(4):899–914.

Shojaie, A. and Michailidis, G. (2009). Analysis of gene sets based on the underlying regulatory network. *Journal of Computational Biology*, 16(3):407–426.

Shojaie, A. and Michailidis, G. (2010a). Discovering graphical Granger causality using the truncating lasso penalty. *Bioinformatics*, 26(18):i517–i523.

Shojaie, A. and Michailidis, G. (2010b). Network enrichment analysis in complex experiments. *Statistical Applications in Genetics and Molecular Biology*, 9(1):Article 22.

Shojaie, A. and Michailidis, G. (2010c). Penalized principal component regression on graphs for analysis of subnetworks. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, volume 23, pages 2155–2163. Curran Associates, Inc., Red Hook, NY.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 48(1):1–48.

Slawski, M., zu Castell, W., and Tutz, G. (2010). Feature selection guided by structural information. *The Annals of Applied Statistics*, 4(2):1056–1080.

Smith, S. M. (2012). The future of FMRI connectivity. *NeuroImage*, 62(2):1257–1266.

Srivastava, M. S. (2005). Some tests concerning the covariance matrix in high dimensional data. *Journal of the Japan Statistical Society*, 35(2):251–272.

Srivastava, M. S. and Yanagihara, H. (2010). Testing the equality of several covariance matrices with fewer observations than the dimension. *Journal of Multivariate Analysis*, 101(6):1319–1329.

Stock, J. H. and Watson, M. W. (2001). Vector autoregressions. *Journal of Economic Perspectives*, 15(4):101–115.

Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika*, 99(4):879–898.

Sun, T. and Zhang, C.-H. (2013). Sparse matrix inversion with scaled Lasso. *Journal of Machine Learning Research*, 14:3385–3418.

Taylor, J. and Tibshirani, R. J. (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society: Series B*, 73(3):273–282.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B*, 67(1):91–108.

Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490.

Tibshirani, R. J. and Taylor, J. (2012). Degree of freedom in lasso problems. *The Annals of Statistics*, 40(2):1198–1232.

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620.

Tropp, J. A. (2006). Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52(3):1030–1051.

Untergasser, G., Plas, E., Pfister, G., Heinrich, E., and Berger, P. (2005). Interferon-γ induces neuroendocrine-like differentiation of human prostate basal-epithelial cells. *The Prostate*, 64(4):419–429.

Vaezi, A. E., Bepler, G., Bhagwat, N. R., Malysa, A., Rubatt, J. M., Chen, W., Hood, B. L., Conrads, T. P., Wang, L., Kemp, C. E., and Niedernhofer, L. J. (2014). Choline phosphate cytidylyltransferase-α is a novel antigen detected by the anti-ERCC1 antibody 8F1

with biomarker value in patients with lung and head and neck squamous cell carcinomas. *Cancer*, 120(12):1898–1907.

van de Geer, S. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

Vardabasso, C., Hasson, D., Ratnakumar, K., Chung, C.-Y., Duarte, L. F., and Bernstein, E. (2014). Histone variants: Emerging players in cancer biology. *Cellular and Molecular Life Sciences*, 71(3):379–404.

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In Eldar, Y. C. and Kutyniok, G., editors, *Compressed Sensing: Theory and Applications*, chapter 5. Cambridge University Press, Cambridge, UK.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programmming (Lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202.

Wang, L., Kim, Y., and Li, R. (2013). Calibrating nonconvex penalized regression in ultra-high dimension. *The Annals of Statistics*, 41(5):2505–2536.

Wang, Z., Liu, H., and Zhang, T. (2014). Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *The Annals of Statistics*, 42(6):2164–2201.

Wasserman, L. and Roeder, K. (2009). High-dimensional variable selection. *The Annals of Statistics*, 37(5A):2178–2201.

Wei, P. and Pan, W. (2008). Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model. *Bioinformatics*, 24(3):404–411.

Weinstein, A., Fithian, W., and Benjamini, Y. (2013). Selection adjusted confidence intervals with more power to determine the sign. *Journal of the American Statistical Association*, 108(501):165–176.

Weisberg, S. (2013). *Applied Lienar Regression*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ.

Wolf, Y. I., Karev, G., and Koonin, E. V. (2002). Scale-free networks in biology: New insights into the fundamentals of evolution? *BioEssays*, 24(2):105–109.

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., and Lin, X. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. *American Journal of Human Genetics*, 86(6):929–942.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare variant association testing for sequencing data using the sequence kernel association test. *American Journal of Human Genetics*, 89(1):82–93.

Wu, T. T. and Lange, K. (2008). Coordinate descent algorithm for lasso penalized regression. *The Annals of Applied Statistics*, 2(1):224–244.

Xia, Y., Cai, T., and Cai, T. T. (2015). Testing differential networks with applications to detecting gene-by-gene interactions. *Biometrika*, 102(2):247–266.

Xia, Y. and Li, L. (2017). Hypothesis testing of matrix graph model with application to brain connectivity analysis. *Biometrics*, to appear.

Yang, S., Yuan, L., Lai, Y.-C., Shen, X., Wonka, P., and Ye, J. (2012). Feature grouping and selection over an undirected graph. In *Proceedings of the 18th ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining, pages 922–930, New York, NY. Association for Computing Machinery.

Ye, F. and Zhang, C.-H. (2010). Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *Journal of Machine Learning Research*, 11:3519–3540.

You, H., Lin, H., and Zhang, Z. (2015). CKS2 in human cancers: Clinical roles and current perspectives (Review). *Molecular and Clinical Oncology*, 3(3):459–463.

Yu, Z., Chen, T., Hébert, J., Li, E., and Richard, S. (2009). A mouse *PRMT1* null allele defines an essential role for arginine methylation in genome maintenance and cell proliferation. *Molecular and Cellular Biology*, 29(11):2982–2996.

Yuan, M. (2008). Efficient computation of ℓ_1 regularized estimates in Gaussian graphical models. *Journal of Computational and Graphical Statistics*, 17(4):809–826.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11:2261–2286.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35.

Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1):Article 17.

Zhang, C.-H. (2010a). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B*, 76(1):217–242.

Zhang, C.-H. and Zhang, T. (2012). A general theory of concave regularization for high dimensional sparse estimation problems. *Statistical Science*, 27(4):576–593.

Zhang, Q., Liu, S., Ge, D., Zhang, Q., Xue, Y., Xiong, Z., Abdel-Mageed, A. B., Myers, L., Hill, S. M., Rowan, B. G., Sartor, O., Melamed, J., Chen, Z., and You, Z. (2012). Interleukin-17 promotes formation and growth of prostate adenocarcinoma in mouse models. *Cancer Research*, 72(10):2589–2599.

Zhang, T. (2009). Some sharp performance bounds for least squares regression with l_1 regularization. *The Annals of Statistics*, 37(5A):2109–2144.

Zhang, T. (2010b). Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11:1081–1107.

Zhang, T. (2013). Multi-stage convex relaxation for feature selection. *Bernoulli*, 19(5B):2277–2293.

Zhang, Y., Zhang, J., Oishi, K., Faria, A. V., Jiang, H., Li, X., Akhter, K., Rosa-Neto, P., Pike, G. B., Evans, A., Toga, A. W., Woods, R., Mazziotta, J. C., Miller, M. I., van Zijl, P. C. M., and Mori, S. (2010). Atlas-guided tract reconstruction for automated and comprehensive examination of the white matter anatomy. *NeuroImage*, 52(4):1289–1301.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563.

Zhao, S. and Shojaie, A. (2016). A significance test for graph-constrained estimation. *Biometrics*, 72(2):484–493.

Zhao, S. D., Cai, T. T., and Li, H. (2014). Direct estimation of differential networks. *Biometrika*, 101(2):253–268.

Zhu, X., Gerstein, M., and Snyder, M. (2007). Getting connected: Analysis and principles of biological networks. *Genes & Development*, 21(9):1010–1024.

Zhu, Y., Shen, X., and Pan, W. (2013). Simultaneous grouping pursuit and feature selection over an undirected graph. *Journal of the American Statistical Association*, 108(502):713–725.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

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

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35(5):2173–2192.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36(4):1509–1533.

Appendix A

TECHNICAL DETAILS FOR "IN DEFENSE OF THE INDEFENSIBLE: A VERY NAÏVE APPROACH TO HIGH-DIMENSIONAL INFERENCE"

A.1 Proof of Proposition 2.2.4

We first state and prove Lemmas A.1.1–A.1.5, which are required to prove Proposition 2.2.4.

Lemma A.1.1. *Suppose (***M1***) holds. Then, $\check{\boldsymbol{\beta}}_\lambda$ and $\hat{\boldsymbol{\beta}}_\lambda$ as defined in* (1.6) *and* (1.2)*, respectively, are unique.*

Proof. First, by, e.g., Lemma 3 of Tibshirani (2013), (**M1**) implies that $\hat{\boldsymbol{\beta}}_\lambda$ is unique. We now prove $\check{\boldsymbol{\beta}}_\lambda$ is also unique. The proof is similar to the proof of Lemmas 1 and 3 in Tibshirani (2013).

To show $\check{\boldsymbol{\beta}}_\lambda$ is unique, we first show that the fitted value of $\boldsymbol{X}\check{\boldsymbol{\beta}}_\lambda$ is unique. This is because, suppose, to the contrary, that we have two solutions to the problem (1.6), $\check{\boldsymbol{\beta}}^{\mathrm{I}}_\lambda$ and $\check{\boldsymbol{\beta}}^{\mathrm{II}}_\lambda$, which give different fitted values, $\boldsymbol{X}\check{\boldsymbol{\beta}}^{\mathrm{I}}_\lambda \neq \boldsymbol{X}\check{\boldsymbol{\beta}}^{\mathrm{II}}_\lambda$, but achieve the same minimum value of the objective function, c, i.e.,

$$\frac{1}{2n}\mathrm{E}\left[\left\|\boldsymbol{y}-\boldsymbol{X}\check{\boldsymbol{\beta}}^{\mathrm{I}}_\lambda\right\|_2^2\right]+\lambda\left\|\check{\boldsymbol{\beta}}^{\mathrm{I}}_\lambda\right\|_1=\frac{1}{2n}\mathrm{E}\left[\left\|\boldsymbol{y}-\boldsymbol{X}\check{\boldsymbol{\beta}}^{\mathrm{II}}_\lambda\right\|_2^2\right]+\lambda\left\|\check{\boldsymbol{\beta}}^{\mathrm{II}}_\lambda\right\|_1=c. \tag{A.1}$$

Then the value of the objective function of $\check{\boldsymbol{\beta}}^{\mathrm{I}}_\lambda/2+\check{\boldsymbol{\beta}}^{\mathrm{II}}_\lambda/2$ is

$$\begin{aligned}&\frac{1}{2n}\mathrm{E}\left[\left\|\boldsymbol{y}-\boldsymbol{X}\left(\frac{1}{2}\check{\boldsymbol{\beta}}^{\mathrm{I}}_\lambda+\frac{1}{2}\check{\boldsymbol{\beta}}^{\mathrm{II}}_\lambda\right)\right\|_2^2\right]+\lambda\left\|\frac{1}{2}\check{\boldsymbol{\beta}}^{\mathrm{I}}_\lambda+\frac{1}{2}\check{\boldsymbol{\beta}}^{\mathrm{II}}_\lambda\right\|_1\\ &<\frac{1}{4n}\mathrm{E}\left[\left\|\boldsymbol{y}-\boldsymbol{X}\check{\boldsymbol{\beta}}^{\mathrm{I}}_\lambda\right\|_2^2\right]+\frac{\lambda}{2}\left\|\check{\boldsymbol{\beta}}^{\mathrm{I}}_\lambda\right\|_1+\frac{1}{4n}\mathrm{E}\left[\left\|\boldsymbol{y}-\boldsymbol{X}\check{\boldsymbol{\beta}}^{\mathrm{II}}_\lambda\right\|_2^2\right]+\frac{\lambda}{2}\left\|\check{\boldsymbol{\beta}}^{\mathrm{II}}_\lambda\right\|_1=c.\end{aligned} \tag{A.2}$$

The inequality is due to the strict convexity of the squared ℓ_2 norm function and the convexity of the ℓ_1 norm function. Thus, $\check{\boldsymbol{\beta}}^{\mathrm{I}}_\lambda/2+\check{\boldsymbol{\beta}}^{\mathrm{II}}_\lambda/2$ achieves a smaller value of the objective function

than either $\check{\boldsymbol{\beta}}_\lambda^{\mathrm{I}}$ or $\check{\boldsymbol{\beta}}_\lambda^{\mathrm{II}}$, which is a contradiction. Hence, all solutions to the problem (1.6) have the same fitted value.

Therefore, based on the stationary condition in (2.12),

$$\lambda n \check{\boldsymbol{\tau}}_\lambda = \boldsymbol{X}^\top \boldsymbol{X} \left(\boldsymbol{\beta} - \check{\boldsymbol{\beta}}_\lambda \right), \tag{A.3}$$

$\check{\boldsymbol{\tau}}_\lambda$ is unique. Define $\check{\mathcal{T}}_\lambda \equiv \{j : |\check{\tau}_{\lambda,j}| = 1\}$. $\check{\mathcal{T}}_\lambda$ is also unique. Furthermore, because $|\check{\boldsymbol{\tau}}_{\lambda,\check{\mathcal{A}}_\lambda}| = \mathbf{1}$, $\check{\mathcal{T}}_\lambda \supseteq \check{\mathcal{A}}_\lambda$, and

$$\check{\boldsymbol{\beta}}_{\lambda,\check{\mathcal{T}}_\lambda^c} = \mathbf{0}. \tag{A.4}$$

Also according to the stationary condition in (2.12), and using the fact that $\mathrm{E}[\boldsymbol{y}] = \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y} - \boldsymbol{\epsilon}$, we have

$$\lambda n \check{\boldsymbol{\tau}}_{\lambda,\check{\mathcal{T}}_\lambda} = \boldsymbol{X}_{\check{\mathcal{T}}_\lambda}^\top \left(\boldsymbol{y} - \boldsymbol{X}\check{\boldsymbol{\beta}}_\lambda - \boldsymbol{\epsilon} \right) = \boldsymbol{X}_{\check{\mathcal{T}}_\lambda}^\top \left(\boldsymbol{y} - \boldsymbol{X}_{\check{\mathcal{T}}_\lambda} \check{\boldsymbol{\beta}}_{\lambda,\check{\mathcal{T}}_\lambda} - \boldsymbol{\epsilon} \right). \tag{A.5}$$

The last equality holds because as shown in (A.4), $\check{\boldsymbol{\beta}}_{\lambda,\check{\mathcal{T}}_\lambda^c} = \mathbf{0}$. Equation (A.5) indicates that $\lambda n \check{\boldsymbol{\tau}}_{\lambda,\check{\mathcal{T}}_\lambda}$ is in the row space of $\boldsymbol{X}_{\check{\mathcal{T}}_\lambda}$, or $\lambda n \check{\boldsymbol{\tau}}_{\lambda,\check{\mathcal{T}}_\lambda} = \boldsymbol{X}_{\check{\mathcal{T}}_\lambda}^\top (\boldsymbol{X}_{\check{\mathcal{T}}_\lambda}^\top)^+ \lambda n \check{\boldsymbol{\tau}}_{\lambda,\check{\mathcal{T}}_\lambda}$, where $(\boldsymbol{X}_{\check{\mathcal{T}}_\lambda}^\top)^+$ is the Moorse-Penrose pseudoinverse of $\boldsymbol{X}_{\check{\mathcal{T}}_\lambda}^\top$. Properties of the Moorse-Penrose pseudoinverse include $(\boldsymbol{X}_{\check{\mathcal{T}}_\lambda}^\top)^+ = (\boldsymbol{X}_{\check{\mathcal{T}}_\lambda}^+)^\top$, $\boldsymbol{X}_{\check{\mathcal{T}}_\lambda}^+ = (\boldsymbol{X}_{\check{\mathcal{T}}_\lambda}^\top \boldsymbol{X}_{\check{\mathcal{T}}_\lambda})^+ \boldsymbol{X}_{\check{\mathcal{T}}_\lambda}^\top$ and $\boldsymbol{X}_{\check{\mathcal{T}}_\lambda} = \boldsymbol{X}_{\check{\mathcal{T}}_\lambda} \boldsymbol{X}_{\check{\mathcal{T}}_\lambda}^+ \boldsymbol{X}_{\check{\mathcal{T}}_\lambda}$.

Rearranging terms in (A.5), and pluging in $\lambda n \check{\boldsymbol{\tau}}_{\lambda,\check{\mathcal{T}}_\lambda} = \boldsymbol{X}_{\check{\mathcal{T}}_\lambda}^\top (\boldsymbol{X}_{\check{\mathcal{T}}_\lambda}^\top)^+ \lambda n \check{\boldsymbol{\tau}}_{\lambda,\check{\mathcal{T}}_\lambda}$, we get

$$\begin{aligned} \boldsymbol{X}_{\check{\mathcal{T}}_\lambda}^\top \boldsymbol{X}_{\check{\mathcal{T}}_\lambda} \check{\boldsymbol{\beta}}_{\lambda,\check{\mathcal{T}}_\lambda} &= \boldsymbol{X}_{\check{\mathcal{T}}_\lambda}^\top (\boldsymbol{y} - \boldsymbol{\epsilon}) - \lambda n \check{\boldsymbol{\tau}}_{\lambda,\check{\mathcal{T}}_\lambda} \\ &= \boldsymbol{X}_{\check{\mathcal{T}}_\lambda}^\top \left(\boldsymbol{y} - \boldsymbol{\epsilon} - \left(\boldsymbol{X}_{\check{\mathcal{T}}_\lambda}^\top \right)^+ \lambda n \check{\boldsymbol{\tau}}_{\lambda,\check{\mathcal{T}}_\lambda} \right). \end{aligned} \tag{A.6}$$

Hence, according to (A.6) and the properties of the Moorse-Penrose pseudoinverse,

$$\begin{aligned} \boldsymbol{X}_{\check{\mathcal{T}}_\lambda} \check{\boldsymbol{\beta}}_{\lambda,\check{\mathcal{T}}_\lambda} &= \boldsymbol{X}_{\check{\mathcal{T}}_\lambda} \boldsymbol{X}_{\check{\mathcal{T}}_\lambda}^+ \boldsymbol{X}_{\check{\mathcal{T}}_\lambda} \check{\boldsymbol{\beta}}_{\lambda,\check{\mathcal{T}}_\lambda} \\ &= \boldsymbol{X}_{\check{\mathcal{T}}_\lambda} \left(\boldsymbol{X}_{\check{\mathcal{T}}_\lambda}^\top \boldsymbol{X}_{\check{\mathcal{T}}_\lambda} \right)^+ \boldsymbol{X}_{\check{\mathcal{T}}_\lambda}^\top \boldsymbol{X}_{\check{\mathcal{T}}_\lambda} \check{\boldsymbol{\beta}}_{\lambda,\check{\mathcal{T}}_\lambda} \\ &= \boldsymbol{X}_{\check{\mathcal{T}}_\lambda} \left(\boldsymbol{X}_{\check{\mathcal{T}}_\lambda}^\top \boldsymbol{X}_{\check{\mathcal{T}}_\lambda} \right)^+ \boldsymbol{X}_{\check{\mathcal{T}}_\lambda}^\top \left(\boldsymbol{y} - \boldsymbol{\epsilon} - \left(\boldsymbol{X}_{\check{\mathcal{T}}_\lambda}^\top \right)^+ \lambda n \check{\boldsymbol{\tau}}_{\lambda,\check{\mathcal{T}}_\lambda} \right) \\ &= \boldsymbol{X}_{\check{\mathcal{T}}_\lambda} \boldsymbol{X}_{\check{\mathcal{T}}_\lambda}^+ \left(\boldsymbol{y} - \boldsymbol{\epsilon} - \left(\boldsymbol{X}_{\check{\mathcal{T}}_\lambda}^\top \right)^+ \lambda n \check{\boldsymbol{\tau}}_{\lambda,\check{\mathcal{T}}_\lambda} \right). \end{aligned} \tag{A.7}$$

Thus,

$$\check{\boldsymbol{\beta}}_{\lambda,\check{\mathcal{T}}_\lambda} = \boldsymbol{X}^+_{\check{\mathcal{T}}_\lambda}\left(\boldsymbol{y} - \boldsymbol{\epsilon} - \left(\boldsymbol{X}^\top_{\check{\mathcal{T}}_\lambda}\right)^+ \lambda n \check{\boldsymbol{\tau}}_{\lambda,\check{\mathcal{T}}_\lambda}\right) + \boldsymbol{d}, \tag{A.8}$$

where $\boldsymbol{X}_{\check{\mathcal{T}}_\lambda}\boldsymbol{d} = \boldsymbol{0}$. Therefore, if $\text{null}(\boldsymbol{X}_{\check{\mathcal{T}}_\lambda}) = \{\boldsymbol{0}\}$, $\boldsymbol{d} = \boldsymbol{0}$. Because $\check{\boldsymbol{\tau}}_\lambda$ and $\check{\mathcal{T}}_\lambda$ are unique based on (A.3), $\text{null}(\boldsymbol{X}_{\check{\mathcal{T}}_\lambda}) = \{\boldsymbol{0}\}$ also implies $\check{\boldsymbol{\beta}}_\lambda$ is unique, i.e.,

$$\check{\boldsymbol{\beta}}_{\lambda,\check{\mathcal{T}}_\lambda} = \boldsymbol{X}^+_{\check{\mathcal{T}}_\lambda}\left(\boldsymbol{y} - \boldsymbol{\epsilon} - \left(\boldsymbol{X}^\top_{\check{\mathcal{T}}_\lambda}\right)^+ \lambda n \check{\boldsymbol{\tau}}_{\lambda,\check{\mathcal{T}}_\lambda}\right), \tag{A.9}$$

$$\check{\boldsymbol{\beta}}_{\lambda,\check{\mathcal{T}}^c_\lambda} = \boldsymbol{0}. \tag{A.10}$$

To see that (**M1**) implies $\text{null}(\boldsymbol{X}_{\check{\mathcal{T}}_\lambda}) = \{\boldsymbol{0}\}$, we use a similar argument as Tibshirani (2013). Specifically, we assume to the contrary, $\text{null}(\boldsymbol{X}_{\check{\mathcal{T}}_\lambda}) \neq \{\boldsymbol{0}\}$. Then, for any $j \in \check{\mathcal{T}}_\lambda$, we can write $\boldsymbol{X}_j = \sum_{k\in\check{\mathcal{T}}_\lambda\setminus\{j\}} a_k \boldsymbol{X}_k$, and multiplying both sides by $\check{\tau}_{\lambda,j}$,

$$\check{\tau}_{\lambda,j}\boldsymbol{X}_j = \sum_{k\in\check{\mathcal{T}}_\lambda\setminus\{j\}} \check{\tau}_{\lambda,j} a_k \boldsymbol{X}_k. \tag{A.11}$$

Multiplying both sides of (A.11) by $\boldsymbol{X}(\boldsymbol{\beta} - \check{\boldsymbol{\beta}}_\lambda)/n$,

$$\check{\tau}_{\lambda,j}\frac{1}{n}\boldsymbol{X}_j^\top \boldsymbol{X}\left(\boldsymbol{\beta} - \check{\boldsymbol{\beta}}_\lambda\right) = \sum_{k\in\check{\mathcal{T}}_\lambda\setminus\{j\}} \check{\tau}_{\lambda,j} a_k \frac{1}{n}\boldsymbol{X}_k^\top \boldsymbol{X}\left(\boldsymbol{\beta} - \check{\boldsymbol{\beta}}_\lambda\right). \tag{A.12}$$

Based on the stationary condition in (2.12),

$$\lambda\check{\tau}_{\lambda,j} = \frac{1}{n}\boldsymbol{X}_j^\top \boldsymbol{X}\left(\boldsymbol{\beta} - \check{\boldsymbol{\beta}}_\lambda\right), \tag{A.13}$$

(A.12) implies that $\check{\tau}^2_{\lambda,j} = \sum_{k\in\check{\mathcal{T}}_\lambda\setminus\{j\}} \check{\tau}_{\lambda,j} a_k \check{\tau}_{\lambda,k}$. Since $\check{\tau}^2_{\lambda,j} = 1$ for any $j \in \check{\mathcal{T}}_\lambda$,

$$\sum_{k\in\check{\mathcal{T}}_\lambda\setminus\{j\}} \check{\tau}_{\lambda,j} a_k \check{\tau}_{\lambda,k} = 1. \tag{A.14}$$

Therefore,

$$\begin{aligned}
\check{\tau}_{\lambda,j}\boldsymbol{X}_j &= \sum_{k\in\check{\mathcal{T}}_\lambda\setminus\{j\}} \check{\tau}_{\lambda,j} a_k \boldsymbol{X}_k \\
&= \sum_{k\in\check{\mathcal{T}}_\lambda\setminus\{j\}} \check{\tau}_{\lambda,j} a_k \check{\tau}^2_{\lambda,k} \boldsymbol{X}_k \\
&\equiv \sum_{k\in\check{\mathcal{T}}_\lambda\setminus\{j\}} c_k \check{\tau}_{\lambda,k} \boldsymbol{X}_k,
\end{aligned} \tag{A.15}$$

where $c_k \equiv \check{\tau}_{\lambda,j} a_k \check{\tau}_{\lambda,k}$. By (A.14), we have $\sum_{k \in \check{\mathcal{T}}_\lambda \setminus \{j\}} c_k = 1$. This shows that $\check{\tau}_{\lambda,j} \boldsymbol{X}_j$, $j \in \check{\mathcal{T}}_\lambda$, is a weighted average of $\check{\tau}_{\lambda,k} \boldsymbol{X}_k$, $k \in \check{\mathcal{T}}_\lambda \setminus \{j\}$, which contradicts (**M1**). Thus, (**M1**) implies that $\check{\boldsymbol{\beta}}_\lambda$ is unique. □

Lemma A.1.2. *Suppose (***M1***) and (***M2***) hold. Then,*

$$\frac{1}{n} \left\| \boldsymbol{X}^\top \boldsymbol{\epsilon} \right\|_\infty = \mathcal{O}_p \left(\sqrt{\frac{\log(p)}{n}} \right).$$

Proof. There is an equivalence between the tail probability of a sub-Gaussian random variable and its moment generating function. For example, according to Lemma 5.5 in Vershynin (2012), since $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$, for the constant $h > 0$ stated in (**M2**), there exists some $k > 0$ such that, $M_{\epsilon_i}(t) \equiv \mathrm{E}[\exp(t\epsilon_i)] \leq \exp(kt^2)$ for all $t \in \mathbb{R}$ if $\Pr[|\epsilon_i| \geq x] \leq \exp(1 - hx^2)$ for any $x > 0$, where $M_{\epsilon_i}(t)$ is the moment generating function of ϵ_i. Thus, for any $j = 1, \ldots, p$, denoting $T_j \equiv \sum_{i=1}^n X_{ij}\epsilon_i / \sqrt{n} = (\boldsymbol{X}^\top \boldsymbol{\epsilon})_j / \sqrt{n}$, we have

$$\begin{aligned} M_{T_j}(t) = \mathrm{E}\left[\exp\left(t \sum_{i=1}^n \frac{1}{\sqrt{n}} X_{ij}\epsilon_i \right) \right] &= \prod_{i=1}^n \mathrm{E}\left[\exp\left(\frac{1}{\sqrt{n}} X_{(i,j)}^2 t\epsilon_i \right) \right] \\ &\leq \prod_{i=1}^n \exp\left(\frac{X_{(i,j)}^2}{n} kt^2 \right) \\ &= \exp\left(\frac{\|\boldsymbol{X}_j\|_2^2}{n} kt^2 \right) = \exp\left(kt^2\right). \end{aligned} \tag{A.16}$$

The last equality holds because columns of $\boldsymbol{X}$ are standardized such that $\|\boldsymbol{X}_j\|_2^2 = n$ for $j = 1, \ldots, p$ by (**M1**). Using Chebyshev's inequality, (A.16) shows that for any $j = 1, \ldots, p$, we have $\Pr[|T_j| \geq x] \leq \exp(1 - h'x^2)$ for some $h' > 0$. Applying Boole's inequality,

$$\begin{aligned} \Pr\left[\frac{1}{\sqrt{n}} \left\| \boldsymbol{X}^\top \boldsymbol{\epsilon} \right\|_\infty \equiv \max_{j=1,\ldots,p} |T_j| > t\sqrt{\log(p)} \right] &= \Pr\left[\bigcup_{j=1,\ldots,p} \left\{ |T_j| > t\sqrt{\log(p)} \right\} \right] \\ &\leq \sum_{j=1}^p \Pr\left[|T_j| > t\sqrt{\log(p)} \right] \\ &\leq p \exp\left(1 - h't^2 \log(p)\right) \\ &= \exp\left(\log(p) \left(1 - h't^2 \log(p)\right) \right) \\ &\leq \exp\left(1 - h't^2\right). \end{aligned}$$

Sicne $\exp(\log(p)(1-h't^2\log(p)))$ is a decreasing function of p with $h't^2 > 1$ and $p \geq e$, the last inequality holds with $h't^2 > 1$ and $p \geq 3$. Thus, for any $\xi > 0$, we can choose a large value of t, such that $\Pr[\|\boldsymbol{X}^\top \boldsymbol{\epsilon}\|_\infty/\sqrt{n} > t\sqrt{\log(p)}] < \xi$. This shows that $\|\boldsymbol{X}^\top \boldsymbol{\epsilon}\|_\infty/\sqrt{n} = \mathcal{O}_p(\sqrt{\log(p)})$. Dividing both sides by $\sqrt{n}$ completes the proof. □

Lemma A.1.3. *Suppose (***E***) holds. Then, $\check{\mathcal{A}}_\lambda \supseteq \mathcal{S}$, where $\check{\mathcal{A}}_\lambda \equiv \mathrm{supp}(\check{\boldsymbol{\beta}}_\lambda)$ and $\mathcal{S} \equiv \{j : |\beta_j| > 3\lambda\sqrt{q}/\phi^2\}$.*

Proof. First, if $q \equiv |\mathcal{A}| = 0$, we trivially have $\check{\mathcal{A}}_\lambda \supseteq \mathcal{S} = \emptyset$.

If $q \geq 1$, by Corollary 2.1 in van de Geer and Bühlmann (2009), (**E**) guarantees that

$$\left\|\check{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\right\|_\infty \leq \left\|\check{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\right\|_2 \leq \frac{2\lambda\sqrt{2q}}{\phi^2}. \tag{A.17}$$

For any $j = 1, \ldots, p$ such that $j \in \mathcal{S}$, we have $|\beta_j| > 3\lambda\sqrt{q}/\phi^2$. Thus, by (A.17), $|\check{\beta}_{\lambda,j}| > (3 - 2\sqrt{2})\lambda\sqrt{q}/\phi^2 > 0$, i.e., $j \in \check{\mathcal{A}}_\lambda$, or, $\check{\mathcal{A}}_\lambda \supseteq \mathcal{S}$. □

Lemma A.1.4. *Suppose (***M1***)–(***M3***), (***E***) and (***T***) hold. Then, the estimator $\hat{\boldsymbol{\beta}}_\lambda$ defined in* (1.2) *satisfies $\|\hat{\boldsymbol{\beta}}_\lambda - \check{\boldsymbol{\beta}}_\lambda\|_1 = \mathcal{O}_p(\sqrt{\log(p)/n})$.*

Proof. Let $Q(\boldsymbol{b}) \equiv \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2/(2n) + \lambda\|\boldsymbol{b}\|_1$. Thus, $\hat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{b}\in\mathbb{R}^p} Q(\boldsymbol{b})$. Lemma A.1.1 shows that $\hat{\boldsymbol{\beta}}_\lambda$ and $\check{\boldsymbol{\beta}}_\lambda$ are unique, even with $p > n$. To prove Lemma A.1.4, we want to show that for all $\xi > 0$, there exists an $m \in \mathbb{R}$, not depending on n, such that for n sufficiently large,

$$\Pr\left[\inf_{\boldsymbol{a}:\|\boldsymbol{a}\|_1=m} Q\left(\check{\boldsymbol{\beta}}_\lambda + \boldsymbol{a}\sqrt{\frac{\log(p)}{n}}\right) > Q\left(\check{\boldsymbol{\beta}}_\lambda\right)\right] \geq 1 - \xi. \tag{A.18}$$

Since Q is convex, (A.18) implies that the minimizer of $Q(\boldsymbol{b})$, $\hat{\boldsymbol{\beta}}_\lambda$, lies in the convex region $\{\check{\boldsymbol{\beta}}_\lambda + \boldsymbol{a}\sqrt{\log(p)/n} : \|\boldsymbol{a}\|_1 < m\}$ with probability at least $1 - \xi$. Hence, for n sufficiently large,

$$\Pr\left[\left\|\hat{\boldsymbol{\beta}}_\lambda - \check{\boldsymbol{\beta}}_\lambda\right\|_1 \geq m\sqrt{\frac{\log(p)}{n}}\right] \leq \xi,$$

i.e., $\|\hat{\boldsymbol{\beta}}_\lambda - \check{\boldsymbol{\beta}}_\lambda\|_1 = \mathcal{O}_p(\sqrt{\log(p)/n})$.

We now proceed to prove (A.18). Let $\boldsymbol{w} \equiv \arg\min_{\boldsymbol{a}:\|\boldsymbol{a}\|_1=m} Q(\check{\boldsymbol{\beta}}_\lambda + \boldsymbol{a}\sqrt{\log(p)/n})$. We want to prove $Q(\check{\boldsymbol{\beta}}_\lambda + \boldsymbol{w}\sqrt{\log(p)/n}) - Q(\check{\boldsymbol{\beta}}_\lambda) > 0$. Expanding terms,

$$\begin{aligned} Q\left(\check{\boldsymbol{\beta}}_\lambda + \boldsymbol{w}\sqrt{\frac{\log(p)}{n}}\right) - Q\left(\check{\boldsymbol{\beta}}_\lambda\right) &= \frac{1}{2n}\left\|\left(\boldsymbol{y} - \boldsymbol{X}\check{\boldsymbol{\beta}}_\lambda\right) - \boldsymbol{X}\boldsymbol{w}\sqrt{\frac{\log(p)}{n}}\right\|_2^2 - \frac{1}{2n}\left\|\boldsymbol{y} - \boldsymbol{X}\check{\boldsymbol{\beta}}_\lambda\right\|_2^2 \\ &\quad + \lambda\left\|\check{\boldsymbol{\beta}}_\lambda + \boldsymbol{w}\sqrt{\frac{\log(p)}{n}}\right\|_1 - \lambda\left\|\check{\boldsymbol{\beta}}_\lambda\right\|_1 \\ &= -\frac{\sqrt{\log(p)}}{n^{3/2}}\boldsymbol{w}^\top\boldsymbol{X}^\top\left(\boldsymbol{y} - \boldsymbol{X}\check{\boldsymbol{\beta}}_\lambda\right) + \frac{\log(p)}{2n^2}\boldsymbol{w}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w} \\ &\quad + \lambda\left\|\check{\boldsymbol{\beta}}_\lambda + \boldsymbol{w}\sqrt{\frac{\log(p)}{n}}\right\|_1 - \lambda\left\|\check{\boldsymbol{\beta}}_\lambda\right\|_1. \end{aligned} \tag{A.19}$$

Now, for any $g, h \in \mathbb{R}$, $g \neq 0$, if g and h have the same sign, $|g+h| = |g| + |h| = |g| + \text{sign}(g)h$. If they have opposite signs, $|g+h| = ||g| - |h|| \geq |g| - |h| = |g| + \text{sign}(g)h$. Finally, if $h = 0$, $|g+h| = |g| + \text{sign}(g)h$. Thus, for any $g, h \in \mathbb{R}$ such that $g \neq 0$, $|g+h| \geq |g| + \text{sign}(g)h$. Given that $\check{\boldsymbol{\tau}}_{\lambda,\check{\mathcal{A}}_\lambda} = \text{sign}(\check{\boldsymbol{\beta}}_{\lambda,\check{\mathcal{A}}_\lambda})$,

$$\begin{aligned} \left\|\check{\boldsymbol{\beta}}_\lambda + \boldsymbol{w}\sqrt{\frac{\log(p)}{n}}\right\|_1 &= \left\|\check{\boldsymbol{\beta}}_{\lambda,\setminus\check{\mathcal{A}}_\lambda} + \boldsymbol{w}_{\setminus\check{\mathcal{A}}_\lambda}\sqrt{\frac{\log(p)}{n}}\right\|_1 + \left\|\check{\boldsymbol{\beta}}_{\lambda,\check{\mathcal{A}}_\lambda} + \boldsymbol{w}_{\check{\mathcal{A}}_\lambda}\sqrt{\frac{\log(p)}{n}}\right\|_1 \\ &= \sqrt{\frac{\log(p)}{n}}\left\|\boldsymbol{w}_{\setminus\check{\mathcal{A}}_\lambda}\right\|_1 + \sum_{j\in\check{\mathcal{A}}_\lambda}\left|\check{\beta}_{\lambda,j} + \sqrt{\frac{\log(p)}{n}}w_j\right| \\ &\geq \sqrt{\frac{\log(p)}{n}}\left\|\boldsymbol{w}_{\setminus\check{\mathcal{A}}_\lambda}\right\|_1 + \sum_{j\in\check{\mathcal{A}}_\lambda}\left(\left|\check{\beta}_{\lambda,j}\right| + \sqrt{\frac{\log(p)}{n}}\check{\tau}_{\lambda,j}w_j\right) \\ &= \sqrt{\frac{\log(p)}{n}}\left\|\boldsymbol{w}_{\setminus\check{\mathcal{A}}_\lambda}\right\|_1 + \left\|\check{\boldsymbol{\beta}}_{\lambda,\check{\mathcal{A}}_\lambda}\right\|_1 + \sqrt{\frac{\log(p)}{n}}\check{\boldsymbol{\tau}}_{\lambda,\check{\mathcal{A}}_\lambda}^\top\boldsymbol{w}_{\check{\mathcal{A}}_\lambda} \\ &= \sqrt{\frac{\log(p)}{n}}\left\|\boldsymbol{w}_{\setminus\check{\mathcal{A}}_\lambda}\right\|_1 + \left\|\check{\boldsymbol{\beta}}_\lambda\right\|_1 + \sqrt{\frac{\log(p)}{n}}\check{\boldsymbol{\tau}}_{\lambda,\check{\mathcal{A}}_\lambda}^\top\boldsymbol{w}_{\check{\mathcal{A}}_\lambda}, \end{aligned} \tag{A.20}$$

where the inequality makes use of $|g+h| \geq |g| + \text{sign}(g)h$, and the second and last equalities

use the fact that $\check{\boldsymbol{\beta}}_{\lambda,\backslash\check{\mathcal{A}}_\lambda} = \mathbf{0}$. Therefore, denoting $\boldsymbol{G} \equiv \boldsymbol{X}^\top \boldsymbol{X}/n$,

$$\begin{aligned} Q\left(\check{\boldsymbol{\beta}}_\lambda + \boldsymbol{w}\sqrt{\frac{\log(p)}{n}}\right) - Q\left(\check{\boldsymbol{\beta}}_\lambda\right) &\geq -\frac{\sqrt{\log(p)}}{n^{3/2}}\boldsymbol{w}^\top \boldsymbol{X}^\top\left(\boldsymbol{y} - \boldsymbol{X}\check{\boldsymbol{\beta}}_\lambda\right) + \frac{\log(p)}{2n^2}\boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} \\ &\quad + \lambda\sqrt{\frac{\log(p)}{n}}\check{\boldsymbol{\tau}}_{\lambda,\check{\mathcal{A}}_\lambda}^\top \boldsymbol{w}_{\check{\mathcal{A}}_\lambda} + \lambda\sqrt{\frac{\log(p)}{n}}\left\|\boldsymbol{w}_{\backslash\check{\mathcal{A}}_\lambda}\right\|_1 \\ &= -\frac{\sqrt{\log(p)}}{n^{3/2}}\boldsymbol{w}^\top \boldsymbol{X}^\top\left(\boldsymbol{y} - \boldsymbol{X}\check{\boldsymbol{\beta}}_\lambda\right) + \frac{\log(p)}{2n}\boldsymbol{w}^\top \boldsymbol{G}\boldsymbol{w} \\ &\quad + \lambda\sqrt{\frac{\log(p)}{n}}\boldsymbol{\tau}_\lambda^\top \boldsymbol{w} + \lambda\sqrt{\frac{\log(p)}{n}}\left(\left\|\boldsymbol{w}_{\backslash\check{\mathcal{A}}_\lambda}\right\|_1 - \check{\boldsymbol{\tau}}_{\lambda,\backslash\check{\mathcal{A}}_\lambda}^\top \boldsymbol{w}_{\backslash\check{\mathcal{A}}_\lambda}\right) \\ &= -\frac{\sqrt{\log(p)}}{n^{3/2}}\boldsymbol{w}^\top\left(\boldsymbol{X}^\top\left(\boldsymbol{y} - \boldsymbol{X}\check{\boldsymbol{\beta}}_\lambda\right) - \lambda n\check{\boldsymbol{\tau}}_\lambda\right) \\ &\quad + \frac{\log(p)}{2n}\boldsymbol{w}^\top \boldsymbol{G}\boldsymbol{w} + \lambda\sqrt{\frac{\log(p)}{n}}\left(\left\|\boldsymbol{w}_{\backslash\check{\mathcal{A}}_\lambda}\right\|_1 - \check{\boldsymbol{\tau}}_{\lambda,\backslash\check{\mathcal{A}}_\lambda}^\top \boldsymbol{w}_{\backslash\check{\mathcal{A}}_\lambda}\right) \end{aligned} \tag{A.21}$$

Since $\limsup_{n\to\infty}\|\check{\boldsymbol{\tau}}_{\lambda,\backslash\check{\mathcal{A}}_\lambda}\|_\infty \leq 1-\delta$ by (**T**), for n sufficiently large, $\|\check{\boldsymbol{\tau}}_{\lambda,\backslash\check{\mathcal{A}}_\lambda}\|_\infty \leq 1-\delta/2$. Thus, $\check{\boldsymbol{\tau}}_{\lambda,\backslash\check{\mathcal{A}}_\lambda}^\top \boldsymbol{w}_{\backslash\check{\mathcal{A}}_\lambda} \leq \|\check{\boldsymbol{\tau}}_{\lambda,\backslash\check{\mathcal{A}}_\lambda}\|_\infty \|\boldsymbol{w}_{\backslash\check{\mathcal{A}}_\lambda}\|_1 \leq (1-\delta/2)\|\boldsymbol{w}_{\backslash\check{\mathcal{A}}_\lambda}\|_1$. Therefore,

$$\begin{aligned} Q\left(\check{\boldsymbol{\beta}}_\lambda + \boldsymbol{w}\sqrt{\frac{\log(p)}{n}}\right) - Q\left(\boldsymbol{\beta}_\lambda\right) &\geq -\frac{\sqrt{\log(p)}}{n^{3/2}}\boldsymbol{w}^\top\left(\boldsymbol{X}^\top\left(\boldsymbol{y} - \boldsymbol{X}\check{\boldsymbol{\beta}}_\lambda\right) - \lambda n\check{\boldsymbol{\tau}}_\lambda\right) \\ &\quad + \left(\frac{\log(p)}{2n}\boldsymbol{w}^\top \boldsymbol{G}\boldsymbol{w} + \lambda\frac{\delta}{2}\sqrt{\frac{\log(p)}{n}}\left\|\boldsymbol{w}_{\backslash\check{\mathcal{A}}_\lambda}\right\|_1\right). \end{aligned} \tag{A.22}$$

The stationary condition of (1.6) in (A.5) gives that $\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\check{\boldsymbol{\beta}}_\lambda) - \lambda n\check{\boldsymbol{\tau}}_\lambda = \boldsymbol{X}^\top\boldsymbol{\epsilon}$. Therefore,

$$\begin{aligned} \boldsymbol{w}^\top\left(\boldsymbol{X}^\top\left(\boldsymbol{y} - \boldsymbol{X}\check{\boldsymbol{\beta}}_\lambda\right) - \lambda n\check{\boldsymbol{\tau}}_\lambda\right) &\leq \left|\boldsymbol{w}^\top\left(\boldsymbol{X}^\top\left(\boldsymbol{y} - \boldsymbol{X}\check{\boldsymbol{\beta}}_\lambda\right) - \lambda n\check{\boldsymbol{\tau}}_\lambda\right)\right| \\ &\leq \|\boldsymbol{w}\|_1\left\|\boldsymbol{X}^\top\left(\boldsymbol{y} - \boldsymbol{X}\check{\boldsymbol{\beta}}_\lambda\right) - \lambda n\check{\boldsymbol{\tau}}_\lambda\right\|_\infty \\ &= \|\boldsymbol{w}\|_1\left\|\boldsymbol{X}^\top\boldsymbol{\epsilon}\right\|_\infty = m\left\|\boldsymbol{X}^\top\boldsymbol{\epsilon}\right\|_\infty. \end{aligned} \tag{A.23}$$

To bound the other term $\log(p)\boldsymbol{w}^\top\boldsymbol{G}\boldsymbol{w}/2n + \lambda\delta\sqrt{\log(p)}\|\boldsymbol{w}_{\backslash\check{\mathcal{A}}_\lambda}\|_1/(2\sqrt{n})$ in (A.22), we consider two cases: $\check{\mathcal{A}}_\lambda = \emptyset$ and $\check{\mathcal{A}}_\lambda \neq \emptyset$.

Case 1: $\check{\mathcal{A}}_\lambda = \emptyset$. In this case, because $\boldsymbol{G}$ is positive semi-definite,

$$\begin{aligned}\frac{\log(p)}{2n}\boldsymbol{w}^\top \boldsymbol{G}\boldsymbol{w} + \lambda\frac{\delta}{2}\sqrt{\frac{\log(p)}{n}}\left\|\boldsymbol{w}_{\setminus\check{\mathcal{A}}_\lambda}\right\|_1 &\geq \lambda\frac{\delta}{2}\sqrt{\frac{\log(p)}{n}}\left\|\boldsymbol{w}\right\|_1 \\ &= \lambda\frac{\delta}{2}\sqrt{\frac{\log(p)}{n}}m. \end{aligned} \tag{A.24}$$

Combining (A.22), (A.23) and (A.24),

$$\begin{aligned} Q\left(\check{\boldsymbol{\beta}}_\lambda + \boldsymbol{w}\sqrt{\frac{\log(p)}{n}}\right) - Q\left(\check{\boldsymbol{\beta}}_\lambda\right) &\geq -\frac{\sqrt{\log(p)}}{n^{3/2}}m\left\|\boldsymbol{X}^\top\boldsymbol{\epsilon}\right\|_\infty + \lambda\frac{\delta}{2}\sqrt{\frac{\log(p)}{n}}m \\ &= m\sqrt{\frac{\log(p)}{n}}\left(\lambda\frac{\delta}{2} - \frac{\left\|\boldsymbol{X}^\top\boldsymbol{\epsilon}\right\|_\infty}{n}\right). \end{aligned} \tag{A.25}$$

By Lemma A.1.2, $\|\boldsymbol{X}^\top\boldsymbol{\epsilon}\|_\infty/n = \mathcal{O}_p(\sqrt{\log(p)/n})$, i.e., there exists a constant $C > 0$, not depending on n, such that with n sufficiently large, $\Pr[\|\boldsymbol{X}^\top\boldsymbol{\epsilon}\|_\infty/n \geq C\sqrt{\log(p)/n}] \leq \xi$. Based on (**T**) that $\sqrt{\log(p)/n}/(\lambda\delta) \to 0$, with n sufficiently large,

$$\lambda\frac{\delta}{2} > C\sqrt{\frac{\log(p)}{n}}, \tag{A.26}$$

which means that with n sufficiently large,

$$\Pr\left[\left(\lambda\frac{\delta}{2} - \frac{\left\|\boldsymbol{X}^\top\boldsymbol{\epsilon}\right\|_\infty}{n}\right) > 0\right] \leq \xi.$$

Thus, for any $m > 0$ and n sufficiently large, (A.18) holds.

Case 2: $\check{\mathcal{A}}_\lambda \neq \emptyset$. We further break down the argument into two cases: $\|\boldsymbol{w}_{\setminus\check{\mathcal{A}}_\lambda}\|_1 > \|\boldsymbol{w}_{\check{\mathcal{A}}_\lambda}\|_1$ and $\|\boldsymbol{w}_{\setminus k}\|_1 \leq |w_k|$, where $\check{\mathcal{A}}_\lambda \equiv \mathrm{supp}(\check{\boldsymbol{\beta}}_\lambda)$ and k is any index such that $k \in \check{\mathcal{A}}_\lambda$. These two cases are not mutually exclusive. However, since $k \in \check{\mathcal{A}}_\lambda$, for any $\boldsymbol{w}$, we have $\|\boldsymbol{w}_{\setminus\check{\mathcal{A}}_\lambda}\|_1 \leq \|\boldsymbol{w}_{\setminus k}\|_1$, and $|w_k| \leq \|\boldsymbol{w}_{\check{\mathcal{A}}_\lambda}\|_1$. Hence $\|\boldsymbol{w}_{\setminus k}\|_1 \leq |w_k|$ implies $\|\boldsymbol{w}_{\setminus\check{\mathcal{A}}_\lambda}\|_1 \leq \|\boldsymbol{w}_{\check{\mathcal{A}}_\lambda}\|_1$. Therefore, although the two cases $\|\boldsymbol{w}_{\setminus\check{\mathcal{A}}_\lambda}\|_1 > \|\boldsymbol{w}_{\check{\mathcal{A}}_\lambda}\|_1$ and $\|\boldsymbol{w}_{\setminus k}\|_1 \leq |w_k|$ are not mutually exclusive, they cover all possibilities.

If $\|\boldsymbol{w}_{\setminus\check{\mathcal{A}}_\lambda}\|_1 \geq \|\boldsymbol{w}_{\check{\mathcal{A}}_\lambda}\|_1$, because $\|\boldsymbol{w}\|_1 = \|\boldsymbol{w}_{\setminus\check{\mathcal{A}}_\lambda}\|_1 + \|\boldsymbol{w}_{\check{\mathcal{A}}_\lambda}\|_1 = m$, $\|\boldsymbol{w}_{\setminus\check{\mathcal{A}}_\lambda}\|_1 \geq m/2$. Also

because $\boldsymbol{G}$ is positive semi-definite,

$$\frac{\log(p)}{2n}\boldsymbol{w}^\top \boldsymbol{G}\boldsymbol{w} + \lambda\frac{\delta}{2}\sqrt{\frac{\log(p)}{n}}\left\|\boldsymbol{w}_{\setminus \check{\mathcal{A}}_\lambda}\right\|_1 \geq \lambda\frac{\delta}{2}\sqrt{\frac{\log(p)}{n}}\left\|\boldsymbol{w}_{\setminus \check{\mathcal{A}}_\lambda}\right\|_1 \geq \lambda\frac{\delta}{4}\sqrt{\frac{\log(p)}{n}}m. \tag{A.27}$$

Combining (A.22), (A.23) and (A.27),

$$Q\left(\check{\boldsymbol{\beta}}_\lambda + \boldsymbol{w}\sqrt{\frac{\log(p)}{n}}\right) - Q\left(\check{\boldsymbol{\beta}}_\lambda\right) \geq -\frac{\sqrt{\log(p)}}{n^{3/2}}m\left\|\boldsymbol{X}^\top\boldsymbol{\epsilon}\right\|_\infty + \lambda\frac{\delta}{4}\sqrt{\frac{\log(p)}{n}}m = m\sqrt{\frac{\log(p)}{n}}\left(\frac{\delta}{4}\lambda - \frac{\left\|\boldsymbol{X}^\top\boldsymbol{\epsilon}\right\|_\infty}{n}\right). \tag{A.28}$$

Based on (**T**) that $\sqrt{\log(p)/n}/(\lambda\delta) \to 0$, for n sufficiently large,

$$\frac{\delta}{4}\lambda > C\sqrt{\frac{\log(p)}{n}}. \tag{A.29}$$

Thus, with any $m > 0$ and n sufficiently large, (A.18) holds.

On the other hand, if $\|\boldsymbol{w}_{\setminus k}\|_1 \leq |w_k|$, because $\|\boldsymbol{w}\|_1 = m$, $|w_k| > m/2$. Hence, taking $\mathcal{I} = \{k\}$, (**E**) implies that for n sufficiently large, $\boldsymbol{w}^\top\boldsymbol{G}\boldsymbol{w} \geq \phi^2\|\boldsymbol{w}_{\mathcal{B}}\|_2^2/2$, where $k \in \mathcal{B}$, or, $\|\boldsymbol{w}_{\mathcal{B}}\|_2^2 \geq w_k^2$. Thus,

$$\boldsymbol{w}^\top\boldsymbol{G}\boldsymbol{w} \geq \frac{\phi^2}{2}\|\boldsymbol{w}_{\mathcal{B}}\|_2^2 \geq \frac{\phi^2}{2}w_k^2 \geq \frac{\phi^2}{8}m^2. \tag{A.30}$$

Hence,

$$\frac{\log(p)}{2n}\boldsymbol{w}^\top \boldsymbol{G}\boldsymbol{w} + \lambda\frac{\delta}{2}\sqrt{\frac{\log(p)}{n}}\left\|\boldsymbol{w}_{\setminus \check{\mathcal{A}}_\lambda}\right\|_1 \geq \frac{\log(p)}{2n}\boldsymbol{w}^\top\boldsymbol{G}\boldsymbol{w} \geq \frac{\log(p)\phi^2}{16n}m^2. \tag{A.31}$$

Combining (A.22), (A.23) and (A.31), for n sufficiently large,

$$Q\left(\check{\boldsymbol{\beta}}_\lambda + \boldsymbol{w}\sqrt{\frac{\log(p)}{n}}\right) - Q\left(\check{\boldsymbol{\beta}}_\lambda\right) \geq -\frac{\sqrt{\log(p)}}{n^{3/2}}m\left\|\boldsymbol{X}^\top\boldsymbol{\epsilon}\right\|_\infty + \frac{\log(p)\phi^2}{16n}m^2 = m\sqrt{\frac{\log(p)}{n}}\left(\frac{\phi^2}{16}\sqrt{\frac{\log(p)}{n}}m - \frac{\left\|\boldsymbol{X}^\top\boldsymbol{\epsilon}\right\|_\infty}{n}\right). \tag{A.32}$$

Since $\Pr[\|\boldsymbol{X}^\top \boldsymbol{\epsilon}\|_\infty/n \geq C\sqrt{\log(p)/n}] \leq \xi$ by Lemma A.1.2, we can choose $m > 16C/\phi^2$, not depending on n, such that for n sufficiently large, (A.18) holds. □

Lemma A.1.5. *Suppose (***M1***), (***M3***), (***M4***), (***E***) and (***T***) hold. For* $\check{\mathcal{A}}_\lambda \neq \emptyset$,

$$\sqrt{\frac{\log(p)}{n}} \frac{1}{\check{b}_{\lambda \min}} \to 0, \tag{A.33}$$

where $\check{b}_{\lambda \min} \equiv \min_{j \in \check{\mathcal{A}}_\lambda} |\check{\beta}_{\lambda,j}|$, *and* $\check{\boldsymbol{\beta}}_\lambda$ *is defined in* (1.6).

Proof. To prove the result, we show that $\sqrt{\log(p)/n}/|\check{\beta}_{\lambda,j}| \to 0$ for any $j \in \check{\mathcal{A}}_\lambda$. This is proven separately for entries in $\mathcal{S} \cap \check{\mathcal{A}}_\lambda$ and in $\check{\mathcal{A}}_\lambda \backslash \mathcal{S}$, where $\mathcal{S} \equiv \{j : |\beta_j| > 3\lambda\sqrt{q}/\phi^2\}$. By Lemma A.1.3, $\mathcal{S} \subseteq \check{\mathcal{A}}_\lambda$ and thus, $\mathcal{S} \cap \check{\mathcal{A}}_\lambda = \mathcal{S}$.

We first show that for any $j \in \mathcal{S} \subseteq \check{\mathcal{A}}_\lambda$, $\sqrt{\log(p)/n}/|\check{\beta}_{\lambda,j}| \to 0$. For n sufficiently large, Lemma A.1.3 gives us

$$\left\|\check{\boldsymbol{\beta}}_{\lambda,\mathcal{S}} - \boldsymbol{\beta}_{\mathcal{S}}\right\|_\infty \leq \left\|\check{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\right\|_\infty \leq \left\|\check{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\right\|_2 \leq \frac{2\lambda\sqrt{2q}}{\phi^2}. \tag{A.34}$$

Thus, for any $j \in \mathcal{S}$, i.e., $|\beta_j| > 3\lambda\sqrt{q}/\phi^2$, $|\check{\beta}_{\lambda,j}| > (3 - 2\sqrt{2})\lambda\sqrt{q}/\phi^2$. Therefore,

$$0 < \sqrt{\frac{\log(p)}{n}} \frac{1}{\left|\check{\beta}_{\lambda,j}\right|} < \sqrt{\frac{\log(p)}{n}} \frac{1}{\lambda} \cdot \frac{\phi^2}{\left(3 - 2\sqrt{2}\right)\sqrt{q}} \to 0,$$

by (**M3**) and (**E**).

If $\check{\mathcal{A}}_\lambda = \mathcal{S}$, then our proof is complete. If $\check{\mathcal{A}}_\lambda \neq \mathcal{S}$, by Lemma A.1.3, (**E**) implies that $\check{\mathcal{A}}_\lambda \supset \mathcal{S}$. We now proceed to show that in the case that $\check{\mathcal{A}}_\lambda \supset \mathcal{S}$, $\sqrt{\log(p)/n}/|\check{\beta}_{\lambda,j}| \to 0$ for $j \in \check{\mathcal{A}}_\lambda \backslash \mathcal{S}$. Consider the stationary condition of (1.6),

$$\begin{aligned} n\lambda \check{\boldsymbol{\tau}}_{\lambda,\check{\mathcal{A}}_\lambda} &= \boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top \boldsymbol{X} \left(\boldsymbol{\beta} - \check{\boldsymbol{\beta}}_\lambda\right) \\ &= \boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top \boldsymbol{X}_{\check{\mathcal{A}}_\lambda} \left(\boldsymbol{\beta}_{\check{\mathcal{A}}_\lambda} - \check{\boldsymbol{\beta}}_{\lambda,\check{\mathcal{A}}_\lambda}\right) + \boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top \boldsymbol{X}_{\backslash \check{\mathcal{A}}_\lambda} \boldsymbol{\beta}_{\backslash \check{\mathcal{A}}_\lambda}, \end{aligned} \tag{A.35}$$

where the second equality holds because $\check{\boldsymbol{\beta}}_{\lambda,\backslash\check{\mathcal{A}}_\lambda} = \boldsymbol{0}$.

Denote $\boldsymbol{G} \equiv \boldsymbol{X}^\top \boldsymbol{X}/n$. By definition,

$$\phi_{\min}^2 \left[\boldsymbol{G}_{(\check{\mathcal{A}}_\lambda,\check{\mathcal{A}}_\lambda)}\right] \equiv \min_{\boldsymbol{a} \in \mathbb{R}^p : \boldsymbol{a}_{\backslash \check{\mathcal{A}}_\lambda} = \boldsymbol{0}} \frac{\boldsymbol{a}^\top \boldsymbol{G} \boldsymbol{a}}{\|\boldsymbol{a}\|_2^2} = \min_{\boldsymbol{a} \in \mathbb{R}^p : \boldsymbol{a}_{\backslash \check{\mathcal{A}}_\lambda} = \boldsymbol{0}} \frac{\boldsymbol{a}^\top \boldsymbol{G} \boldsymbol{a}}{\|\boldsymbol{a}_{\check{\mathcal{A}}_\lambda}\|_2^2}.$$

For any $\boldsymbol{a} \in \mathbb{R}^p$ such that $\boldsymbol{a}_{\setminus \check{\mathcal{A}}_\lambda} = \mathbf{0}$, we trivially have $\|\boldsymbol{a}_{\setminus \check{\mathcal{A}}_\lambda}\|_1 \leq \|\boldsymbol{a}_{\check{\mathcal{A}}_\lambda}\|_1$, and by (**E**),

$$\liminf_{n\to\infty} \phi^2_{\min}\left[\boldsymbol{G}_{(\check{\mathcal{A}}_\lambda,\check{\mathcal{A}}_\lambda)}\right] = \liminf_{n\to\infty} \min_{\boldsymbol{a}\in\mathbb{R}^p:\boldsymbol{a}_{\setminus\check{\mathcal{A}}_\lambda}=\mathbf{0}} \frac{\boldsymbol{a}^\top \boldsymbol{G}\boldsymbol{a}}{\|\boldsymbol{a}_{\check{\mathcal{A}}_\lambda}\|_2^2} \geq \phi^2 > 0.$$

Thus, with n sufficiently large, $\phi^2_{\min}[\boldsymbol{G}_{(\check{\mathcal{A}}_\lambda,\check{\mathcal{A}}_\lambda)}] > \phi^2/2 > 0$. Thus, $\boldsymbol{G}_{(\check{\mathcal{A}}_\lambda,\check{\mathcal{A}}_\lambda)}$ is invertible.

Rearranging terms of (A.35),

$$\check{\boldsymbol{\beta}}_{\lambda,\check{\mathcal{A}}_\lambda} = \boldsymbol{\beta}_{\check{\mathcal{A}}_\lambda} + \left(\boldsymbol{X}^\top_{\check{\mathcal{A}}_\lambda}\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\right)^{-1}\boldsymbol{X}^\top_{\check{\mathcal{A}}_\lambda}\boldsymbol{X}_{\setminus\check{\mathcal{A}}_\lambda}\boldsymbol{\beta}_{\setminus\check{\mathcal{A}}_\lambda} - n\lambda\left(\boldsymbol{X}^\top_{\check{\mathcal{A}}_\lambda}\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\right)^{-1}\check{\boldsymbol{\tau}}_{\lambda,\check{\mathcal{A}}_\lambda}. \tag{A.36}$$

Thus, for any $j \in \check{\mathcal{A}}_\lambda \setminus \mathcal{S}$, we have

$$\begin{aligned}
\left|\check{\boldsymbol{\beta}}_{\lambda,j}\right| &= \left|\left[n\lambda\left(\boldsymbol{X}^\top_{\check{\mathcal{A}}_\lambda}\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\right)^{-1}\check{\boldsymbol{\tau}}_{\lambda,\check{\mathcal{A}}_\lambda}\right]_j - \boldsymbol{\beta}_j - \left[\left(\boldsymbol{X}^\top_{\check{\mathcal{A}}_\lambda}\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\right)^{-1}\boldsymbol{X}^\top_{\check{\mathcal{A}}_\lambda}\boldsymbol{X}_{\setminus\check{\mathcal{A}}_\lambda}\boldsymbol{\beta}_{\setminus\check{\mathcal{A}}_\lambda}\right]_j\right| \\
&\geq \left|\left|\lambda\left(\boldsymbol{G}_{(\check{\mathcal{A}}_\lambda,\check{\mathcal{A}}_\lambda)}\right)^{-1}\check{\boldsymbol{\tau}}_{\lambda,\check{\mathcal{A}}_\lambda}\right|_j - \left|\beta_j + \left[\left(\boldsymbol{X}^\top_{\check{\mathcal{A}}_\lambda}\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\right)^{-1}\boldsymbol{X}^\top_{\check{\mathcal{A}}_\lambda}\boldsymbol{X}_{\setminus\check{\mathcal{A}}_\lambda}\boldsymbol{\beta}_{\setminus\check{\mathcal{A}}_\lambda}\right]_j\right|\right|.
\end{aligned} \tag{A.37}$$

We now bound the term $(\boldsymbol{X}^\top_{\check{\mathcal{A}}_\lambda}\boldsymbol{X}_{\check{\mathcal{A}}_\lambda})^{-1}\boldsymbol{X}^\top_{\check{\mathcal{A}}_\lambda}\boldsymbol{X}_{\setminus\check{\mathcal{A}}_\lambda}\boldsymbol{\beta}_{\setminus\check{\mathcal{A}}_\lambda}$. We first consider the term $(\boldsymbol{X}^\top_{\check{\mathcal{A}}_\lambda}\boldsymbol{X}_{\check{\mathcal{A}}_\lambda})^{-1}\boldsymbol{X}^\top_{\check{\mathcal{A}}_\lambda}$. For n sufficiently large,

$$\begin{aligned}
\left\|\left(\boldsymbol{X}^\top_{\check{\mathcal{A}}_\lambda}\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\right)^{-1}\boldsymbol{X}^\top_{\check{\mathcal{A}}_\lambda}\right\|_2 &= \sqrt{\phi^2_{\max}\left[\left(\boldsymbol{X}^\top_{\check{\mathcal{A}}_\lambda}\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\right)^{-1}\boldsymbol{X}^\top_{\check{\mathcal{A}}_\lambda}\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\left(\boldsymbol{X}^\top_{\check{\mathcal{A}}_\lambda}\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\right)^{-1}\right]} \\
&= \sqrt{\phi^2_{\max}\left[\left(\boldsymbol{X}^\top_{\check{\mathcal{A}}_\lambda}\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\right)^{-1}\right]} = \sqrt{\frac{1}{n}\phi^2_{\max}\left[\left(\boldsymbol{G}_{(\check{\mathcal{A}}_\lambda,\check{\mathcal{A}}_\lambda)}\right)^{-1}\right]} \\
&= \sqrt{\frac{1}{n}\phi^{-2}_{\min}\left[\boldsymbol{G}_{(\check{\mathcal{A}}_\lambda,\check{\mathcal{A}}_\lambda)}\right]} \leq \sqrt{\frac{2}{n\phi^2}} = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right),
\end{aligned} \tag{A.38}$$

where the inequality is based on the fact that with n sufficiently large, $\phi^2_{\min}[\boldsymbol{G}_{(\check{\mathcal{A}}_\lambda,\check{\mathcal{A}}_\lambda)}] > \phi^2/2$. Thus,

$$\begin{aligned}
\left\|\left(\boldsymbol{X}^\top_{\check{\mathcal{A}}_\lambda}\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\right)^{-1}\boldsymbol{X}^\top_{\check{\mathcal{A}}_\lambda}\boldsymbol{X}_{\setminus\check{\mathcal{A}}_\lambda}\boldsymbol{\beta}_{\setminus\check{\mathcal{A}}_\lambda}\right\|_\infty &\leq \left\|\left(\boldsymbol{X}^\top_{\check{\mathcal{A}}_\lambda}\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\right)^{-1}\boldsymbol{X}^\top_{\check{\mathcal{A}}_\lambda}\boldsymbol{X}_{\setminus\check{\mathcal{A}}_\lambda}\boldsymbol{\beta}_{\setminus\check{\mathcal{A}}_\lambda}\right\|_2 \\
&\leq \left\|\left(\boldsymbol{X}^\top_{\check{\mathcal{A}}_\lambda}\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\right)^{-1}\boldsymbol{X}^\top_{\check{\mathcal{A}}_\lambda}\right\|_2 \left\|\boldsymbol{X}_{\setminus\check{\mathcal{A}}_\lambda}\boldsymbol{\beta}_{\setminus\check{\mathcal{A}}_\lambda}\right\|_2 \\
&= \mathcal{O}\left(\frac{1}{\sqrt{n}}\left\|\boldsymbol{X}_{\setminus\check{\mathcal{A}}_\lambda}\boldsymbol{\beta}_{\setminus\check{\mathcal{A}}_\lambda}\right\|_2\right).
\end{aligned} \tag{A.39}$$

If $\check{\mathcal{A}}_\lambda \supseteq \mathcal{A}$, we have $\|\boldsymbol{X}_{\setminus\check{\mathcal{A}}_\lambda}\boldsymbol{\beta}_{\setminus\check{\mathcal{A}}_\lambda}\|_2 = 0$. Otherwise, if $\mathcal{S} \subset \check{\mathcal{A}}_\lambda \not\supseteq \mathcal{A}$, based on (**M4**),

$$\left\|\boldsymbol{X}_{\setminus\check{\mathcal{A}}_\lambda}\boldsymbol{\beta}_{\setminus\check{\mathcal{A}}_\lambda}\right\|_2 = \left\|\boldsymbol{X}_{\mathcal{A}\setminus\check{\mathcal{A}}_\lambda}\boldsymbol{\beta}_{\mathcal{A}\setminus\check{\mathcal{A}}_\lambda}\right\|_2 = \mathcal{O}\left(\sqrt{\log(p)}\right), \tag{A.40}$$

and based on (A.39),

$$\left\|\left(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\right)^{-1}\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\setminus\check{\mathcal{A}}_\lambda}\boldsymbol{\beta}_{\setminus\check{\mathcal{A}}_\lambda}\right\|_\infty = \mathcal{O}\left(\sqrt{\frac{\log(p)}{n}}\right). \tag{A.41}$$

Thus, for any $j \in \check{\mathcal{A}}_\lambda\setminus\mathcal{S}$, based on (**M4**) that $\|\boldsymbol{\beta}_{\check{\mathcal{A}}_\lambda\setminus\mathcal{S}}\|_\infty = \mathcal{O}(\sqrt{\log(p)/n})$,

$$\begin{aligned}\left|\boldsymbol{\beta}_j + \left[\left(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\right)^{-1}\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\setminus\check{\mathcal{A}}_\lambda}\boldsymbol{\beta}_{\setminus\check{\mathcal{A}}_\lambda}\right]_j\right| &\le \left\|\boldsymbol{\beta}_{\check{\mathcal{A}}_\lambda\setminus\mathcal{S}}\right\|_\infty + \left\|\left(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\right)^{-1}\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\setminus\check{\mathcal{A}}_\lambda}\boldsymbol{\beta}_{\setminus\check{\mathcal{A}}_\lambda}\right\|_\infty \\ &= \mathcal{O}\left(\sqrt{\frac{\log(p)}{n}}\right).\end{aligned} \tag{A.42}$$

Now, by (A.37), for any $j \in \check{\mathcal{A}}_\lambda\setminus\mathcal{S}$,

$$\sqrt{\frac{\log(p)}{n}}\frac{1}{\left|\check{\beta}_{\lambda,j}\right|} \le \frac{\sqrt{\log(p)/n}}{\left|\left|n\lambda\left(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\right)^{-1}\check{\boldsymbol{\tau}}_{\lambda,\check{\mathcal{A}}_\lambda}\right|_j - \left|\boldsymbol{\beta}_j + \left[\left(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\right)^{-1}\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\setminus\check{\mathcal{A}}_\lambda}\boldsymbol{\beta}_{\setminus\check{\mathcal{A}}_\lambda}\right]_j\right|\right|}.$$

But, by (A.42) and (**T**),

$$\frac{\left|\boldsymbol{\beta}_j + \left[\left(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\right)^{-1}\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\setminus\check{\mathcal{A}}_\lambda}\boldsymbol{\beta}_{\setminus\check{\mathcal{A}}_\lambda}\right]_j\right|}{\min_{j\in\check{\mathcal{A}}_\lambda\setminus\mathcal{S}}\left|\lambda\left(\boldsymbol{G}_{(\check{\mathcal{A}}_\lambda,\check{\mathcal{A}}_\lambda)}\right)^{-1}\check{\boldsymbol{\tau}}_{\lambda,\check{\mathcal{A}}_\lambda}\right|_j} \to 0$$

Therefore,

$$\begin{aligned}&\frac{\sqrt{\log(p)/n}}{\left|\min_{j\in\check{\mathcal{A}}_\lambda\setminus\mathcal{S}}\left|\lambda\left(\boldsymbol{G}_{(\check{\mathcal{A}}_\lambda,\check{\mathcal{A}}_\lambda)}\right)^{-1}\check{\boldsymbol{\tau}}_{\lambda,\check{\mathcal{A}}_\lambda}\right|_j - \left|\boldsymbol{\beta}_j + \left[\left(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\right)^{-1}\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\setminus\check{\mathcal{A}}_\lambda}\boldsymbol{\beta}_{\setminus\check{\mathcal{A}}_\lambda}\right]_j\right|\right|} \\ \to &\frac{\sqrt{\log(p)/n}}{\min_{j\in\check{\mathcal{A}}_\lambda\setminus\mathcal{S}}\left|\lambda\left(\boldsymbol{G}_{(\check{\mathcal{A}}_\lambda,\check{\mathcal{A}}_\lambda)}\right)^{-1}\check{\boldsymbol{\tau}}_{\lambda,\check{\mathcal{A}}_\lambda}\right|_j} \to 0.\end{aligned}$$

Thus,

$$\sqrt{\frac{\log(p)}{n}}\frac{1}{\left|\check{\beta}_{\lambda,j}\right|} \to 0,$$

as desired. □

Proof of Proposition 2.2.4. According to the stationary conditions of (1.6) and (1.2), respectively,

$$\lambda n \check{\boldsymbol{\tau}}_\lambda = \boldsymbol{X}^\top \left(\boldsymbol{y} - \boldsymbol{X} \check{\boldsymbol{\beta}}_\lambda \right) - \boldsymbol{X}^\top \boldsymbol{\epsilon}, \tag{A.43}$$

$$\lambda n \hat{\boldsymbol{\tau}}_\lambda = \boldsymbol{X}^\top \left(\boldsymbol{y} - \boldsymbol{X} \hat{\boldsymbol{\beta}}_\lambda \right). \tag{A.44}$$

This implies that

$$\hat{\boldsymbol{\tau}}_\lambda - \check{\boldsymbol{\tau}}_\lambda = \frac{1}{n\lambda} \boldsymbol{X}^\top \boldsymbol{X} \left(\check{\boldsymbol{\beta}}_\lambda - \hat{\boldsymbol{\beta}}_\lambda \right) + \frac{1}{n\lambda} \boldsymbol{X}^\top \boldsymbol{\epsilon}. \tag{A.45}$$

We now bound both terms on the right hand side of (A.45). By Lemma A.1.2,

$$\frac{\left\| \boldsymbol{X}^\top \boldsymbol{\epsilon} \right\|_\infty}{n\lambda} = \mathcal{O}_p \left(\frac{1}{\lambda} \sqrt{\frac{\log(p)}{n}} \right).$$

In addition, Lemma A.1.4 shows that

$$\left\| \hat{\boldsymbol{\beta}}_\lambda - \check{\boldsymbol{\beta}}_\lambda \right\|_1 = \mathcal{O}_p \left(\sqrt{\frac{\log(p)}{n}} \right).$$

Also, because columns of $\boldsymbol{X}$ are standardized by (**M1**), i.e., $\boldsymbol{X}_j^\top \boldsymbol{X}_j = n$ for all $j = 1, \ldots, p$,

$$\frac{1}{n} \left\| \boldsymbol{X}^\top \boldsymbol{X} \right\|_{\max} = 1.$$

Therefore,

$$\begin{aligned}
\frac{1}{n\lambda} \left\| \boldsymbol{X}^\top \boldsymbol{X} \left(\check{\boldsymbol{\beta}}_\lambda - \hat{\boldsymbol{\beta}}_\lambda \right) \right\|_\infty &= \frac{1}{n\lambda} \max_{j=1,\ldots,p} \left\{ \left| \boldsymbol{X}_j^\top \boldsymbol{X} \left(\check{\boldsymbol{\beta}}_\lambda - \hat{\boldsymbol{\beta}}_\lambda \right) \right| \right\} \\
&\leq \frac{1}{n\lambda} \max_{j=1,\ldots,p} \left\{ \left\| \boldsymbol{X}_j^\top \boldsymbol{X} \right\|_\infty \left\| \check{\boldsymbol{\beta}}_\lambda - \hat{\boldsymbol{\beta}}_\lambda \right\|_1 \right\} \\
&= \frac{1}{n} \left\| \boldsymbol{X}^\top \boldsymbol{X} \right\|_{\max} \frac{\left\| \check{\boldsymbol{\beta}}_\lambda - \hat{\boldsymbol{\beta}}_\lambda \right\|_1}{\lambda} \\
&= \mathcal{O}_p \left(\frac{1}{\lambda} \sqrt{\frac{\log(p)}{n}} \right).
\end{aligned} \tag{A.46}$$

Therefore, it follows from (A.45) that $\| \check{\boldsymbol{\tau}}_\lambda - \hat{\boldsymbol{\tau}}_\lambda \|_\infty = \mathcal{O}_p(\sqrt{\log(p)/n}/\lambda)$. By (**T**), $\limsup_{n\to\infty} \| \check{\boldsymbol{\tau}}_{\lambda, \setminus \check{\mathcal{A}}_\lambda} \|_\infty \leq 1 - \delta$ for $\sqrt{\log(p)/n}/(\delta\lambda) > 0$. Since $\| \check{\boldsymbol{\tau}}_\lambda - \hat{\boldsymbol{\tau}}_\lambda \|_\infty = \mathcal{O}_p(\sqrt{\log(p)/n}/\lambda)$, $\lim_{n\to\infty} \Pr[\| \hat{\boldsymbol{\tau}}_{\lambda, \setminus \check{\mathcal{A}}_\lambda} \|_\infty < 1] = 1$. Hence,

$$\lim_{n\to\infty} \Pr \left[\check{\mathcal{A}}_\lambda \supseteq \hat{\mathcal{A}}_\lambda \right] = 1. \tag{A.47}$$

To prove the other direction, if $\check{\mathcal{A}}_\lambda = \emptyset$, then, $\check{\mathcal{A}}_\lambda \subseteq \hat{\mathcal{A}}_\lambda$, Otherwise, if $\check{\mathcal{A}}_\lambda \neq \emptyset$, by Lemma A.1.5,

$$\sqrt{\frac{\log(p)}{n}}\frac{1}{\check{b}_{\lambda\min}} \to 0. \tag{A.48}$$

Based on Lemma A.1.4, $\|\hat{\boldsymbol{\beta}}_\lambda - \check{\boldsymbol{\beta}}_\lambda\|_\infty \leq \|\hat{\boldsymbol{\beta}}_\lambda - \check{\boldsymbol{\beta}}_\lambda\|_1 = \mathcal{O}_p(\sqrt{\log(p)/n})$. Thus, for any $\xi > 0$, there exists a constant $C > 0$, not depending on n, such that for n sufficiently large,

$$\Pr\left[\left\|\hat{\boldsymbol{\beta}}_\lambda - \check{\boldsymbol{\beta}}_\lambda\right\|_\infty > C\sqrt{\frac{\log(p)}{n}}\right] < \xi. \tag{A.49}$$

Based on (A.48), for n sufficiently large, $\check{b}_{\lambda\min} > C\sqrt{\log(p)/n}$, where $\check{b}_{\lambda\min} \equiv \min_{j\in\check{\mathcal{A}}_\lambda}\{|\check{\beta}_{\lambda,j}|\}$. Thus, combining (A.48) and (A.49), for n sufficiently large, whenever $|\check{\beta}_{\lambda,j}| > 0$, $|\check{\beta}_{\lambda,j}| > C\sqrt{\log(p)/n}$ and hence $\Pr[|\hat{\beta}_{\lambda,j}| > 0] > 1 - \xi$. Therefore

$$\lim_{n\to\infty}\Pr\left[\check{\mathcal{A}}_\lambda \subseteq \hat{\mathcal{A}}_\lambda\right] = 1, \tag{A.50}$$

which completes the proof. □

A.2 Proof of Corollary 2.2.5

Proof of Corollary 2.2.5. To verify that we can replace (**M4**) with (**M4a**) and (**M4b**), recall that the condition $\|\boldsymbol{X}_{\mathcal{A}\setminus(\check{\mathcal{A}}_\lambda\cup\mathcal{S})}\boldsymbol{\beta}_{\mathcal{A}\setminus(\check{\mathcal{A}}_\lambda\cup\mathcal{S})}\|_2 = \mathcal{O}(\sqrt{\log(p)})$ in (**M4**) is required only in the proof of Lemma A.1.5 to show (A.41),

$$\left\|\left(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\right)^{-1}\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\setminus\check{\mathcal{A}}_\lambda}\boldsymbol{\beta}_{\setminus\check{\mathcal{A}}_\lambda}\right\|_\infty = \mathcal{O}\left(\sqrt{\frac{\log(p)}{n}}\right).$$

With (**M4a**) and (**M4b**),

$$\begin{aligned}
\left\|\left(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\right)^{-1}\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\setminus\check{\mathcal{A}}_\lambda}\boldsymbol{\beta}_{\setminus\check{\mathcal{A}}_\lambda}\right\|_\infty &= \left\|\left(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\right)^{-1}\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\mathcal{A}\setminus\check{\mathcal{A}}_\lambda}\boldsymbol{\beta}_{\mathcal{A}\setminus\check{\mathcal{A}}_\lambda}\right\|_\infty \\
&\leq \left\|\left(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\right)^{-1}\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\mathcal{A}\setminus\check{\mathcal{A}}_\lambda}\right\|_\infty \left\|\boldsymbol{\beta}_{\mathcal{A}\setminus\check{\mathcal{A}}_\lambda}\right\|_\infty \\
&\leq \left\|\left[\boldsymbol{G}_{(\check{\mathcal{A}}_\lambda,\check{\mathcal{A}}_\lambda)}\right]^{-1}\boldsymbol{G}_{(\check{\mathcal{A}}_\lambda,\mathcal{A}\setminus\check{\mathcal{A}}_\lambda)}\right\|_\infty \left\|\boldsymbol{\beta}_{\mathcal{A}\setminus\mathcal{S}}\right\|_\infty \\
&= \mathcal{O}\left(\sqrt{\frac{\log(p)}{n}}\right),
\end{aligned}$$

where the first equality holds because $\boldsymbol{\beta}_{\setminus\mathcal{A}} \equiv \mathbf{0}$ and the second inequality holds because by Lemma A.1.3, $\mathcal{S} \subseteq \check{\mathcal{A}}_\lambda$, or $\|\boldsymbol{\beta}_{\mathcal{A}\setminus\mathcal{S}}\|_\infty \geq \|\boldsymbol{\beta}_{\mathcal{A}\setminus\check{\mathcal{A}}_\lambda}\|_\infty$. Hence, (**M4a**) and (**M4b**) also imply (A.41). □

A.3 Proof of Remarks 2.2.2 and 2.2.3

Proof of Remark 2.2.2. When $\boldsymbol{G} \equiv \boldsymbol{X}^\top \boldsymbol{X}/n = \boldsymbol{I}$, based on the the stationary condition of the noiseless lasso,

$$\lambda \check{\boldsymbol{\tau}}_\lambda = \boldsymbol{G}\left(\boldsymbol{\beta} - \check{\boldsymbol{\beta}}_\lambda\right) = \boldsymbol{\beta} - \check{\boldsymbol{\beta}}_\lambda. \tag{A.51}$$

Rearranging terms,

$$\check{\boldsymbol{\beta}}_\lambda = \boldsymbol{\beta} - \lambda \check{\boldsymbol{\tau}}_\lambda. \tag{A.52}$$

Thus, $\check{\boldsymbol{\beta}}_\lambda = \text{sign}(\boldsymbol{\beta})(|\boldsymbol{\beta}| - \lambda)_+$. Since for $j \in \check{\mathcal{A}}_\lambda \setminus \mathcal{S}$, $\beta_j \preceq \sqrt{\log(p)/n} \prec \lambda$, for n sufficiently large, we have $\boldsymbol{\beta}_{\lambda,\check{\mathcal{A}}_\lambda\setminus\mathcal{S}} = \mathbf{0}$, or $\check{\mathcal{A}}_\lambda \subseteq \mathcal{S}$. Also based on Lemma A.1.3 that $\check{\mathcal{A}}_\lambda \supseteq \mathcal{S}$, we have $\check{\mathcal{A}}_\lambda = \mathcal{S}$.

In addition, since $\boldsymbol{G} = \boldsymbol{I}$, we have

$$\min_{j\in\check{\mathcal{A}}_\lambda\setminus\mathcal{S}} \left| \left[\boldsymbol{G}_{(\check{\mathcal{A}}_\lambda,\check{\mathcal{A}}_\lambda)}\right]^{-1} \check{\boldsymbol{\tau}}_{\lambda,\check{\mathcal{A}}_\lambda} \right|_j = \min_{j\in\check{\mathcal{A}}_\lambda\setminus\mathcal{S}} |\check{\boldsymbol{\tau}}_{\lambda,j}| = 1.$$

Thus based on (**M3**),

$$\frac{\sqrt{\log(p)/n}/\lambda}{\min_{j\in\check{\mathcal{A}}_\lambda\setminus\mathcal{S}} \left| \left[\boldsymbol{G}_{(\check{\mathcal{A}}_\lambda,\check{\mathcal{A}}_\lambda)}\right]^{-1} \check{\boldsymbol{\tau}}_{\lambda,\check{\mathcal{A}}_\lambda} \right|_j} \to 0,$$

□

Proof of Remark 2.2.3. Based on the stationary condition of the noiseless lasso,

$$\lambda \check{\boldsymbol{\tau}}_\lambda = \boldsymbol{G}\left(\boldsymbol{\beta} - \check{\boldsymbol{\beta}}_\lambda\right). \tag{A.53}$$

Since $\boldsymbol{G}_{(\mathcal{A},\setminus\mathcal{A})} = \mathbf{0}$, we get

$$\lambda \check{\boldsymbol{\tau}}_{\lambda,\mathcal{A}} = \boldsymbol{G}_{(\mathcal{A},\mathcal{A})}\left(\boldsymbol{\beta}_\mathcal{A} - \check{\boldsymbol{\beta}}_{\lambda,\mathcal{A}}\right), \tag{A.54}$$

$$\lambda \check{\boldsymbol{\tau}}_{\lambda,\setminus\mathcal{A}} = \boldsymbol{G}_{(\setminus\mathcal{A},\setminus\mathcal{A})}\left(\boldsymbol{\beta}_{\setminus\mathcal{A}} - \check{\boldsymbol{\beta}}_{\lambda,\setminus\mathcal{A}}\right) = -\boldsymbol{G}_{(\setminus\mathcal{A},\setminus\mathcal{A})}\check{\boldsymbol{\beta}}_{\lambda,\setminus\mathcal{A}}. \tag{A.55}$$

Observe that (A.54) is the stationary condition of the noiseless lasso applied on the data $(\boldsymbol{X}_{\mathcal{A}}, \boldsymbol{y})$ with tuning parameter λ. Thus,

$$\check{\boldsymbol{\beta}}_{\lambda,\mathcal{A}} = \underset{\boldsymbol{b}\in\mathbb{R}^q}{\arg\min} \left\{ \mathrm{E}\left[\|\boldsymbol{y} - \boldsymbol{X}_{\mathcal{A}}\boldsymbol{b}\|_2^2 \right] + \lambda \|\boldsymbol{b}\| \right\}. \tag{A.56}$$

On the other hand, $\check{\boldsymbol{\beta}}_{\lambda,\backslash\mathcal{A}} = \boldsymbol{0}$ is the solution to (A.55). Hence, when $\boldsymbol{G}_{(\mathcal{A},\backslash\mathcal{A})} = \boldsymbol{0}$, $\check{\mathcal{A}}_\lambda \subseteq \mathcal{A}$. Also based on Lemma A.1.3, $\check{\mathcal{A}}_\lambda \supseteq \mathcal{S}$. Thus, $\mathcal{S} \subseteq \check{\mathcal{A}}_\lambda \subseteq \mathcal{A}$.

□

A.4 Proof of Theorem 2.2.6

Proof of Theorem 2.2.6. By Proposition 2.2.4, $\Pr[\hat{\mathcal{A}}_\lambda = \check{\mathcal{A}}_\lambda] \to 1$. Therefore, with probability tending to one,

$$\begin{aligned}
\bar{\beta}_j^{(\hat{\mathcal{A}}_\lambda)} &\equiv \left[\left(\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda} \right)^{-1} \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top \boldsymbol{y} \right]_j \\
&= \left[\left(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top \boldsymbol{X}_{\check{\mathcal{A}}_\lambda} \right)^{-1} \boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top \boldsymbol{y} \right]_j . \qquad (A.57) \\
&= \left[\left(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top \boldsymbol{X}_{\check{\mathcal{A}}_\lambda} \right)^{-1} \boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \right]_j \\
&= \left[\left(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top \boldsymbol{X}_{\check{\mathcal{A}}_\lambda} \right)^{-1} \boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top \boldsymbol{\epsilon} \right]_j + \left[\left(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top \boldsymbol{X}_{\check{\mathcal{A}}_\lambda} \right)^{-1} \boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top \boldsymbol{X}\boldsymbol{\beta} \right]_j . \qquad (A.58)
\end{aligned}$$

We proceed to prove the asymptotic distribution of $[(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top \boldsymbol{X}_{\check{\mathcal{A}}_\lambda})^{-1} \boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top \boldsymbol{\epsilon}]_j$. Dividing it by its standard deviation, $\sigma_{\boldsymbol{\epsilon}} \sqrt{[(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top \boldsymbol{X}_{\check{\mathcal{A}}_\lambda})^{-1}]_{(j,j)}}$, where $\sigma_{\boldsymbol{\epsilon}}$ is the error standard deviation, we get

$$\frac{\left[\left(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top \boldsymbol{X}_{\check{\mathcal{A}}_\lambda} \right)^{-1} \boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top \boldsymbol{\epsilon} \right]_j}{\sigma_{\boldsymbol{\epsilon}} \sqrt{\left[\left(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top \boldsymbol{X}_{\check{\mathcal{A}}_\lambda} \right)^{-1} \right]_{(j,j)}}} = \frac{\boldsymbol{r}^w \boldsymbol{\epsilon}}{\sigma_{\boldsymbol{\epsilon}} \|\boldsymbol{r}^w\|_2}, \tag{A.59}$$

where $\boldsymbol{r}^w \equiv \boldsymbol{e}^j (\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top \boldsymbol{X}_{\check{\mathcal{A}}_\lambda})^{-1} \boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top \in \mathbb{R}^n$, and $\boldsymbol{e}^j$ is the row vector of length $|\check{\mathcal{A}}_\lambda|$ with the entry that corresponds to β_j equal to one, and zero otherwise. In order to use the Lindeberg-Feller

Central Limit Theorem to prove the asymptotic normality of (A.59), we need to show that the Lindeberg's condition holds, i.e.,

$$\lim_{n\to\infty}\sum_{i=1}^{n}\mathrm{E}\left[\frac{\left(r_i^w\epsilon_i\right)^2}{\sigma_{\boldsymbol{\epsilon}}^2\left\|\boldsymbol{r}^w\right\|_2^2}\mathbb{1}\left[\frac{|r_i^w\epsilon_i|}{\sigma_{\boldsymbol{\epsilon}}\left\|\boldsymbol{r}^w\right\|_2}>\eta\right]\right]=0,\quad\forall\eta>0.$$

Given that $|r_i^w|\leq\|\boldsymbol{r}^w\|_\infty$, and that the ϵ_i's are identically distributed,

$$\begin{aligned}0\leq\sum_{i=1}^{n}\mathrm{E}\left[\frac{\left(r_i^w\epsilon_i\right)^2}{\sigma_{\boldsymbol{\epsilon}}^2\left\|\boldsymbol{r}^w\right\|_2^2}\mathbb{1}\left[\frac{|r_i^w\epsilon_i|}{\sigma_{\boldsymbol{\epsilon}}\left\|\boldsymbol{r}^w\right\|_2}>\eta\right]\right]&\leq\sum_{i=1}^{n}\mathrm{E}\left[\frac{\left(r_i^w\epsilon_i\right)^2}{\sigma_{\boldsymbol{\epsilon}}^2\left\|\boldsymbol{r}^w\right\|_2^2}\mathbb{1}\left[\frac{|\epsilon_i|\left\|\boldsymbol{r}^w\right\|_\infty}{\sigma_{\boldsymbol{\epsilon}}\left\|\boldsymbol{r}^w\right\|_2}>\eta\right]\right]\\&=\sum_{i=1}^{n}\frac{r_i^{w2}}{\sigma_{\boldsymbol{\epsilon}}^2\left\|\boldsymbol{r}^w\right\|_2^2}\mathrm{E}\left[\epsilon_i^2\mathbb{1}\left[\frac{|\epsilon_i|\left\|\boldsymbol{r}^w\right\|_\infty}{\sigma_{\boldsymbol{\epsilon}}\left\|\boldsymbol{r}^w\right\|_2}>\eta\right]\right]\\&=\frac{1}{\sigma_{\boldsymbol{\epsilon}}^2}\mathrm{E}\left[\epsilon_1^2\mathbb{1}\left[\frac{|\epsilon_1|\left\|\boldsymbol{r}^w\right\|_\infty}{\sigma_{\boldsymbol{\epsilon}}\left\|\boldsymbol{r}^w\right\|_2}>\eta\right]\right].\end{aligned}$$

Since $\|\boldsymbol{r}^w\|_\infty/\|\boldsymbol{r}^w\|_2\to 0$ by Condition (**W**), $\epsilon_1^2\mathbb{1}\left[|\epsilon_1|\|\boldsymbol{r}^w\|_\infty/(\sigma_{\boldsymbol{\epsilon}}\|\boldsymbol{r}^w\|_2)>\eta\right]\to_p 0$. Thus, because $\epsilon_1^2\geq\epsilon_1^2\mathbb{1}\left[|\epsilon_1|\|\boldsymbol{r}^w\|_\infty/(\sigma_{\boldsymbol{\epsilon}}\|\boldsymbol{r}^w\|_2)>\eta\right]$ with probability one and $\mathrm{E}[\epsilon_1^2]=\sigma_\epsilon^2<\infty$, we use ϵ_1^2 as the dominant random variable, and apply the Dominated Convergence Theorem,

$$\lim_{n\to\infty}\frac{1}{\sigma_{\boldsymbol{\epsilon}}^2}\mathrm{E}\left[\epsilon_1^2\mathbb{1}\left[\frac{|\epsilon_1|\left\|\boldsymbol{r}^w\right\|_\infty}{\sigma_{\boldsymbol{\epsilon}}\left\|\boldsymbol{r}^w\right\|_2}>\eta\right]\right]=\frac{1}{\sigma_{\boldsymbol{\epsilon}}^2}\mathrm{E}\left[\lim_{n\to\infty}\epsilon_1^2\mathbb{1}\left[\frac{|\epsilon_1|\left\|\boldsymbol{r}^w\right\|_\infty}{\sigma_{\boldsymbol{\epsilon}}\left\|\boldsymbol{r}^w\right\|_2}>\eta\right]\right]=0,$$

which gives the Lindeberg's condition.

Thus,

$$\frac{\left[\left(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\right)^{-1}\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top\boldsymbol{\epsilon}\right]_j}{\sigma_{\boldsymbol{\epsilon}}\sqrt{\left[\left(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\right)^{-1}\right]_{(j,j)}}}\to_d\mathcal{N}(0,1).\tag{A.60}$$

Using, again, the fact that by Proposition 2.2.4, $\lim_{n\to\infty}\Pr\left[\check{\mathcal{A}}_\lambda=\hat{\mathcal{A}}_\lambda\right]=1$, we can write

$$\begin{aligned}\frac{\bar{\beta}_j^{(\hat{\mathcal{A}}_\lambda)}-\beta_j^{(\hat{\mathcal{A}}_\lambda)}}{\sigma_{\boldsymbol{\epsilon}}\sqrt{\left[\left(\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}\right)^{-1}\right]_{(j,j)}}}&\equiv\frac{\bar{\beta}_j^{(\hat{\mathcal{A}}_\lambda)}-\left[\left(\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}\right)^{-1}\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top\boldsymbol{X}\boldsymbol{\beta}\right]_j}{\sigma_{\boldsymbol{\epsilon}}\sqrt{\left[\left(\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}\right)^{-1}\right]_{(j,j)}}}\\&\to_p\frac{\left[\left(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\right)^{-1}\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top\boldsymbol{\epsilon}\right]_j}{\sigma_{\boldsymbol{\epsilon}}\sqrt{\left[\left(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\right)^{-1}\right]_{(j,j)}}}\to_d\mathcal{N}(0,1).\end{aligned}$$

□

A.5 Proof of Theorem 2.2.7

Proof. Based on Proposition 2.2.4 that $\Pr[\hat{\mathcal{A}}_\lambda = \check{\mathcal{A}}_\lambda] \to 1$, we have

$$\frac{1}{n-\hat{q}_\lambda}\left\|\boldsymbol{y}-\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}\bar{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda)}\right\|_2^2 \to_p \frac{1}{n-\check{q}_\lambda}\left\|\boldsymbol{y}-\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\bar{\boldsymbol{\beta}}^{(\check{\mathcal{A}}_\lambda)}\right\|_2^2, \tag{A.61}$$

where $\hat{q}_\lambda \equiv |\hat{\mathcal{A}}_\lambda|$, $\check{q}_\lambda \equiv |\check{\mathcal{A}}_\lambda|$. Denoting $\boldsymbol{P}^{(\check{\mathcal{A}}_\lambda)} \equiv \boldsymbol{X}_{\check{\mathcal{A}}_\lambda}(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top \boldsymbol{X}_{\check{\mathcal{A}}_\lambda})^{-1}\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top$,

$$\begin{aligned}
&\left\|\boldsymbol{y}-\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\bar{\boldsymbol{\beta}}^{(\check{\mathcal{A}}_\lambda)}\right\|_2^2 \\
\equiv &\left\|\boldsymbol{y}-\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\left(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}^\top\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\right)^{-1}\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\boldsymbol{y}\right\|_2^2 \\
= &\,\boldsymbol{y}^\top\left(\boldsymbol{I}-\boldsymbol{P}^{(\check{\mathcal{A}}_\lambda)}\right)^2\boldsymbol{y} \\
= &\,\boldsymbol{y}^\top\left(\boldsymbol{I}-\boldsymbol{P}^{(\check{\mathcal{A}}_\lambda)}\right)\boldsymbol{y} \\
= &\left(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\boldsymbol{\beta}_{\check{\mathcal{A}}_\lambda}+\boldsymbol{X}_{\setminus\check{\mathcal{A}}_\lambda}\boldsymbol{\beta}_{\setminus\check{\mathcal{A}}_\lambda}+\boldsymbol{\epsilon}\right)^\top\left(\boldsymbol{I}-\boldsymbol{P}^{(\check{\mathcal{A}}_\lambda)}\right)\left(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\boldsymbol{\beta}_{\check{\mathcal{A}}_\lambda}+\boldsymbol{X}_{\setminus\check{\mathcal{A}}_\lambda}\boldsymbol{\beta}_{\setminus\check{\mathcal{A}}_\lambda}+\boldsymbol{\epsilon}\right) \\
= &\left(\boldsymbol{X}_{\setminus\check{\mathcal{A}}_\lambda}\boldsymbol{\beta}_{\setminus\check{\mathcal{A}}_\lambda}+\boldsymbol{\epsilon}\right)^\top\left(\boldsymbol{I}-\boldsymbol{P}^{(\check{\mathcal{A}}_\lambda)}\right)\left(\boldsymbol{X}_{\setminus\check{\mathcal{A}}_\lambda}\boldsymbol{\beta}_{\setminus\check{\mathcal{A}}_\lambda}+\boldsymbol{\epsilon}\right) \\
= &\left(\boldsymbol{X}_{\mathcal{A}\setminus\check{\mathcal{A}}_\lambda}\boldsymbol{\beta}_{\mathcal{A}\setminus\check{\mathcal{A}}_\lambda}+\boldsymbol{\epsilon}\right)^\top\left(\boldsymbol{I}-\boldsymbol{P}^{(\check{\mathcal{A}}_\lambda)}\right)\left(\boldsymbol{X}_{\mathcal{A}\setminus\check{\mathcal{A}}_\lambda}\boldsymbol{\beta}_{\mathcal{A}\setminus\check{\mathcal{A}}_\lambda}+\boldsymbol{\epsilon}\right) &\text{(A.62)} \\
= &\left(\boldsymbol{X}_{\mathcal{A}\setminus(\check{\mathcal{A}}_\lambda\cup\mathcal{S})}\boldsymbol{\beta}_{\mathcal{A}\setminus(\check{\mathcal{A}}_\lambda\cup\mathcal{S})}+\boldsymbol{\epsilon}\right)^\top\left(\boldsymbol{I}-\boldsymbol{P}^{(\check{\mathcal{A}}_\lambda)}\right)\left(\boldsymbol{X}_{\mathcal{A}\setminus(\check{\mathcal{A}}_\lambda\cup\mathcal{S})}\boldsymbol{\beta}_{\mathcal{A}\setminus(\check{\mathcal{A}}_\lambda\cup\mathcal{S})}+\boldsymbol{\epsilon}\right) &\text{(A.63)}
\end{aligned}$$

where (A.62) is based on the fact that $\boldsymbol{\beta}_{\setminus\mathcal{A}} \equiv \boldsymbol{0}$ and (A.63) holds because based on Lemma A.1.3, $\check{\mathcal{A}}_\lambda \supseteq \mathcal{S}$. To simplify notations, denote $\boldsymbol{\theta} \equiv \boldsymbol{X}_{\mathcal{A}\setminus(\check{\mathcal{A}}_\lambda\cup\mathcal{S})}\boldsymbol{\beta}_{\mathcal{A}\setminus(\check{\mathcal{A}}_\lambda\cup\mathcal{S})}$ and $\boldsymbol{Q} \equiv \boldsymbol{I}-\boldsymbol{P}^{(\check{\mathcal{A}}_\lambda)}$. Expanding (A.63),

$$(\boldsymbol{\theta}+\boldsymbol{\epsilon})^\top\boldsymbol{Q}(\boldsymbol{\theta}+\boldsymbol{\epsilon}) = \boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\theta}+2\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\epsilon}+\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}.$$

Because $\boldsymbol{Q}$ is an idempotent matrix, $\boldsymbol{Q}$ is positive semidefinite, whose eigenvalues are all zeros and ones. Thus,

$$0 \le \boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\theta} \le \phi_{\max}^2[\boldsymbol{Q}]\,\|\boldsymbol{\theta}\|_2^2 = \mathcal{O}(\log(p)),$$

where the last equality is based on (**M4**). Since $\log(p)/(n-\check{q}_\lambda) \to 0$,

$$0 \leq \frac{1}{n-\check{q}_\lambda}\boldsymbol{\theta}^\top \boldsymbol{Q}\boldsymbol{\theta} = \mathcal{O}\left(\frac{\log(p)}{n-\check{q}_\lambda}\right) = o(1),$$

which means that

$$\frac{1}{n-\check{q}_\lambda}\boldsymbol{\theta}^\top \boldsymbol{Q}\boldsymbol{\theta} = o(1). \tag{A.64}$$

Therefore,

$$\frac{1}{n-\check{q}_\lambda}\left\|\boldsymbol{y} - \boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\bar{\boldsymbol{\beta}}^{(\check{\mathcal{A}}_\lambda)}\right\|_2^2 = \frac{2}{n-\check{q}_\lambda}\boldsymbol{\theta}^\top \boldsymbol{Q}\boldsymbol{\epsilon} + \frac{1}{n-\check{q}_\lambda}\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon} + o(1). \tag{A.65}$$

We now derive the expected value and variance of $\|\boldsymbol{y} - \boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\bar{\boldsymbol{\beta}}^{(\check{\mathcal{A}}_\lambda)}\|_2^2/(n-\check{q}_\lambda)$. First, because $\mathrm{E}[\boldsymbol{\epsilon}] = \boldsymbol{0}$ and $\mathrm{Cov}[\boldsymbol{\epsilon}] = \sigma_{\boldsymbol{\epsilon}}^2\boldsymbol{I}$, according to the formula of the expectation of quadratic forms,

$$\begin{aligned}
\mathrm{E}\left[\frac{1}{n-\check{q}_\lambda}\left\|\boldsymbol{y} - \boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\bar{\boldsymbol{\beta}}^{(\check{\mathcal{A}}_\lambda)}\right\|_2^2\right] &= \mathrm{E}\left[\frac{2}{n-\check{q}_\lambda}\boldsymbol{\theta}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\right] + \mathrm{E}\left[\frac{1}{n-\check{q}_\lambda}\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\right] + o(1) \\
&= \mathrm{E}\left[\frac{1}{n-\check{q}_\lambda}\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\right] + o(1) \\
&= \frac{1}{n-\check{q}_\lambda}\sigma_{\boldsymbol{\epsilon}}^2 \mathrm{tr}\left(\boldsymbol{Q}\right) + o(1).
\end{aligned} \tag{A.66}$$

Because $\boldsymbol{Q}$ is an idempotent matrix, we have $\mathrm{tr}(\boldsymbol{Q}) = n - \check{q}_\lambda$. Thus

$$\mathrm{E}\left[\frac{1}{n-\check{q}_\lambda}\left\|\boldsymbol{y} - \boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\bar{\boldsymbol{\beta}}^{(\check{\mathcal{A}}_\lambda)}\right\|_2^2\right] \to \sigma_{\boldsymbol{\epsilon}}^2. \tag{A.67}$$

We now calculate the variance of $\|\boldsymbol{y} - \boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\bar{\boldsymbol{\beta}}^{(\check{\mathcal{A}}_\lambda)}\|_2^2/(n-\check{q}_\lambda)$. Since

$$\begin{aligned}
&\mathrm{Var}\left[\frac{1}{n-\check{q}_\lambda}\left\|\boldsymbol{y} - \boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\bar{\boldsymbol{\beta}}^{(\check{\mathcal{A}}_\lambda)}\right\|_2^2\right] \\
&= \frac{1}{\left(n-\check{q}_\lambda\right)^2}\left(\mathrm{E}\left[\left\|\boldsymbol{y} - \boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\bar{\boldsymbol{\beta}}^{(\check{\mathcal{A}}_\lambda)}\right\|_2^4\right] - \mathrm{E}\left[\left\|\boldsymbol{y} - \boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\bar{\boldsymbol{\beta}}^{(\check{\mathcal{A}}_\lambda)}\right\|_2^2\right]^2\right) \\
&\to \frac{1}{\left(n-\check{q}_\lambda\right)^2}\mathrm{E}\left[\left\|\boldsymbol{y} - \boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\bar{\boldsymbol{\beta}}^{(\check{\mathcal{A}}_\lambda)}\right\|_2^4\right] - \sigma_{\boldsymbol{\epsilon}}^4,
\end{aligned} \tag{A.68}$$

where the second term is derived in (A.67), we now derive the expected value of $\|\boldsymbol{y} - \boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\bar{\boldsymbol{\beta}}^{(\check{\mathcal{A}}_\lambda)}\|_2^4/(n-\check{q}_\lambda)^2$. Based on (A.65),

$$\begin{aligned}\frac{1}{(n-\check{q}_\lambda)^2}\mathrm{E}\left[\left\|\boldsymbol{y} - \boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\bar{\boldsymbol{\beta}}^{(\check{\mathcal{A}}_\lambda)}\right\|_2^4\right] &= o(1)\cdot\frac{1}{n-\check{q}_\lambda}\mathrm{E}\left[2\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right] \\ &+ \frac{4}{(n-\check{q}_\lambda)^2}\mathrm{E}\left[\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\theta}\right] \\ &+ \frac{2}{(n-\check{q}_\lambda)^2}\mathrm{E}\left[\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right] \\ &+ \frac{1}{(n-\check{q}_\lambda)^2}\mathrm{E}\left[\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right] + o(1). \qquad \text{(A.69)}\end{aligned}$$

We now consider each term in the above formulation. First,

$$\frac{1}{n-\check{q}_\lambda}\mathrm{E}\left[2\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right] = \frac{1}{n-\check{q}_\lambda}\mathrm{E}\left[\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right] = \sigma_{\boldsymbol{\epsilon}}^2, \qquad \text{(A.70)}$$

where the last equality is based on (A.66) and (A.67). For the second term, based on (A.64),

$$\frac{4}{(n-\check{q}_\lambda)^2}\mathrm{E}\left[\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\theta}\right] = \frac{4\sigma_{\boldsymbol{\epsilon}}^2}{(n-\check{q}_\lambda)^2}\mathrm{E}\left[\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\theta}\right] = o\left(\frac{1}{n-\check{q}_\lambda}\right). \qquad \text{(A.71)}$$

For the third term.

$$\begin{aligned}&\frac{2}{(n-\check{q}_\lambda)^2}\mathrm{E}\left[\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right] \\ &= \frac{2}{(n-\check{q}_\lambda)^2}\mathrm{E}\left[\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right]\mathrm{E}\left[\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right] + \mathrm{Cor}\left[\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\epsilon}, \boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right]\sqrt{\frac{2\mathrm{Var}\left[\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right]}{(n-\check{q}_\lambda)^2}\frac{2\mathrm{Var}\left[\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right]}{(n-\check{q}_\lambda)^2}} \\ &= \mathrm{Cor}\left[\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\epsilon}, \boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right]\sqrt{\frac{2\mathrm{Var}\left[\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right]}{(n-\check{q}_\lambda)^2}\frac{2\mathrm{Var}\left[\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right]}{(n-\check{q}_\lambda)^2}},\end{aligned}$$

where, based on (A.64),

$$\begin{aligned}\frac{2}{(n-\check{q}_\lambda)^2}\mathrm{Var}\left[\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right] &= \frac{2}{(n-\check{q}_\lambda)^2}\left(\mathrm{E}\left[\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\theta}\right] - \mathrm{E}\left[\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right]^2\right) \\ &= \frac{2\sigma_{\boldsymbol{\epsilon}}^2}{(n-\check{q}_\lambda)^2}\mathrm{E}\left[\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\theta}\right] = o\left(\frac{1}{n-\check{q}_\lambda}\right), \qquad \text{(A.72)}\end{aligned}$$

and

$$\begin{aligned}\frac{2}{(n-\check{q}_\lambda)^2}\mathrm{Var}\left[\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right] &= \frac{2}{(n-\check{q}_\lambda)^2}\left(\mathrm{E}\left[\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right] - \mathrm{E}\left[\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right]^2\right) \\ &= \frac{2}{(n-\check{q}_\lambda)^2}\mathrm{E}\left[\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\boldsymbol{Q}\boldsymbol{\epsilon}\right] - 2\sigma_{\boldsymbol{\epsilon}}^4. \qquad \text{(A.73)}\end{aligned}$$

The last equality is based on (A.66) and (A.67). Since $-1 \leq \mathrm{Cor}[\boldsymbol{\theta}^\top \boldsymbol{Q}\boldsymbol{\epsilon}, \boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}] \leq 1$, based on (A.72) and (A.73), we have

$$
\begin{aligned}
&- \sqrt{\mathcal{O}\left(\frac{1}{(n-\check{q}_\lambda)^3}\right) \mathrm{E}\left[\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\right] - \mathcal{O}\left(\frac{1}{n-\check{q}_\lambda}\right)} \\
&\leq \frac{2}{(n-\check{q}_\lambda)^2} \mathrm{E}\left[\boldsymbol{\theta}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\right] \\
&\leq \sqrt{\mathcal{O}\left(\frac{1}{(n-\check{q}_\lambda)^3}\right) \mathrm{E}\left[\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\right] - \mathcal{O}\left(\frac{1}{n-\check{q}_\lambda}\right)}.
\end{aligned}
$$

Therefore, collecting (A.70), (A.71), (A.72) and (A.73), we have

$$
\begin{aligned}
0 &\leq \frac{1}{(n-\check{q}_\lambda)^2} \mathrm{E}\left[\left\| \boldsymbol{y} - \boldsymbol{X}_{\check{\mathcal{A}}_\lambda} \bar{\boldsymbol{\beta}}^{(\check{\mathcal{A}}_\lambda)} \right\|_2^4\right] \\
&\leq \mathcal{O}\left(\frac{1}{(n-\check{q}_\lambda)^{3/2}}\right) \sqrt{\mathrm{E}\left[\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\right]} + \frac{1}{(n-\check{q}_\lambda)^2} \mathrm{E}\left[\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\right] + \mathcal{O}(1). \qquad \text{(A.74)}
\end{aligned}
$$

We now calculate $\mathrm{E}[\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}]$.

$$
\begin{aligned}
\mathrm{E}\left[\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\right] &= \mathrm{E}\left[\sum_{i,j,l,k}^{n} \epsilon_i \epsilon_j \epsilon_l \epsilon_k Q_{(i,j)} Q_{(l,k)}\right] \\
&= \mathrm{E}\left[\sum_{i=1}^{n} \epsilon_i^4 Q_{(i,i)}^2\right] + \mathrm{E}\left[\sum_{i=l\neq j=k}^{n} \epsilon_i^2 \epsilon_j^2 Q_{(i,j)} Q_{(l,k)}\right] \\
&\quad + \mathrm{E}\left[\sum_{i=k\neq j=l}^{n} \epsilon_i^2 \epsilon_j^2 Q_{(i,j)} Q_{(l,k)}\right] + \mathrm{E}\left[\sum_{i=j\neq l=k}^{n} \epsilon_i^2 \epsilon_l^2 Q_{(i,j)} Q_{(l,k)}\right]. \qquad \text{(A.75)}
\end{aligned}
$$

Because $\boldsymbol{\epsilon}$ is independent and identically distributed with $\mathrm{E}[\epsilon_1^2] = \sigma_{\boldsymbol{\epsilon}}^2$,

$$
\begin{aligned}
\mathrm{E}\left[\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\right] &= \mathrm{E}\left[\epsilon_1^4\right] \sum_{i=1}^{n} Q_{(i,i)}^2 + \mathrm{E}\left[\epsilon_1^2\right]^2 \sum_{i=l\neq j=k}^{n} Q_{(i,j)} Q_{(l,k)} \\
&\quad + \mathrm{E}\left[\epsilon_1^2\right]^2 \sum_{i=k\neq j=l}^{n} Q_{(i,j)} Q_{(l,k)} + \mathrm{E}\left[\epsilon_1^2\right]^2 \sum_{i=j\neq l=k}^{n} Q_{(i,j)} Q_{(l,k)} \\
&= \mathrm{E}\left[\epsilon_1^4\right] \sum_{i=1}^{n} Q_{(i,i)}^2 + \sigma_{\boldsymbol{\epsilon}}^4 \sum_{i\neq j}^{n} Q_{(i,j)}^2 + \sigma_{\boldsymbol{\epsilon}}^4 \sum_{i\neq j}^{n} Q_{(i,j)}^2 + \sigma_{\boldsymbol{\epsilon}}^4 \sum_{i\neq j}^{n} Q_{(i,i)} Q_{(j,j)} \\
&= \mathrm{E}\left[\epsilon_1^4\right] \sum_{i=1}^{n} Q_{(i,i)}^2 + 2\sigma_{\boldsymbol{\epsilon}}^4 \sum_{i\neq j}^{n} Q_{(i,j)}^2 + \sigma_{\boldsymbol{\epsilon}}^4 \sum_{i\neq j}^{n} Q_{(i,i)} Q_{(j,j)}. \qquad \text{(A.76)}
\end{aligned}
$$

Because $(\sum_{i=1}^n Q_{(i,i)})^2 = \sum_{i\neq j}^n Q_{(i,i)}Q_{(j,j)} + \sum_{i=1}^n Q_{(i,i)}^2$,

$$\begin{aligned}
\mathrm{E}\left[\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\right] &= \mathrm{E}\left[\epsilon_1^4\right]\sum_{i=1}^n Q_{(i,i)}^2 + 2\sigma_{\boldsymbol{\epsilon}}^4 \sum_{i\neq j}^n Q_{(i,j)}^2 + \sigma_{\boldsymbol{\epsilon}}^4\left(\sum_{i=1}^n Q_{(i,i)}\right)^2 - \sigma_{\boldsymbol{\epsilon}}^4\sum_{i=1}^n Q_{(i,i)}^2 \\
&= \mathrm{E}\left[\epsilon_1^4\right]\sum_{i=1}^n Q_{(i,i)}^2 + \sigma_{\boldsymbol{\epsilon}}^4 \sum_{i\neq j}^n Q_{(i,j)}^2 + \sigma_{\boldsymbol{\epsilon}}^4\left(\sum_{i=1}^n Q_{(i,i)}\right)^2 \\
&\leq \mathrm{E}\left[\epsilon_1^4\right]\left(\sum_{i=1}^n Q_{(i,i)}^2 + \sum_{i\neq j}^n Q_{(i,j)}^2\right) + \sigma_{\boldsymbol{\epsilon}}^4\left(\sum_{i=1}^n Q_{(i,i)}\right)^2 \\
&= \mathrm{E}\left[\epsilon_1^4\right]\sum_{i,j}^n Q_{(i,j)}^2 + \sigma_{\boldsymbol{\epsilon}}^4 \mathrm{tr}\left(\boldsymbol{Q}\right)^2, \qquad \text{(A.77)}
\end{aligned}$$

where the inequality is based on Jensen's inequality that $\mathrm{E}[\epsilon_1^2]^2 \leq \mathrm{E}[\epsilon_1^4]$. Because $\sum_{i,j}^n Q_{(i,j)}^2 = \mathrm{tr}(\boldsymbol{Q}^2) = \mathrm{tr}(\boldsymbol{Q}) = n - \check{q}_\lambda$, we have

$$\mathrm{E}\left[\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \boldsymbol{Q}\boldsymbol{\epsilon}\right] = (n-\check{q}_\lambda)\,\mathrm{E}\left[\epsilon_1^4\right] + (n-\check{q}_\lambda)^2\,\sigma_{\boldsymbol{\epsilon}}^4. \qquad \text{(A.78)}$$

Since ϵ_1 has sub-Gaussian tails, we have $\mathrm{E}[\epsilon_1^4] = \mathcal{O}(1)$ (see, e.g., Lemma 5.5 in Vershynin, 2012). Thus, based on (A.74),

$$0 \leq \frac{1}{(n-\check{q}_\lambda)^2}\mathrm{E}\left[\left\|\boldsymbol{y} - \boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\bar{\boldsymbol{\beta}}^{(\check{\mathcal{A}}_\lambda)}\right\|_2^4\right] \leq \sigma_{\boldsymbol{\epsilon}}^4 + o(1), \qquad \text{(A.79)}$$

and hence, based on (A.68),

$$\mathrm{Var}\left[\frac{1}{n-\check{q}_\lambda}\left\|\boldsymbol{y} - \boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\bar{\boldsymbol{\beta}}^{(\check{\mathcal{A}}_\lambda)}\right\|_2^2\right] = o(1). \qquad \text{(A.80)}$$

Finally, applying Chebyshev's inequality, we obtain

$$\begin{aligned}
\frac{1}{n-\hat{q}_\lambda}\left\|\boldsymbol{y} - \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}\bar{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda)}\right\|_2^2 &\to_p \frac{1}{n-\check{q}_\lambda}\left\|\boldsymbol{y} - \boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\bar{\boldsymbol{\beta}}^{(\check{\mathcal{A}}_\lambda)}\right\|_2^2 \\
&\to_p \mathrm{E}\left[\frac{1}{n-\check{q}_\lambda}\left\|\boldsymbol{y} - \boldsymbol{X}_{\check{\mathcal{A}}_\lambda}\bar{\boldsymbol{\beta}}^{(\check{\mathcal{A}}_\lambda)}\right\|_2^2\right] = \sigma_{\boldsymbol{\epsilon}}^2. \qquad \text{(A.81)}
\end{aligned}$$

□

A.6 Proof of Theorem 2.4.1

Proof of Theorem 2.4.1. By Proposition 2.2.4, $\Pr[\hat{\mathcal{A}}_\lambda = \check{\mathcal{A}}_\lambda] \to 1$. Therefore, we also have $\Pr[(\hat{\mathcal{A}}_\lambda \backslash \{j\}) = (\check{\mathcal{A}}_\lambda \backslash \{j\})] \to 1$, and with probability tending to one,

$$\begin{aligned} S^j \equiv \boldsymbol{X}_j^\top \left(\boldsymbol{I} - \boldsymbol{P}^{(\hat{\mathcal{A}}_\lambda \backslash \{j\})}\right) \boldsymbol{y} &= \boldsymbol{X}_j^\top \left(\boldsymbol{I} - \boldsymbol{P}^{(\check{\mathcal{A}}_\lambda \backslash \{j\})}\right) \boldsymbol{y} \\ &= \boldsymbol{X}_j^\top \left(\boldsymbol{I} - \boldsymbol{P}^{(\check{\mathcal{A}}_\lambda \backslash \{j\})}\right) \left(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda} \boldsymbol{\beta}_{\check{\mathcal{A}}_\lambda} + \boldsymbol{X}_{\backslash \check{\mathcal{A}}_\lambda} \boldsymbol{\beta}_{\backslash \check{\mathcal{A}}_\lambda} + \boldsymbol{\epsilon}\right), \end{aligned} \tag{A.82}$$

where

$$\boldsymbol{P}^{(\check{\mathcal{A}}_\lambda \backslash \{j\})} \equiv \boldsymbol{X}_{\check{\mathcal{A}}_\lambda \backslash \{j\}} \left(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda \backslash \{j\}}^\top \boldsymbol{X}_{\check{\mathcal{A}}_\lambda \backslash \{j\}}\right)^{-1} \boldsymbol{X}_{\check{\mathcal{A}}_\lambda \backslash \{j\}}^\top.$$

Thus, under the null hypothesis $H_{0,j} : \beta_j = 0$, (A.82) is equal to

$$\begin{aligned} &\boldsymbol{X}_j^\top \left(\boldsymbol{I} - \boldsymbol{P}^{(\check{\mathcal{A}}_\lambda \backslash \{j\})}\right) \left(\boldsymbol{X}_{\check{\mathcal{A}}_\lambda \backslash \{j\}} \boldsymbol{\beta}_{\check{\mathcal{A}}_\lambda \backslash \{j\}} + \boldsymbol{X}_{\backslash \check{\mathcal{A}}_\lambda \backslash \{j\}} \boldsymbol{\beta}_{\backslash \check{\mathcal{A}}_\lambda \backslash \{j\}} + \boldsymbol{\epsilon}\right) \\ &= \boldsymbol{X}_j^\top \left(\boldsymbol{I} - \boldsymbol{P}^{(\check{\mathcal{A}}_\lambda \backslash \{j\})}\right) \boldsymbol{\epsilon} + \boldsymbol{X}_j^\top \left(\boldsymbol{I} - \boldsymbol{P}^{(\check{\mathcal{A}}_\lambda \backslash \{j\})}\right) \boldsymbol{X}_{\backslash \check{\mathcal{A}}_\lambda \backslash \{j\}} \boldsymbol{\beta}_{\backslash \check{\mathcal{A}}_\lambda \backslash \{j\}}. \end{aligned} \tag{A.83}$$

The equality holds because $(\boldsymbol{I} - \boldsymbol{P}^{(\check{\mathcal{A}}_\lambda \backslash \{j\})}) \boldsymbol{X}_{\check{\mathcal{A}}_\lambda \backslash \{j\}} \boldsymbol{\beta}_{\check{\mathcal{A}}_\lambda \backslash \{j\}} = \boldsymbol{0}$.

We first show the asymptotic distribution of $\boldsymbol{X}_j^\top (\boldsymbol{I} - \boldsymbol{P}^{(\check{\mathcal{A}}_\lambda \backslash \{j\})}) \boldsymbol{\epsilon}$. Dividing it by its standard deviation $\sigma_{\boldsymbol{\epsilon}} \sqrt{\boldsymbol{X}_j^\top (\boldsymbol{I} - \boldsymbol{P}^{(\check{\mathcal{A}}_\lambda \backslash \{j\})}) \boldsymbol{X}_j}$, where $\sigma_{\boldsymbol{\epsilon}}$ is the error standard deviation,

$$\frac{\boldsymbol{X}_j^\top \left(\boldsymbol{I} - \boldsymbol{P}^{(\check{\mathcal{A}}_\lambda \backslash \{j\})}\right) \boldsymbol{\epsilon}}{\sigma_{\boldsymbol{\epsilon}} \sqrt{\boldsymbol{X}_j^\top \left(\boldsymbol{I} - \boldsymbol{P}^{(\check{\mathcal{A}}_\lambda \backslash \{j\})}\right) \boldsymbol{X}_j}} = \frac{\boldsymbol{r}^{s\top} \boldsymbol{\epsilon}}{\sigma_{\boldsymbol{\epsilon}} \left\|\boldsymbol{r}^s\right\|_2}, \tag{A.84}$$

where $\boldsymbol{r}^{s\top} \equiv \boldsymbol{X}_j^\top (\boldsymbol{I} - \boldsymbol{P}^{(\check{\mathcal{A}}_\lambda \backslash \{j\})})$. Now, we use the Lindeberg-Feller Central Limit Theorem to prove the asymptotic normality of (A.84). Similar to the proof of Theorem 2.2.6, we need to prove that the Lindeberg's condition holds, i.e.,

$$\lim_{n \to \infty} \sum_{i=1}^n \mathrm{E}\left[\frac{\left(r_i^s \epsilon_i\right)^2}{\sigma_{\boldsymbol{\epsilon}}^2 \left\|\boldsymbol{r}^s\right\|_2^2} \mathbb{1}\left[\frac{|r_i^s \epsilon_i|}{\sigma_{\boldsymbol{\epsilon}} \left\|\boldsymbol{r}^s\right\|_2} > \eta\right]\right] = 0, \quad \forall \eta > 0.$$

Given that $|r_i^s| \leq \left\|\boldsymbol{r}^s\right\|_\infty$, and that the ϵ_i's are identically distributed,

$$0 \leq \sum_{i=1}^n \mathrm{E}\left[\frac{\left(r_i^s \epsilon_i\right)^2}{\sigma_{\boldsymbol{\epsilon}}^2 \left\|\boldsymbol{r}^s\right\|_2^2} \mathbb{1}\left[\frac{|r_i^s \epsilon_i|}{\sigma_{\boldsymbol{\epsilon}} \left\|\boldsymbol{r}^s\right\|_2} > \eta\right]\right] \leq \frac{1}{\sigma_{\boldsymbol{\epsilon}}^2} \mathrm{E}\left[\epsilon_1^2 \mathbb{1}\left[\frac{|\epsilon_1| \left\|\boldsymbol{r}^s\right\|_\infty}{\sigma_{\boldsymbol{\epsilon}} \left\|\boldsymbol{r}^s\right\|_2} > \eta\right]\right].$$

Since $\|\boldsymbol{r}^s\|_\infty/\|\boldsymbol{r}^s\|_2 \to 0$ by Condition (**S**), $\epsilon_1^2\mathbb{1}[|\epsilon_1|\|\boldsymbol{r}^s\|_\infty/(\sigma_{\boldsymbol{\epsilon}}\|\boldsymbol{r}^s\|_2) > \eta] \to_p 0$. Thus, because $\epsilon_1^2 \geq \epsilon_1^2\mathbb{1}[|\epsilon_1|\|\boldsymbol{r}^s\|_\infty/(\sigma_{\boldsymbol{\epsilon}}\|\boldsymbol{r}^s\|_2) > \eta]$ with probability one and $\mathrm{E}[\epsilon_1^2] = \sigma_\epsilon^2 < \infty$, we use ϵ_1^2 as the dominant random variable, and apply the Dominated Convergence Theorem,

$$\lim_{n\to\infty} \frac{1}{\sigma_{\boldsymbol{\epsilon}}^2}\mathrm{E}\left[\epsilon_1^2\mathbb{1}\left[\frac{|\epsilon_1|\,\|\boldsymbol{r}^s\|_\infty}{\sigma_{\boldsymbol{\epsilon}}\,\|\boldsymbol{r}^s\|_2} > \eta\right]\right] = 0,$$

which in turn gives the Lindeberg's condition.

Thus,

$$\frac{\boldsymbol{X}_j^\top\left(\boldsymbol{I} - \boldsymbol{P}^{(\check{\mathcal{A}}_\lambda\setminus\{j\})}\right)\boldsymbol{\epsilon}}{\sigma_{\boldsymbol{\epsilon}}\sqrt{\boldsymbol{X}_j^\top\left(\boldsymbol{I} - \boldsymbol{P}^{(\check{\mathcal{A}}_\lambda\setminus\{j\})}\right)\boldsymbol{X}_j}} \to_d \mathcal{N}(0,1), \tag{A.85}$$

We now prove the asymptotic unbiasedness of the naïve score test on $\boldsymbol{\beta}$. Dividing the second term in (A.83) by $\sigma_{\boldsymbol{\epsilon}}\sqrt{\boldsymbol{X}_j^\top(\boldsymbol{I} - \boldsymbol{P}^{(\check{\mathcal{A}}_\lambda\setminus\{j\})})\boldsymbol{X}_j}$, we get

$$\begin{aligned}\left|\frac{\boldsymbol{X}_j^\top\left(\boldsymbol{I} - \boldsymbol{P}^{(\check{\mathcal{A}}_\lambda\setminus\{j\})}\right)\boldsymbol{X}_{\setminus\check{\mathcal{A}}_\lambda\setminus\{j\}}\boldsymbol{\beta}_{\setminus\check{\mathcal{A}}_\lambda\setminus\{j\}}}{\sigma_{\boldsymbol{\epsilon}}\sqrt{\boldsymbol{X}_j^\top\left(\boldsymbol{I} - \boldsymbol{P}^{(\check{\mathcal{A}}_\lambda\setminus\{j\})}\right)\boldsymbol{X}_j}}\right| &= \left|\frac{\boldsymbol{r}^{s\top}\boldsymbol{X}_{\setminus\check{\mathcal{A}}_\lambda\setminus\{j\}}\boldsymbol{\beta}_{\setminus\check{\mathcal{A}}_\lambda\setminus\{j\}}}{\sigma_{\boldsymbol{\epsilon}}\,\|\boldsymbol{r}^s\|_2}\right| \\ &\leq \frac{\|\boldsymbol{r}^s\|_2}{\|\boldsymbol{r}^s\|_2}\frac{\left\|\boldsymbol{X}_{\setminus\check{\mathcal{A}}_\lambda\setminus\{j\}}\boldsymbol{\beta}_{\setminus\check{\mathcal{A}}_\lambda\setminus\{j\}}\right\|_2}{\sigma_{\boldsymbol{\epsilon}}} \\ &= \frac{\left\|\boldsymbol{X}_{\setminus\check{\mathcal{A}}_\lambda\setminus\{j\}}\boldsymbol{\beta}_{\setminus\check{\mathcal{A}}_\lambda\setminus\{j\}}\right\|_2}{\sigma_{\boldsymbol{\epsilon}}}.\end{aligned} \tag{A.86}$$

By (**M4***),

$$\begin{aligned}\left\|\boldsymbol{X}_{\setminus\check{\mathcal{A}}_\lambda\setminus\{j\}}\boldsymbol{\beta}_{\setminus\check{\mathcal{A}}_\lambda\setminus\{j\}}\right\|_2 = \left\|\boldsymbol{X}_{(\mathcal{A}\setminus\check{\mathcal{A}}_\lambda)\setminus\{j\}}\boldsymbol{\beta}_{(\mathcal{A}\setminus\check{\mathcal{A}}_\lambda)\setminus\{j\}}\right\|_2 &= \left\|\boldsymbol{X}_{\mathcal{A}\setminus\check{\mathcal{A}}_\lambda}\boldsymbol{\beta}_{\mathcal{A}\setminus\check{\mathcal{A}}_\lambda}\right\|_2 \\ &= \left\|\boldsymbol{X}_{\mathcal{A}\setminus(\check{\mathcal{A}}_\lambda\cup\mathcal{S})}\boldsymbol{\beta}_{\mathcal{A}\setminus(\check{\mathcal{A}}_\lambda\cup\mathcal{S})}\right\|_2 = o(1),\end{aligned}$$

where the second equality holds under $H_{0,j}: \beta_j = 0$, and the third equality is based on the fact that $\check{\mathcal{A}}_\lambda \supseteq \mathcal{S}$, proved in Lemma A.1.3. Thus, the naïve score test is asymptotically unbiased in testing $H_{0,j}: \beta_j = 0$.

Using, again, the fact that by Proposition 2.2.4, $\lim_{n\to\infty}\Pr\left[\check{\mathcal{A}}_\lambda = \hat{\mathcal{A}}_\lambda\right] = 1$, we get

$$\frac{\boldsymbol{X}_j^\top\left(\boldsymbol{I} - \boldsymbol{P}^{(\hat{\mathcal{A}}_\lambda\setminus\{j\})}\right)\boldsymbol{y}}{\sigma_{\boldsymbol{\epsilon}}\sqrt{\boldsymbol{X}_j^\top\left(\boldsymbol{I} - \boldsymbol{P}^{(\hat{\mathcal{A}}_\lambda\setminus\{j\})}\right)\boldsymbol{X}_j}} \to_p \frac{\boldsymbol{X}_j^\top\left(\boldsymbol{I} - \boldsymbol{P}^{(\check{\mathcal{A}}_\lambda\setminus\{j\})}\right)\boldsymbol{\epsilon}}{\sigma_{\boldsymbol{\epsilon}}\sqrt{\boldsymbol{X}_j^\top\left(\boldsymbol{I} - \boldsymbol{P}^{(\check{\mathcal{A}}_\lambda\setminus\{j\})}\right)\boldsymbol{X}_j}} \to_d \mathcal{N}(0,1). \tag{A.87}$$

□

A.7 Proof of Lemma 2.6.1

Proof of Lemma 2.6.1. First, according to the stationary conditions of (1.2),

$$\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda \right) = \lambda n \hat{\boldsymbol{\tau}}_{\lambda,\hat{\mathcal{A}}_\lambda}. \tag{A.88}$$

By definition, $\hat{\boldsymbol{\beta}}_{\lambda,\setminus\hat{\mathcal{A}}_\lambda} \equiv \mathbf{0}$. Thus, $\boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda = \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}\hat{\boldsymbol{\beta}}_{\lambda,\hat{\mathcal{A}}_\lambda}$. Rearranging terms,

$$\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top \boldsymbol{y} = \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}\hat{\boldsymbol{\beta}}_{\lambda,\hat{\mathcal{A}}_\lambda} + \lambda n \hat{\boldsymbol{\tau}}_{\lambda,\hat{\mathcal{A}}_\lambda}. \tag{A.89}$$

Multiplying (A.89) on the left by $(\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda})^{-1}$,

$$\bar{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda)} \equiv \left(\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda} \right)^{-1} \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top \boldsymbol{y} = \hat{\boldsymbol{\beta}}_{\lambda,\hat{\mathcal{A}}_\lambda} + \left(\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda} \right)^{-1} \lambda n \hat{\boldsymbol{\tau}}_{\lambda,\hat{\mathcal{A}}_\lambda}, \tag{A.90}$$

where $\bar{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda)}$ is the post-selection OLS estimator in (2.6). Plugging (A.88) into (A.90),

$$\begin{aligned}
\bar{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda)} &= \hat{\boldsymbol{\beta}}_{\lambda,\hat{\mathcal{A}}_\lambda} + \left(\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda} \right)^{-1} \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda \right) \\
&= \hat{\boldsymbol{\beta}}_{\lambda,\hat{\mathcal{A}}_\lambda} + \frac{1}{n} \left(\boldsymbol{G}_{(\hat{\mathcal{A}}_\lambda,\hat{\mathcal{A}}_\lambda)} \right)^{-1} \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda \right) \\
&= \hat{\boldsymbol{\beta}}_{\lambda,\hat{\mathcal{A}}_\lambda} + \frac{1}{n} \left[\begin{array}{cc} \left(\boldsymbol{G}_{(\hat{\mathcal{A}}_\lambda,\hat{\mathcal{A}}_\lambda)} \right)^{-1} & \mathbf{0} \end{array} \right] \boldsymbol{X}^\top \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda \right).
\end{aligned} \tag{A.91}$$

Combining (A.91) with the fact that $\hat{\boldsymbol{\beta}}_{\lambda,\setminus\hat{\mathcal{A}}_\lambda} \equiv \mathbf{0}$, we get

$$\left[\begin{array}{c} \bar{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda)} \\ \mathbf{0} \end{array} \right] = \left[\begin{array}{c} \hat{\boldsymbol{\beta}}_{\lambda,\hat{\mathcal{A}}_\lambda} \\ \hat{\boldsymbol{\beta}}_{\lambda,\setminus\hat{\mathcal{A}}_\lambda} \end{array} \right] + \frac{1}{n} \left[\begin{array}{cc} \left(\boldsymbol{G}_{(\hat{\mathcal{A}}_\lambda,\hat{\mathcal{A}}_\lambda)} \right)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{array} \right] \boldsymbol{X}^\top \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda \right). \tag{A.92}$$

Because we assume that, without loss of generality $\hat{\mathcal{A}}_\lambda = \{1, \ldots, |\hat{\mathcal{A}}_\lambda|\}$, we have

$$\left[\begin{array}{c} \bar{\boldsymbol{\beta}}^{(\hat{\mathcal{A}}_\lambda)} \\ \mathbf{0} \end{array} \right] = \hat{\boldsymbol{\beta}}_\lambda + \frac{1}{n} \left[\begin{array}{cc} \left(\boldsymbol{G}_{(\hat{\mathcal{A}}_\lambda,\hat{\mathcal{A}}_\lambda)} \right)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{array} \right] \boldsymbol{X}^\top \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda \right). \tag{A.93}$$

□

A.8 Proof of Lemma 2.6.2

Proof of Lemma 2.6.2. Based on Equation (21) of Tibshirani and Taylor (2012), under the null hypothesis $H_{0,j} : \beta_j = 0$, we have

$$\boldsymbol{P}^{(\hat{\mathcal{A}}_\lambda^{(\setminus j)})} \boldsymbol{y} = \boldsymbol{X}_{\setminus j} \hat{\boldsymbol{\beta}}_\lambda^{(\setminus j)} + n\lambda \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda^{(\setminus j)}} \left(\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda^{(\setminus j)}}^\top \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda^{(\setminus j)}} \right)^{-1} \hat{\boldsymbol{\tau}}_{\lambda,\hat{\mathcal{A}}_\lambda^{(\setminus j)}}^{(\setminus j)}, \tag{A.94}$$

where $\boldsymbol{P}^{(\hat{\mathcal{A}}_\lambda^{(\setminus j)})} \equiv \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda^{(\setminus j)}}(\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda^{(\setminus j)}}^\top \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda^{(\setminus j)}})^{-1}\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda^{(\setminus j)}}^\top$, $\hat{\boldsymbol{\beta}}_\lambda^{(\setminus j)}$ is the lasso estimator under the null hypothesis, defined in (2.28), and $\hat{\boldsymbol{\tau}}_{\lambda,\hat{\mathcal{A}}_\lambda^{(\setminus j)}}^{(\setminus j)}$ is defined based on the stationary condition of (2.28):

$$n\lambda\hat{\boldsymbol{\tau}}_{\lambda,\hat{\mathcal{A}}_\lambda^{(\setminus j)}}^{(\setminus j)} = \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda^{(\setminus j)}}^\top \left(\boldsymbol{y} - \boldsymbol{X}_{\setminus j}\hat{\boldsymbol{\beta}}_\lambda^{(\setminus j)}\right). \tag{A.95}$$

Equation (A.94) implies that

$$\boldsymbol{X}_j^\top \left(\boldsymbol{I} - \boldsymbol{P}^{(\hat{\mathcal{A}}_\lambda^{(\setminus j)})}\right)\boldsymbol{y} = \boldsymbol{X}_j^\top \left(\boldsymbol{y} - \boldsymbol{X}_{\setminus j}\hat{\boldsymbol{\beta}}_\lambda^{(\setminus j)}\right) - \boldsymbol{X}_j^\top \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda^{(\setminus j)}} \left(\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda^{(\setminus j)}}^\top \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda^{(\setminus j)}}\right)^{-1} n\lambda\hat{\boldsymbol{\tau}}_{\lambda,\hat{\mathcal{A}}_\lambda^{(\setminus j)}}^{(\setminus j)}. \tag{A.96}$$

Plugging (A.95) into (A.96),

$$\begin{aligned} S^0 &\equiv \boldsymbol{X}_j^\top \left(\boldsymbol{I} - \boldsymbol{P}^{(\hat{\mathcal{A}}_\lambda^{(\setminus j)})}\right)\boldsymbol{y} \\ &= \boldsymbol{X}_j^\top \left(\boldsymbol{y} - \boldsymbol{X}_{\setminus j}\hat{\boldsymbol{\beta}}_\lambda^{(\setminus j)}\right) - \boldsymbol{X}_j^\top \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda^{(\setminus j)}}(\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda^{(\setminus j)}}^\top \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda^{(\setminus j)}})^{-1}\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda^{(\setminus j)}}^\top \left(\boldsymbol{y} - \boldsymbol{X}_{\setminus j}\hat{\boldsymbol{\beta}}_\lambda^{(\setminus j)}\right) \\ &= \boldsymbol{X}_j^\top \left(\boldsymbol{y} - \boldsymbol{X}_{\setminus j}\hat{\boldsymbol{\beta}}_\lambda^{(\setminus j)}\right) - \frac{1}{n}\boldsymbol{X}_j^\top \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda^{(\setminus j)}} \left(\boldsymbol{G}_{(\hat{\mathcal{A}}_\lambda^{(\setminus j)},\hat{\mathcal{A}}_\lambda^{(\setminus j)})}\right)^{-1} \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda^{(\setminus j)}}^\top \left(\boldsymbol{y} - \boldsymbol{X}_{\setminus j}\hat{\boldsymbol{\beta}}_\lambda^{(\setminus j)}\right). \end{aligned} \tag{A.97}$$

□

Appendix B

TECHNICAL DETAILS FOR "A SIGNIFICANCE TEST FOR GRAPH-CONSTRAINED ESTIMATION"

B.1 Proof of Lemma 3.2.1

Proof of Lemma 3.2.1. Denoting the scaled Gramian matrix $\boldsymbol{G} \equiv \boldsymbol{X}^\top \boldsymbol{X}/n$. Given that $(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L})$ is invertible and $\lambda_{\boldsymbol{L}} > 0$,

$$\begin{aligned}
\mathbf{Bias}\left[\tilde{\boldsymbol{\beta}}^{\boldsymbol{L}}_{\lambda_{\boldsymbol{L}}} | \boldsymbol{X}\right] &= \mathrm{E}\left(\tilde{\boldsymbol{\beta}}^{\boldsymbol{L}}_{\lambda_{\boldsymbol{L}}} | \boldsymbol{X}\right) - \boldsymbol{\beta} \\
&= \left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1} n\boldsymbol{G}\boldsymbol{\beta} - \left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1}\left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right)\boldsymbol{\beta} \\
&= -\left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1} \lambda_{\boldsymbol{L}}\boldsymbol{L}\boldsymbol{\beta}, \qquad \text{(B.1)}
\end{aligned}$$

which is equal to $\mathbf{0}$ if and only if $\boldsymbol{L}\boldsymbol{\beta} = \mathbf{0}$ because $(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L})$ is invertible. Based on the properties of eigenvalues,

$$\left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1} \preceq \frac{1}{\phi^2_{\min}\left[n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right]}\boldsymbol{I}. \qquad \text{(B.2)}$$

Therefore,

$$\begin{aligned}
\left\|\mathbf{Bias}\left[\tilde{\boldsymbol{\beta}}^{\boldsymbol{L}}_{\lambda_{\boldsymbol{L}}} | \boldsymbol{X}\right]\right\|_2 &= \lambda_{\boldsymbol{L}} \sqrt{\left(\boldsymbol{L}\boldsymbol{\beta}\right)^\top \left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-2} \left(\boldsymbol{L}\boldsymbol{\beta}\right)} \\
&\leq \lambda_{\boldsymbol{L}} \sqrt{\left(\boldsymbol{L}\boldsymbol{\beta}\right)^\top \frac{1}{\phi^4_{\min}\left[n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right]} \left(\boldsymbol{L}\boldsymbol{\beta}\right)} \\
&= \frac{\lambda_{\boldsymbol{L}} \left\|\boldsymbol{L}\boldsymbol{\beta}\right\|_2}{\phi^2_{\min}\left[n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right]}. \qquad \text{(B.3)}
\end{aligned}$$

□

B.2 Proof of Theorem 3.2.2

Proof of Theorem 3.2.2. Let $\hat{\boldsymbol{\beta}}_\lambda$ be the lasso initial estimator with tuning parameter λ, defined in (1.2). We first consider the case without condition (**O**). The bias of the test statistic

$$\begin{aligned}\left|\gamma_j^G\right| &\equiv \lambda_{\boldsymbol{L}} \left|(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L})^{-1}\boldsymbol{L}\left(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\right)\right|_j \\ &\leq \lambda_{\boldsymbol{L}} \left\|\left[(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L})^{-1}\boldsymbol{L}\right]_j\right\|_\infty \left\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\right\|_1 . \end{aligned} \tag{B.4}$$

Based on Bühlmann and van de Geer (2011), Chapter 6.12, with Gaussian design, if the $(\mathcal{A}, 3, \boldsymbol{\Sigma})$-compatibility condition is met with compatibility constant ϕ^2 as in (**E**) and $q = o((n/\log(p))^\xi)$, $\xi \leq 0.5$ as in (**A2**), then with probability tending to one, the condition is also met for $\boldsymbol{G}$; also see Raskutti et al. (2010), Rudelson and Zhou (2013) for additional references. In that case, with $\lambda \asymp \sqrt{\log(p)/n}$,

$$\lim_{n\to\infty} \Pr\left[\left\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\right\|_1 \leq 16\frac{\lambda q}{\phi^2}\right] = 1. \tag{B.5}$$

Then, because $q = o((n/\log(p))^\xi)$, $\lambda \asymp \sqrt{\log(p)/n}$ and $\liminf_{n\to\infty} \phi^2 > 0$, we get

$$\left\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\right\|_1 = o_p\left(\left(\frac{\log(p)}{n}\right)^{\frac{1}{2}-\xi}\right). \tag{B.6}$$

Thus, without condition (**O**),

$$\lim_{n\to\infty} \Pr\left[\left|\gamma_j^G\right| \leq \lambda_{\boldsymbol{L}} \left\|\left[(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L})^{-1}\boldsymbol{L}\right]_j\right\|_\infty \left(\frac{\log(p)}{n}\right)^{\frac{1}{2}-\xi}\right] = 1. \tag{B.7}$$

Hence, with probability tending to one,

$$\begin{aligned}\left|z_j^G\right| = \left|Z_j^G + \gamma_j^G\right| &\leq \left|Z_j^G\right| + \left|\gamma_j^G\right| \\ &\precsim \left|Z_j^G\right| + \lambda_{\boldsymbol{L}} \left\|\left[(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L})^{-1}\boldsymbol{L}\right]_j\right\|_\infty \left(\frac{\log(p)}{n}\right)^{\frac{1}{2}-\xi}. \end{aligned} \tag{B.8}$$

On the other hand, with condition (**O**),

$$
\begin{aligned}
\left|\gamma_j^G\right| &\equiv \lambda_{\boldsymbol{L}} \left| \left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1} \boldsymbol{L} \left(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\right) \right|_j \\
&= \lambda_{\boldsymbol{L}} \left| \sum_{k=1}^{p} \left[\left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1} \boldsymbol{L}\right]_{(j,k)} \left(\hat{\beta}_{\lambda,k} - \beta_k\right) \right| \\
&= \lambda_{\boldsymbol{L}} \left| \sum_{k:k\neq j} \left[\left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1} \boldsymbol{L}\right]_{(j,k)} \left(\hat{\beta}_{\lambda,k} - \beta_k\right) + \left[\left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1} \boldsymbol{L}\right]_{(j,j)} \left(\hat{\beta}_{\lambda,j} - \beta_j\right) \right| \\
&\leq \lambda_{\boldsymbol{L}} \left| \sum_{k:k\neq j} \left[\left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1} \boldsymbol{L}\right]_{(j,k)} \left(\hat{\beta}_{\lambda,k} - \beta_k\right) \right| \\
&\quad + \lambda_{\boldsymbol{L}} \left| \left[\left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1} \boldsymbol{L}\right]_{(j,j)} \left(\hat{\beta}_{\lambda,j} - \beta_j\right) \right| \\
&\leq \lambda_{\boldsymbol{L}} \left\| \left[\left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1} \boldsymbol{L}\right]_{(j,\setminus j)} \right\|_\infty \left\| \hat{\boldsymbol{\beta}}_{\lambda,-j} - \boldsymbol{\beta}_{\setminus j} \right\|_1 \\
&\quad + \lambda_{\boldsymbol{L}} \left| \left[\left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1} \boldsymbol{L}\right]_{(j,j)} \left(\hat{\beta}_{\lambda,j} - \beta_j\right) \right|. \\
&\leq \lambda_{\boldsymbol{L}} \left\| \left[\left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1} \boldsymbol{L}\right]_{(j,\setminus j)} \right\|_\infty \left\| \hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta} \right\|_1 \\
&\quad + \lambda_{\boldsymbol{L}} \left| \left[\left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1} \boldsymbol{L}\right]_{(j,j)} \left(\hat{\beta}_{\lambda,j} - \beta_j\right) \right|. \qquad \text{(B.9)}
\end{aligned}
$$

Condition (**O**) states that

$$
\frac{\left[\lambda_{\boldsymbol{L}} \left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1} \boldsymbol{L}\right]_{(j,j)}}{\sqrt{n\sigma_{\boldsymbol{\epsilon}}^2 \left[\left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1} \boldsymbol{G} \left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1}\right]_{(j,j)}}} = \mathcal{O}_p\left(\left(\frac{n}{\log(p)}\right)^{\frac{1}{2}-\xi}\right).
$$

We now bound both terms in (B.9). For the first term, since $\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\|_1 = \mathcal{o}_p((\log(p)/n)^{0.5-\xi})$ by (B.6),

$$
\lambda_{\boldsymbol{L}} \left\| \left[\left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1} \boldsymbol{L}\right]_{(j,\setminus j)} \right\|_\infty \left\| \hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta} \right\|_1 = \lambda_{\boldsymbol{L}} \left\| \left[\left(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1} \boldsymbol{L}\right]_{(j,\setminus j)} \right\|_\infty \cdot \mathcal{o}_p\left(\left(\frac{\log(p)}{n}\right)^{\frac{1}{2}-\xi}\right). \qquad \text{(B.10)}
$$

For the second term, (B.9) also implies that $|\hat{\beta}_{\lambda,j} - \beta_j| = \mathcal{o}_p((\log(p)/n)^{0.5-\xi})$. Thus,

$$\begin{aligned}
\lambda_{\boldsymbol{L}} \left| \left[(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L})^{-1} \boldsymbol{L} \right]_{(j,j)} \left(\hat{\beta}_{\lambda,j} - \beta_j \right) \right| &= \lambda_{\boldsymbol{L}} \left| \left[(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L})^{-1} \boldsymbol{L} \right]_{(j,j)} \right| \left| \hat{\beta}_{\lambda,j} - \beta_j \right| \\
&= \sqrt{n\sigma_{\boldsymbol{\epsilon}}^2 \left[(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L})^{-1} \boldsymbol{G} (n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L})^{-1} \right]_{(j,j)}} \\
&\quad \cdot \mathcal{O}_p \left(\left(\frac{n}{\log(p)} \right)^{\frac{1}{2}-\xi} \right) \cdot \mathcal{o}_p \left(\left(\frac{\log(p)}{n} \right)^{\frac{1}{2}-\xi} \right) \\
&= \sqrt{n\sigma_{\boldsymbol{\epsilon}}^2 \left[(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L})^{-1} \boldsymbol{G} (n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L})^{-1} \right]_{(j,j)}} \cdot \mathcal{o}_p(1),
\end{aligned} \tag{B.11}$$

where the second equality is based on Condition (**O**). Since by (3.6),

$$|Z_j^G| = \sqrt{n\sigma_{\boldsymbol{\epsilon}}^2 \left[(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L})^{-1} \boldsymbol{G} (n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L})^{-1} \right]_{(j,j)}} \cdot \mathcal{O}_p(1),$$

we have $|Z_j^G|$ dominating $\lambda_{\boldsymbol{L}}|[(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L})^{-1}\boldsymbol{L}]_{(j,j)}(\hat{\beta}_{\lambda,j} - \beta_j)|$. Therefore,

$$\begin{aligned}
\left| z_j^G \right| &= \left| Z_j^G + \gamma_j^G \right| \\
&\leq \left| Z_j^G \right| + \left| \gamma_j^G \right| \\
&\leq \left| Z_j^G \right| + \lambda_{\boldsymbol{L}} \left\| \left[(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L})^{-1} \boldsymbol{L} \right]_{(j,\setminus j)} \right\|_\infty \left\| \hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta} \right\|_1 + \lambda_{\boldsymbol{L}} \left| \left[(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L})^{-1} \boldsymbol{L} \right]_{(j,j)} \left(\hat{\beta}_{\lambda,j} - \beta_j \right) \right| \\
&\asymp \left| Z_j^G \right| + \lambda_{\boldsymbol{L}} \left\| \left[(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L})^{-1} \boldsymbol{L} \right]_{(j,\setminus j)} \right\|_\infty \left\| \hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta} \right\|_1 \\
&\precsim \left| Z_j^G \right| + \lambda_{\boldsymbol{L}} \left\| \left[(n\boldsymbol{G} + \lambda_{\boldsymbol{L}}\boldsymbol{L})^{-1} \boldsymbol{L} \right]_{(j,\setminus j)} \right\|_\infty \left(\frac{\log(p)}{n} \right)^{\frac{1}{2}-\xi} \equiv \left| Z_j^G \right| + \Gamma_j^G,
\end{aligned} \tag{B.12}$$

which completes the proof. □

B.3 Proof of Theorem 3.3.1

Proof of Theorem 3.3.1. Given (3.13), conditional on $\boldsymbol{X}$, $p_j^G \leq \alpha$ if

$$\left| z_j^G \right| \geq \Gamma_j^G + \Phi_{\mathcal{N}}^{-1} \left[1 - \frac{\alpha}{2} \right] \sqrt{\mathrm{Var}\left[Z_j^G | \boldsymbol{X} \right]}, \tag{B.13}$$

where $\Phi_{\mathcal{N}}^{-1}[\cdot]$ is the standard normal quantile function. According to (3.5), this is equivalent of

$$\left| \beta_j + Z_j^G + \gamma_j^G \right| \geq \Gamma_j^G + \Phi_{\mathcal{N}}^{-1} \left[1 - \frac{\alpha}{2} \right] \sqrt{\mathrm{Var}\left[Z_j^G | \boldsymbol{X} \right]}, \tag{B.14}$$

which is satisfied if

$$|\beta_j| - \left|\gamma_j^G\right| \geq \left|Z_j^G\right| + \Gamma_j^G + \Phi_{\mathcal{N}}^{-1}\left[1-\frac{\alpha}{2}\right]\sqrt{\mathrm{Var}\left[Z_j^G|\boldsymbol{X}\right]}. \tag{B.15}$$

Because $Z_j^G|\boldsymbol{X}$ is a normal random variable,

$$\Pr\left[\left|Z_j^G\right| \leq \Phi_{\mathcal{N}}^{-1}\left[1-\frac{\psi}{2}\right]\sqrt{\mathrm{Var}\left[Z_j^G|\boldsymbol{X}\right]}\,\middle|\,\boldsymbol{X}\right] = 1-\psi.$$

Therefore, (B.15) is satisfied with probability $1-\psi$ if

$$|\beta_j| - \left|\gamma_j^G\right| \geq \Gamma_j^G + \left(\Phi_{\mathcal{N}}^{-1}\left[1-\frac{\alpha}{2}\right] + \Phi_{\mathcal{N}}^{-1}\left[1-\frac{\psi}{2}\right]\right)\sqrt{\mathrm{Var}\left[Z_j^G|\boldsymbol{X}\right]}. \tag{B.16}$$

Based on Theorem 3.2.2, given Condition (**O**), with probability tending to one, $|\gamma_j^G| \precsim \Gamma_j^G + \sqrt{\mathrm{Var}[Z_j^G|\boldsymbol{X}]}\cdot\mathcal{O}_p(1)$. Therefore, conditional on $\boldsymbol{X}$, we have $p_j^G \leq \alpha$ with probability tending to at least $1-\psi$, if

$$|\beta_j| > 2\Gamma_j^G + \left(\Phi_{\mathcal{N}}^{-1}\left[1-\frac{\alpha}{2}\right] + \Phi_{\mathcal{N}}^{-1}\left[1-\frac{\psi}{2}\right]\right)\sqrt{\mathrm{Var}\left[Z_j^G|\boldsymbol{X}\right]}. \tag{B.17}$$

□

B.4 Proof of Theorem 3.3.2

Proof of Theorem 3.3.2. **a)** First, note that $p_1^G/p_1^{GI} \leq 1$ is equivalent of

$$\frac{\left(\left|z_1^{GI}\right| - \Gamma_1^{GI}\right)_+ / \sqrt{\mathrm{Var}\left[Z_1^{GI}|\boldsymbol{X}\right]}}{\left(|z_1^G| - \Gamma_1^G\right)_+ / \sqrt{\mathrm{Var}\left[Z_1^G|\boldsymbol{X}\right]}} \leq 1. \tag{B.18}$$

Let $\hat{\boldsymbol{\beta}}_\lambda$ be the initial lasso estimator with tuning parameter λ. We first write out those components for the Grace test:

$$\begin{aligned} z_1^G &= \left[\left(\boldsymbol{X}^\top\boldsymbol{X}+\lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1}\left(\boldsymbol{X}^\top\boldsymbol{y}+\lambda_{\boldsymbol{L}}\boldsymbol{L}\hat{\boldsymbol{\beta}}_\lambda\right)\right]_1 \\ &= \frac{(n+\lambda_{\boldsymbol{L}})\,\boldsymbol{X}_1^\top\boldsymbol{y}-(n\rho+\lambda_{\boldsymbol{L}}l)\,\boldsymbol{X}_2^\top\boldsymbol{y}+\lambda_{\boldsymbol{L}}\hat{\beta}_{\lambda,1}\,(n+\lambda_{\boldsymbol{L}}-n\rho l-\lambda_{\boldsymbol{L}}l^2)+n\lambda_{\boldsymbol{L}}\hat{\beta}_{\lambda,2}\,(l-\rho)}{(n+\lambda_{\boldsymbol{L}})^2-(n\rho+\lambda_{\boldsymbol{L}}l)^2}; \end{aligned} \tag{B.19}$$

$$\begin{aligned} \Gamma_1^G &= \left|\lambda_{\boldsymbol{L}}\left[\left(\boldsymbol{X}^\top\boldsymbol{X}+\lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1}\boldsymbol{L}\right]_{(1,-1)}\right|\left(\frac{\log(p)}{n}\right)^{\frac{1}{2}-\xi} \\ &= \left|\lambda_{\boldsymbol{L}}\left[\left(\boldsymbol{X}^\top\boldsymbol{X}+\lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1}\boldsymbol{L}\right]_{(1,2)}\right|\left(\frac{\log(p)}{n}\right)^{\frac{1}{2}-\xi} \\ &= \frac{|n\lambda_{\boldsymbol{L}}l-n\lambda_{\boldsymbol{L}}\rho|}{(n+\lambda_{\boldsymbol{L}})^2-(n\rho+\lambda_{\boldsymbol{L}}l)^2}\left(\frac{\log(p)}{n}\right)^{\frac{1}{2}-\xi}; \end{aligned} \tag{B.20}$$

$$\begin{aligned} \mathrm{Var}\left[Z_1^G|\boldsymbol{X}\right] &= \sigma_{\boldsymbol{\epsilon}}^2\left[\left(\boldsymbol{X}^\top\boldsymbol{X}+\lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}+\lambda_{\boldsymbol{L}}\boldsymbol{L}\right)^{-1}\right]_{(1,1)} \\ &= \sigma_{\boldsymbol{\epsilon}}^2\frac{(n^3+2\lambda_{\boldsymbol{L}}n^2)\,(1-\rho^2)+n\lambda_{\boldsymbol{L}}^2\,(1+l^2-2l\rho)}{\left((n+\lambda_{\boldsymbol{L}})^2-(n\rho+\lambda_{\boldsymbol{L}}l)^2\right)^2}. \end{aligned} \tag{B.21}$$

We can also write out those components for the GraceI test likewise with $l=0$.

Equation (B.6) shows that $\|\hat{\boldsymbol{\beta}}_\lambda-\boldsymbol{\beta}\|_1=\mathcal{O}_p((\log(p)/n)^{0.5-\xi})$. With $p=\mathcal{O}(\exp(n^\nu))$ for some $0\le\nu<1$, $\|\hat{\boldsymbol{\beta}}_\lambda-\boldsymbol{\beta}\|_1=\mathcal{O}_p(1)$. Thus,

$$\hat{\beta}_{\lambda,1}=\beta_1+\mathcal{O}_p(1),\qquad \hat{\beta}_{\lambda,2}=\beta_2+\mathcal{O}_p(1). \tag{B.22}$$

Since our design matrix is centered and scaled, i.e., $\boldsymbol{X}_1^\top\boldsymbol{X}_1=\boldsymbol{X}_2^\top\boldsymbol{X}_2=n$, $\boldsymbol{X}_1^\top\boldsymbol{X}_2=n\rho$,

$$\boldsymbol{X}_1^\top\boldsymbol{y}=\boldsymbol{X}_1^\top\boldsymbol{X}_1\beta_1+\boldsymbol{X}_1^\top\boldsymbol{X}_2\beta_2+\boldsymbol{X}_1^\top\boldsymbol{\epsilon}=n\beta_1+n\rho\beta_2+ne_1, \tag{B.23}$$

$$\boldsymbol{X}_2^\top\boldsymbol{y}=\boldsymbol{X}_2^\top\boldsymbol{X}_1\beta_1+\boldsymbol{X}_2^\top\boldsymbol{X}_2\beta_2+\boldsymbol{X}_2^\top\boldsymbol{\epsilon}=n\rho\beta_1+n\beta_2+ne_2, \tag{B.24}$$

where $\boldsymbol{e}\sim\mathcal{N}_2(\boldsymbol{0},\sigma_{\boldsymbol{\epsilon}}^2/n\boldsymbol{I}_2)=\mathcal{O}_p(1)$.

Define $k_{\boldsymbol{L}}\equiv\lambda_{\boldsymbol{L}}/n$ and $k_{\boldsymbol{I}}\equiv\lambda_{\boldsymbol{I}}/n$.

$$\frac{\left(\left|z_1^G\right|-\Gamma_1^G\right)_+}{\sqrt{\mathrm{Var}\left[Z_1^G|\boldsymbol{X}\right]}}=\frac{\sqrt{n}\left(\left|(k_{\boldsymbol{L}}+1)^2-(\rho+lk_{\boldsymbol{L}})^2+\mathcal{O}_p(1)\right|\,|\beta_1|-(\log(p)/n)^{1/2-\xi}\,|k_{\boldsymbol{L}}\,(l-\rho)|\right)_+}{\sigma_{\boldsymbol{\epsilon}}\sqrt{(1+2k_{\boldsymbol{L}})\,(1-\rho^2)+(k_{\boldsymbol{L}})^2\,(1+l^2-2l\rho)}}. \tag{B.25}$$

Similarly for the GraceI, we get

$$\frac{\left(\left|z_1^{GI}\right| - \Gamma_1^{GI}\right)_+}{\sqrt{\mathrm{Var}\left[Z_1^{GI}|\boldsymbol{X}\right]}} = \frac{\sqrt{n}\left(\left|\left(k_{\boldsymbol{I}}+1\right)^2 - \rho^2 + \mathcal{O}_p(1)\right|\left|\beta_1\right| - \left(\log(p)/n\right)^{1/2-\xi}\left|k_{\boldsymbol{I}}\rho\right|\right)_+}{\sigma_{\boldsymbol{\epsilon}}\sqrt{\left(1+2k_{\boldsymbol{I}}\right)\left(1-\rho^2\right)+k_{\boldsymbol{I}}^2}}. \tag{B.26}$$

Since $|\rho| \leq 1$, $|l| \leq 1$, $k_{\boldsymbol{L}} > 0$ and $k_{\boldsymbol{I}} > 0$, $(k_{\boldsymbol{L}}+1)^2 - (\rho + lk_{\boldsymbol{L}})^2 > 0$ and $(k_{\boldsymbol{I}}+1)^2 - \rho^2 > 0$. Therefore, based on (B.25) and (B.26), conditional on the design matrix $\boldsymbol{X}$, $p_1^G/p_1^{GI} \leq 1$ with probability tending to one if

$$\begin{aligned}
&\lim_{n\to\infty} \frac{\left(\left(\left(k_{\boldsymbol{L}}+1\right)^2 - \left(\rho + lk_{\boldsymbol{L}}\right)^2\right)\left|\beta_1\right| - \left(\log(p)/n\right)^{1/2-\xi}\left|k_{\boldsymbol{L}}\left(l-\rho\right)\right|\right)_+}{\sqrt{\left(1+2k_{\boldsymbol{L}}\right)\left(1-\rho^2\right)+k_{\boldsymbol{L}}^2\left(1+l^2-2l\rho\right)}} \\
&\geq \lim_{n\to\infty} \frac{\left(\left(\left(k_{\boldsymbol{I}}+1\right)^2 - \rho^2\right)\left|\beta_1\right| - \left(\log(p)/n\right)^{1/2-\xi}\left|k_{\boldsymbol{I}}\rho\right|\right)_+}{\sqrt{\left(1+2k_{\boldsymbol{I}}\right)\left(1-\rho^2\right)+k_{\boldsymbol{I}}^2}}.
\end{aligned} \tag{B.27}$$

For any two real numbers a and b, $a \geq b$ implies $a_+ \geq b_+$. Thus, conditional on the design matrix $\boldsymbol{X}$, $p_1^G/p_1^{GI} \leq 1$ with probability tending to one if

$$\begin{aligned}
&\lim_{n\to\infty} \frac{\left(\left(k_{\boldsymbol{L}}+1\right)^2 - \left(\rho + lk_{\boldsymbol{L}}\right)^2\right)\left|\beta_1\right| - \left(\log(p)/n\right)^{1/2-\xi}\left|k_{\boldsymbol{L}}\left(l-\rho\right)\right|}{\sqrt{\left(1+2k_{\boldsymbol{L}}\right)\left(1-\rho^2\right)+k_{\boldsymbol{L}}^2\left(1+l^2-2l\rho\right)}} \\
&\geq \lim_{n\to\infty} \frac{\left(\left(k_{\boldsymbol{I}}+1\right)^2 - \rho^2\right)\left|\beta_1\right| - \left(\log(p)/n\right)^{1/2-\xi}\left|k_{\boldsymbol{I}}\rho\right|}{\sqrt{\left(1+2k_{\boldsymbol{I}}\right)\left(1-\rho^2\right)+k_{\boldsymbol{I}}^2}}.
\end{aligned} \tag{B.28}$$

If we assume $k_{\boldsymbol{L}} = k_{\boldsymbol{I}} = k \to \infty$, Inequality (B.28) is satisfied if

$$\begin{aligned}
1 &\leq \lim_{n\to\infty} \frac{\left(\left(k+1\right)^2 - \left(\rho + lk\right)^2\right)\left|\beta_1\right| - \left(\log(p)/n\right)^{1/2-\xi}\left|k\left(l-\rho\right)\right|}{\left(\left(k+1\right)^2 - \rho^2\right)\left|\beta_1\right| - \left(\log(p)/n\right)^{1/2-\xi}\left|k\rho\right|} \\
&\quad \times \frac{\sqrt{\left(1+2k\right)\left(1-\rho^2\right)+k^2}}{\sqrt{\left(1+2k\right)\left(1-\rho^2\right)+k^2\left(1+l^2-2l\rho\right)}} \\
&= \lim_{n\to\infty} \frac{\left(\left(1-l^2\right)+\left(2-2l\rho\right)/k+\left(1-\rho^2\right)/k^2\right)\left|\beta_1\right| - \left(\log(p)/n\right)^{1/2-\xi}\left|\left(l-\rho\right)/k\right|}{\left(1+2/k+\left(1-\rho^2\right)/k^2\right)\left|\beta_1\right| - \left(\log(p)/n\right)^{1/2-\xi}\left|\rho/k\right|} \\
&\quad \times \frac{\sqrt{1+\left(2-2\rho^2\right)/k+\left(1-\rho^2\right)/k^2}}{\sqrt{\left(1+l^2-2l\rho\right)+\left(2-2\rho^2\right)/k+\left(1-\rho^2\right)/k^2}} \\
&= \frac{\left(1-l^2\right)}{\sqrt{\left(1+l^2-2l\rho\right)}}.
\end{aligned} \tag{B.29}$$

The last equality holds because $p = \mathcal{O}(\exp(n^\nu))$ for some $0 \leq \nu < 1$ implies that $\log(p)/n \to 0$.

For the ridge test, we assume its tuning parameter has $\mathcal{O}(1)$. Thus, we can similarly write out the ridge test objective:

$$\frac{\left|z_1^R\right|}{\sqrt{\text{Var}\left[Z_1^R|\boldsymbol{X}\right]}} = \frac{\sqrt{n}\left|1-\rho^2+\mathcal{O}_p\left(1\right)\right|\left|\beta_1\right|}{\sigma_{\boldsymbol{\epsilon}}\sqrt{\left(1-\rho^2\right)+\mathcal{O}\left(1\right)}}. \tag{B.30}$$

b) Thus, conditional on $\boldsymbol{X}$, we get $p_1^G/p_1^R \leq 1$ with probability tending to one if

$$\lim_{n\to\infty} \frac{\left(\left(k_{\boldsymbol{L}}+1\right)^2-\left(\rho+lk_{\boldsymbol{L}}\right)^2\right)\left|\beta_1\right|-\left(\log(p)/n\right)^{1/2-\xi}\left|k_{\boldsymbol{L}}\left(l-\rho\right)\right|}{\sqrt{\left(1+2k_{\boldsymbol{L}}\right)\left(1-\rho^2\right)+k_{\boldsymbol{L}}^2\left(1+l^2-2l\rho\right)}} \geq \sqrt{1-\rho^2}\left|\beta_1\right|. \tag{B.31}$$

c) We also have $p_1^{GI}/p_1^R \leq 1$ with probability tending to one if

$$\lim_{n\to\infty} \frac{\left(\left(k_{\boldsymbol{I}}+1\right)^2-\rho^2\right)\left|\beta_1\right|-\left(\log(p)/n\right)^{1/2-\xi}\left|k_{\boldsymbol{I}}\rho\right|}{\sqrt{\left(1+2k_{\boldsymbol{I}}\right)\left(1-\rho^2\right)+k_{\boldsymbol{I}}^2}} \geq \sqrt{1-\rho^2}\left|\beta_1\right|. \tag{B.32}$$

□

Appendix C

TECHNICAL DETAILS FOR "DIFFERENTIAL CONNECTIVITY ANALYSIS: TESTING DIFFERENCES IN STRUCTURES OF HIGH-DIMENSIONAL NETWORKS"

C.1 Proof of Proposition 4.4.1

In this section, we prove that conditions (**A1**) and (**A2**) for variable $j \in \mathcal{V}$ imply

$$\lim_{n\to\infty} \Pr\left[\hat{ne}_j^{\mathrm{I}} \supseteq ne_j^{\mathrm{I}}\right] = 1,$$

where $\hat{ne}_j^{\mathrm{I}} \equiv \mathrm{supp}(\hat{\boldsymbol{\beta}}^{\mathrm{I},j})$, with $\hat{\boldsymbol{\beta}}^{\mathrm{I},j}$ defined in (4.9). The result $\lim_{n\to\infty} \Pr[\hat{ne}_j^{\mathrm{II}} \supseteq ne_j^{\mathrm{II}}] = 1$ can be proved using exactly the same procedure and is thus omitted. Since $\hat{ne}_j^0 \equiv \hat{ne}_j^{\mathrm{I}} \cap \hat{ne}_j^{\mathrm{II}}$ and $ne_j^0 \equiv ne_j^{\mathrm{I}} \cap ne_j^{\mathrm{II}}$, the above two results imply

$$\lim_{n\to\infty} \Pr\left[\hat{ne}_j^0 \supseteq ne_j^0\right] = 1,$$

Because our proof only concerns population I, for simplicity, we omit the superscript "I" from the subsequent proofs.

We first prove Lemma C.1.1, which shows that under (**A1**) and (**A2**), the $(ne_j, q_j, 3)$-restricted eigenvalue condition (Bickel et al., 2009, van de Geer and Bühlmann, 2009) is satisfied for each $j \in \mathcal{V}$.

Lemma C.1.1. *Suppose (**A1**) and (**A2**) for variable $j \in \mathcal{V}$ hold. Suppose $q_j \equiv |ne_j| = |\mathrm{supp}(\boldsymbol{\beta}^j)| \geq 1$. For all $\boldsymbol{b} \in \mathbb{R}^{p-1}$ and any index set $\mathcal{I}$, such that $|\mathcal{I}| \leq q_j$, $\|\boldsymbol{b}_{\setminus\mathcal{I}}\|_1 \leq 3\|\boldsymbol{b}_{\mathcal{I}}\|_1$ and $\|\boldsymbol{b}_{\setminus\mathcal{S}}\|_\infty \leq \min_{j\in(\mathcal{S}\setminus\mathcal{I})} |b_j|$, where $\mathcal{S}$ is any index set such that $\mathcal{S} \supseteq \mathcal{I}$ and $|\mathcal{S}| \leq 2q_j$, we have*

$$\lim_{n\to\infty} \Pr\left[\|\boldsymbol{b}_{\mathcal{S}}\|_2^2 \leq \boldsymbol{b}^\top \boldsymbol{G}_{(\setminus j,\setminus j)} \boldsymbol{b} \frac{1}{\phi^2}\right] = 1, \tag{C.1}$$

where $\boldsymbol{G}_{(\setminus j,\setminus j)} \equiv \boldsymbol{X}_{\setminus j}^\top \boldsymbol{X}_{\setminus j}/n$ and $\liminf_{n\to\infty} \phi^2 = \kappa^2 > 0$.

Proof. Based on, e.g., Corollary 1 in Raskutti et al. (2010) and Theorem 6 in Rudelson and Zhou (2013), with Gaussian data, (C.1) holds if:

(**C1**) $\|\boldsymbol{b}_{\mathcal{S}}\|_2^2 \leq \boldsymbol{b}^\top \boldsymbol{\Sigma}_{(\backslash j,\backslash j)} \boldsymbol{b}/(64\phi^2)$ with $\liminf_{n\to\infty} \phi^2 > 0$;

(**C2**) For any $\boldsymbol{v} \in \mathbb{R}^{p-1}$ such that $\|\boldsymbol{v}\|_2 = 1$ and $|\text{supp}(\boldsymbol{v})| = 1$, we have $\boldsymbol{v}^\top \boldsymbol{\Sigma}_{(\backslash j,\backslash j)} \boldsymbol{v} = \mathcal{O}(1)$;

(**C3**) $q_j \log(p)/n \to 0$.

We now proceed to show that these three requirements hold.

(**C1**) For any $\boldsymbol{b} \in \mathbb{R}^{p-1}$, we have $\boldsymbol{b}^\top \boldsymbol{\Sigma}_{(\backslash j,\backslash j)} \boldsymbol{b}/\|\boldsymbol{b}\|_2^2 \geq \phi^2_{\min}[\boldsymbol{\Sigma}_{(\backslash j,\backslash j)}] \geq \phi^2_{\min}[\boldsymbol{\Sigma}] > 0$. The inequality $\phi^2_{\min}[\boldsymbol{\Sigma}_{(\backslash j,\backslash j)}] \geq \phi^2_{\min}[\boldsymbol{\Sigma}]$ is based on the interlacing property of eigenvalues of principal sub-matrices (see, e.g., Theorem 2.1 in Haemers, 1995), while the last equality is guaranteed by (**A1**). Thus, for any $\mathcal{S} \subseteq \mathcal{V}$, we have

$$\|\boldsymbol{b}_{\mathcal{S}}\|_2^2 \leq \|\boldsymbol{b}\|_2^2 \leq \frac{1}{\phi^2_{\min}\left[\boldsymbol{\Sigma}_{(\backslash j,\backslash j)}\right]} \boldsymbol{b}^\top \boldsymbol{\Sigma}_{(\backslash j,\backslash j)} \boldsymbol{b},$$

with $\liminf_{n\to\infty} \phi^2_{\min}[\boldsymbol{\Sigma}_{(\backslash j,\backslash j)}] > 0$. Thus, (**C1**) is satisfied with $\phi = \phi_{\min}[\boldsymbol{\Sigma}_{(\backslash j,\backslash j)}]/8$.

(**C2**) For any $\boldsymbol{v} \in \mathbb{R}^{p-1}$ such that $\|\boldsymbol{v}\|_2 = 1$ and $|\text{supp}(\boldsymbol{v})| = 1$, we have $\boldsymbol{v}^\top \boldsymbol{\Sigma}_{(\backslash j,\backslash j)} \boldsymbol{v} \leq \phi^2_{\max}[\boldsymbol{\Sigma}_{(\backslash j,\backslash j)}] \leq \phi^2_{\max}[\boldsymbol{\Sigma}] < \infty$, where the inequality $\phi^2_{\max}[\boldsymbol{\Sigma}_{(\backslash j,\backslash j)}] \leq \phi^2_{\max}[\boldsymbol{\Sigma}]$ is also based on the interlacing property of eigenvalues of principal sub-matrices, and the last equality is guaranteed by (**A1**).

(**C3**) Combining conditions in (**A2**), we get $\sqrt{q_j \log(p)/n}/b^j_{\min} \to 0$, which implies that $q_j \log(p)/n \to 0$.

□

We now proceed to prove Proposition 4.4.1 for population I.

Proof of Proposition 4.4.1. First, if $q_j \equiv |ne_j| = 0$, then we trivially have $\hat{ne}_j \supseteq ne_j$.

If $q_j \equiv |ne_j| \geq 1$, we write $\boldsymbol{X}_j = \boldsymbol{X}_{\setminus j}\boldsymbol{\beta}^j + \boldsymbol{\epsilon}^j$. With Gaussian $\boldsymbol{X}$ as required in (**A1**), $\boldsymbol{\epsilon}^j$ follows a Gaussian distribution.

Theorem 7.2 in Bickel et al. (2009) shows that with Gaussian design, $(ne_j, q_j, 3)$-restricted eigenvalue condition proved in Lemma C.1.1 and $\lambda_j \gtrsim \sqrt{\log(p)/n}$,

$$\lim_{n\to\infty} \Pr\left[\left\|\hat{\boldsymbol{\beta}}^j - \boldsymbol{\beta}^j\right\|_2 \leq \frac{\lambda_j\sqrt{8q_j}}{\phi^2}\right] = 1. \tag{C.2}$$

In addition, given that $\liminf_{n\to\infty} \phi^2 > 0$, which is guaranteed by Lemma C.1.1, for n sufficiently large, (**A2**) implies that $b^j_{\min} > 3\lambda_j\sqrt{q_j}/\phi^2$. Thus, for any k such that $|\beta^j_k| > 0$, in the event that

$$\left\|\hat{\boldsymbol{\beta}}^j - \boldsymbol{\beta}^j\right\|_\infty \leq \left\|\hat{\boldsymbol{\beta}}^j - \boldsymbol{\beta}^j\right\|_2 \leq \frac{\lambda_j\sqrt{8q_j}}{\phi^2}, \tag{C.3}$$

$|\beta^j_k| > 0$ implies $|\hat{\beta}^j_k| > 0$. Therefore, by (C.2),

$$\lim_{n\to\infty} \Pr\left[\hat{ne}_j \supseteq ne_j\right] = 1.$$

□

C.2 Proof of Proposition 4.4.2

Similar to Section C.1, in this section, we prove that (**A1**)–(**A3**) for some variable $j \in \mathcal{V}$ imply

$$\lim_{n\to\infty} \Pr\left[\hat{ne}^{\mathrm{I}}_j = \check{ne}^{\mathrm{I}}_j\right] = 1.$$

The counterpart for population II can be proved using the same technique. Together these imply

$$\lim_{n\to\infty} \Pr\left[\hat{ne}^{0}_j = \check{ne}^{0}_j\right] = 1.$$

For brevity, we drop the superscript "I" in the subsequent proofs. We first state and prove lemmas needed for the proof of Proposition 4.4.2.

Lemma C.2.2. *Suppose (**A1**) and (**A2**) for variable $j \in \mathcal{V}$ hold. Then if $b^j_{\min} > 3\lambda_j\sqrt{q_j}/\phi^2$, we have $\check{ne}_j \supseteq ne_j$, where $\check{ne}_j \equiv \mathrm{supp}(\check{\boldsymbol{\beta}}^j)$ and $ne_j \equiv \mathrm{supp}(\boldsymbol{\beta}^j)$.*

Proof. First, if $|ne_j| = 0$, we trivially have $\check{ne}_j \supseteq ne_j$.

If $|ne_j| \geq 1$, (**A2**) implies that given $\liminf_{n\to\infty} \phi^2 > 0$, for n sufficiently large, $b^j_{\min} > 3\lambda_j\sqrt{q_j}/\phi^2$.

On the other hand, by Corollary 2.1 in van de Geer and Bühlmann (2009), (**C1**) in the proof of Lemma C.1.1 guarantees that

$$\left\|\check{\boldsymbol{\beta}}^j - \boldsymbol{\beta}^j\right\|_\infty \leq \left\|\check{\boldsymbol{\beta}}^j - \boldsymbol{\beta}^j\right\|_2 \leq \frac{\lambda_j\sqrt{8q_j}}{\phi^2}. \tag{C.4}$$

Therefore, similar to the proof of Proposition 4.4.1, if $b^j_{\min} > 3\lambda_j\sqrt{q_j}/\phi^2$, for any k such that $|\beta^j_k| > 0$, we have $|\check{\beta}^j_k| > 0$ by (C.4), which implies that $\check{ne}_j \supseteq ne_j$. □

Lemma C.2.3. *Suppose (**A1**)–(**A3**) hold. Then the estimator $\hat{\boldsymbol{\beta}}^j$ defined in (4.9) satisfies $\|\hat{\boldsymbol{\beta}}^j - \check{\boldsymbol{\beta}}^j\|_1 = \mathcal{O}_p(\sqrt{\log(p)/n})$.*

Proof. Let $Q(\boldsymbol{b}) \equiv \|\boldsymbol{X}_j - \boldsymbol{X}_{\setminus j}\boldsymbol{b}\|_2^2/(2n) + \lambda_j\|\boldsymbol{b}\|_1$, i.e., $\hat{\boldsymbol{\beta}}^j = \arg\min_{\boldsymbol{b}\in\mathbb{R}^{p-1}} Q(\boldsymbol{b})$. To prove Lemma C.2.3, we show that for all $\xi > 0$, there exists a constant $m > 0$, such that

$$\lim_{n\to\infty} \Pr\left[\inf_{\boldsymbol{b}:\|\boldsymbol{b}\|_1=m} Q\left(\check{\boldsymbol{\beta}}^j + \boldsymbol{b}\sqrt{\frac{\log(p)}{n}}\right) > Q\left(\check{\boldsymbol{\beta}}^j\right)\right] = 1. \tag{C.5}$$

Because Q is convex, (C.5) implies that $\hat{\boldsymbol{\beta}}^j$ lies in the convex region $\{\check{\boldsymbol{\beta}}^j + \boldsymbol{b}\sqrt{\log(p)/n} : \|\boldsymbol{b}\|_1 < m\}$ with probability tending to one. Therefore, we have

$$\lim_{n\to\infty} \Pr\left[\left\|\hat{\boldsymbol{\beta}}^j - \check{\boldsymbol{\beta}}^j\right\|_1 \leq m\sqrt{\frac{\log(p)}{n}}\right] = 1,$$

i.e., $\|\hat{\boldsymbol{\beta}}^j - \check{\boldsymbol{\beta}}^j\|_1 = \mathcal{O}_p(\sqrt{\log(p)/n})$.

To prove (C.5), we denote $\boldsymbol{w} = \arg\min_{\boldsymbol{b}:\|\boldsymbol{b}\|_1=m} Q(\check{\boldsymbol{\beta}}^j + \boldsymbol{b}\sqrt{\log(p)/n})$. Expanding terms,

we get

$$
\begin{aligned}
Q\left(\check{\boldsymbol{\beta}}^j + \boldsymbol{w}\sqrt{\frac{\log(p)}{n}}\right) - Q\left(\check{\boldsymbol{\beta}}^j\right) &= \frac{1}{2n}\left\|\left(\boldsymbol{X}_j - \boldsymbol{X}_{\setminus j}\check{\boldsymbol{\beta}}^j\right) - \boldsymbol{X}_{\setminus j}\boldsymbol{w}\sqrt{\frac{\log(p)}{n}}\right\|_2^2 - \frac{1}{2n}\left\|\boldsymbol{X}_j - \boldsymbol{X}_{\setminus j}\check{\boldsymbol{\beta}}^j\right\|_2^2 \\
&\quad + \lambda_j\left\|\check{\boldsymbol{\beta}}^j + \boldsymbol{w}\sqrt{\frac{\log(p)}{n}}\right\|_1 - \lambda_j\left\|\check{\boldsymbol{\beta}}^j\right\|_1 \\
&= -\frac{\sqrt{\log(p)}}{n^{3/2}}\boldsymbol{w}^\top \boldsymbol{X}_{\setminus j}^\top\left(\boldsymbol{X}_j - \boldsymbol{X}_{\setminus j}\check{\boldsymbol{\beta}}^j\right) + \frac{\log(p)}{2n^2}\boldsymbol{w}^\top \boldsymbol{X}_{\setminus j}^\top \boldsymbol{X}_{\setminus j}\boldsymbol{w} \\
&\quad + \lambda_j\left\|\check{\boldsymbol{\beta}}^j + \boldsymbol{w}\sqrt{\frac{\log(p)}{n}}\right\|_1 - \lambda_j\left\|\check{\boldsymbol{\beta}}^j\right\|_1. \qquad \text{(C.6)}
\end{aligned}
$$

Now, for any $g \neq 0, h \in \mathbb{R}$, we have $|g+h| \geq |g| + \text{sign}(g)h$. This is because, 1) if g and h have the same sign, $|g+h| = |g| + |h| = |g| + \text{sign}(h)h = |g| + \text{sign}(g)h$; 2) if they have the opposite signs, $|g+h| = ||g| - |h|| \geq |g| - |h| = |g| - \text{sign}(h)h = |g| + \text{sign}(g)h$; 3) if $h = 0$, $|g+h| = |g| = |g| + \text{sign}(g)h$. Thus,

$$
\begin{aligned}
\left\|\check{\boldsymbol{\beta}}^j + \boldsymbol{w}\sqrt{\frac{\log(p)}{n}}\right\|_1 &= \left\|\check{\boldsymbol{\beta}}^j_{\setminus \check{n}e_j} + \boldsymbol{w}_{\setminus \check{n}e_j}\sqrt{\frac{\log(p)}{n}}\right\|_1 + \left\|\check{\boldsymbol{\beta}}^j_{\check{n}e_j} + \boldsymbol{w}_{\check{n}e_j}\sqrt{\frac{\log(p)}{n}}\right\|_1 \\
&= \sqrt{\frac{\log(p)}{n}}\left\|\boldsymbol{w}_{\setminus \check{n}e_j}\right\|_1 + \left\|\check{\boldsymbol{\beta}}^j_{\check{n}e_j} + \boldsymbol{w}_{\check{n}e_j}\sqrt{\frac{\log(p)}{n}}\right\|_1 \\
&\geq \sqrt{\frac{\log(p)}{n}}\left\|\boldsymbol{w}_{\setminus \check{n}e_j}\right\|_1 + \left\|\check{\boldsymbol{\beta}}^j_{\check{n}e_j}\right\|_1 + \sqrt{\frac{\log(p)}{n}}\check{\boldsymbol{\tau}}^{j\top}_{\check{n}e_j}\boldsymbol{w}_{\check{n}e_j} \\
&= \sqrt{\frac{\log(p)}{n}}\left\|\boldsymbol{w}_{\setminus \check{n}e_j}\right\|_1 + \left\|\check{\boldsymbol{\beta}}^j\right\|_1 + \sqrt{\frac{\log(p)}{n}}\check{\boldsymbol{\tau}}^{j\top}_{\check{n}e_j}\boldsymbol{w}_{\check{n}e_j}. \qquad \text{(C.7)}
\end{aligned}
$$

In the second line and the fourth line, we use the fact that $\check{\boldsymbol{\beta}}^j_{\setminus \check{n}e_j} = \mathbf{0}$, and in the third line, we use the fact that $\check{\boldsymbol{\tau}}^j_{\check{n}e_j} = \text{sign}(\check{\boldsymbol{\beta}}^j_{\check{n}e_j})$ and the inequality $|g+h| \geq |g| + \text{sign}(g)h$ shown

above. Therefore, combining (C.6) and (C.7), we have

$$\begin{aligned} Q\left(\check{\boldsymbol{\beta}}^j + \boldsymbol{w}\sqrt{\frac{\log(p)}{n}}\right) - Q\left(\check{\boldsymbol{\beta}}^j\right) &\geq -\frac{\sqrt{\log(p)}}{n^{3/2}}\boldsymbol{w}^\top \boldsymbol{X}_{\setminus j}^\top\left(\boldsymbol{X}_j - \boldsymbol{X}_{\setminus j}\check{\boldsymbol{\beta}}^j\right) + \frac{\log(p)}{2n^2}\boldsymbol{w}^\top \boldsymbol{X}_{\setminus j}^\top \boldsymbol{X}_{\setminus j}\boldsymbol{w} \\ &\quad + \lambda_j\sqrt{\frac{\log(p)}{n}}\check{\boldsymbol{\tau}}_{\check{n}e_j}^{j\top}\boldsymbol{w}_{\check{n}e_j} + \lambda_j\sqrt{\frac{\log(p)}{n}}\left\|\boldsymbol{w}_{\setminus \check{n}e_j}\right\|_1 \\ &= -\frac{\sqrt{\log(p)}}{n^{3/2}}\boldsymbol{w}^\top \boldsymbol{X}_{\setminus j}^\top\left(\boldsymbol{X}_j - \boldsymbol{X}_{\setminus j}\check{\boldsymbol{\beta}}^j\right) + \frac{\log(p)}{2n}\boldsymbol{w}^\top \boldsymbol{G}_{(\setminus j,\setminus j)}\boldsymbol{w} \\ &\quad + \lambda_j\sqrt{\frac{\log(p)}{n}}\check{\boldsymbol{\tau}}^{j\top}\boldsymbol{w} + \lambda_j\sqrt{\frac{\log(p)}{n}}\left(\left\|\boldsymbol{w}_{\setminus \check{n}e_j}\right\|_1 - \check{\boldsymbol{\tau}}_{\setminus \check{n}e_j}^{j\top}\boldsymbol{w}_{\setminus \check{n}e_j}\right), \end{aligned}$$

where, as before, $\boldsymbol{G}_{(\setminus j,\setminus j)} = \boldsymbol{X}_{\setminus j}^\top \boldsymbol{X}_{\setminus j}/n$. Since by (**A3**), $\limsup_{n\to\infty}\|\check{\boldsymbol{\tau}}_{\setminus \check{n}e_j}^j\|_\infty \leq 1-\delta$, for n sufficiently large, $\|\check{\boldsymbol{\tau}}_{\setminus \check{n}e_j}^j\|_\infty \leq 1-\delta/2$. Thus, $\check{\boldsymbol{\tau}}_{\setminus \check{n}e_j}^{j\top}\boldsymbol{w}_{\setminus \check{n}e_j} \leq |\check{\boldsymbol{\tau}}_{\setminus \check{n}e_j}^{j\top}\boldsymbol{w}_{\setminus \check{n}e_j}| \leq \|\check{\boldsymbol{\tau}}_{\setminus \check{n}e_j}^j\|_\infty \|\boldsymbol{w}_{\setminus \check{n}e_j}\|_1 \leq (1-\delta/2)\|\boldsymbol{w}_{\setminus \check{n}e_j}\|_1$, and

$$\begin{aligned} Q\left(\check{\boldsymbol{\beta}}^j + \boldsymbol{w}\sqrt{\frac{\log(p)}{n}}\right) - Q\left(\check{\boldsymbol{\beta}}^j\right) &\geq -\frac{\sqrt{\log(p)}}{n^{3/2}}\boldsymbol{w}^\top\left[\boldsymbol{X}_{\setminus j}^\top\left(\boldsymbol{X}_j - \boldsymbol{X}_{\setminus j}\check{\boldsymbol{\beta}}^j\right) - \lambda_j n\check{\boldsymbol{\tau}}^j\right] \\ &\quad + \frac{\log(p)}{2n}\boldsymbol{w}^\top \boldsymbol{G}_{(\setminus j,\setminus j)}\boldsymbol{w} + \lambda_j\frac{\delta}{2}\sqrt{\frac{\log(p)}{n}}\left\|\boldsymbol{w}_{\setminus \check{n}e_j}\right\|_1. \end{aligned} \quad \text{(C.8)}$$

To bound $\boldsymbol{w}^\top[\boldsymbol{X}_{\setminus j}^\top(\boldsymbol{X}_j - \boldsymbol{X}_{\setminus j}\check{\boldsymbol{\beta}}^j) - \lambda_j n\check{\boldsymbol{\tau}}^j]$ in (C.8), writing $\boldsymbol{X}_j = \boldsymbol{X}_{\setminus j}\boldsymbol{\beta}^j + \boldsymbol{\epsilon}^j$, we observe

$$\begin{aligned} \frac{1}{n}\boldsymbol{w}^\top\left[\boldsymbol{X}_{\setminus j}^\top\left(\boldsymbol{X}_j - \boldsymbol{X}_{\setminus j}\check{\boldsymbol{\beta}}^j\right) - \lambda_j\check{\boldsymbol{\tau}}^j\right] &= \boldsymbol{w}^\top\left[\boldsymbol{G}_{(\setminus j,\setminus j)}\left(\boldsymbol{\beta}^j - \check{\boldsymbol{\beta}}^j\right) - \lambda_j\check{\boldsymbol{\tau}}^j + \frac{1}{n}\boldsymbol{X}_{\setminus j}^\top\boldsymbol{\epsilon}^j\right] \\ &\leq \left|\boldsymbol{w}^\top\left[\boldsymbol{G}_{(\setminus j,\setminus j)}\left(\boldsymbol{\beta}^j - \check{\boldsymbol{\beta}}^j\right) - \lambda_j\check{\boldsymbol{\tau}}^j + \frac{1}{n}\boldsymbol{X}_{\setminus j}^\top\boldsymbol{\epsilon}^j\right]\right| \\ &= \Big|\boldsymbol{w}^\top\Big[\boldsymbol{G}_{(\setminus j,\setminus j)}\left(\boldsymbol{\beta}^j - \check{\boldsymbol{\beta}}^j\right) - \boldsymbol{\Sigma}_{(\setminus j,\setminus j)}\left(\boldsymbol{\beta}^j - \check{\boldsymbol{\beta}}^j\right) \\ &\quad + \boldsymbol{\Sigma}_{(\setminus j,\setminus j)}\left(\boldsymbol{\beta}^j - \check{\boldsymbol{\beta}}^j\right) - \lambda_j\check{\boldsymbol{\tau}}^j + \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\epsilon}^j\Big]\Big|. \end{aligned}$$

Based on the stationary condition of (4.10), we have $\boldsymbol{\Sigma}_{(\backslash j,\backslash j)}(\boldsymbol{\beta}^j - \check{\boldsymbol{\beta}}^j) - \lambda_j \check{\boldsymbol{\tau}}^j = \mathbf{0}$. Thus,

$$\begin{aligned}
\frac{1}{n}\boldsymbol{w}^\top \left[\boldsymbol{X}_{\backslash j}^\top \left(\boldsymbol{X}_j - \boldsymbol{X}_{\backslash j}\check{\boldsymbol{\beta}}^j\right) - \lambda_j \check{\boldsymbol{\tau}}^j\right] &= \left|\boldsymbol{w}^\top \left[\left(\boldsymbol{G}_{(\backslash j,\backslash j)} - \boldsymbol{\Sigma}_{(\backslash j,\backslash j)}\right)\left(\boldsymbol{\beta}^j - \check{\boldsymbol{\beta}}^j\right) + \frac{1}{n}\boldsymbol{X}_{\backslash j}^\top \boldsymbol{\epsilon}^j\right]\right| \\
&\leq \left|\boldsymbol{w}^\top \left(\boldsymbol{G}_{(\backslash j,\backslash j)} - \boldsymbol{\Sigma}_{(\backslash j,\backslash j)}\right)\left(\boldsymbol{\beta}^j - \check{\boldsymbol{\beta}}^j\right)\right| + \left|\frac{1}{n}\boldsymbol{w}^\top \boldsymbol{X}_{\backslash j}^\top \boldsymbol{\epsilon}^j\right| \\
&\leq \|\boldsymbol{w}\|_1 \left\|\left(\boldsymbol{G}_{(\backslash j,\backslash j)} - \boldsymbol{\Sigma}_{(\backslash j,\backslash j)}\right)\left(\boldsymbol{\beta}^j - \check{\boldsymbol{\beta}}^j\right)\right\|_\infty + \frac{1}{n}\left\|\boldsymbol{w}\right\|_1 \left\|\boldsymbol{X}_{\backslash j}^\top \boldsymbol{\epsilon}^j\right\|_\infty \\
&= m\left(\left\|\left(\boldsymbol{G}_{(\backslash j,\backslash j)} - \boldsymbol{\Sigma}_{(\backslash j,\backslash j)}\right)\left(\boldsymbol{\beta}^j - \check{\boldsymbol{\beta}}^j\right)\right\|_\infty + \frac{1}{n}\left\|\boldsymbol{X}_{\backslash j}^\top \boldsymbol{\epsilon}^j\right\|_\infty\right).
\end{aligned} \tag{C.9}$$

Based on e.g., Lemma 1 in Ravikumar et al. (2011), assuming (**A1**), we have $\|\boldsymbol{G}_{(\backslash j,\backslash j)} - \boldsymbol{\Sigma}_{(\backslash j,\backslash j)}\|_{\max} = \mathcal{O}_p(\sqrt{\log(p)/n})$, where $\|\cdot\|_{\max}$ is the entry-wise infinity norm. In addition, according to Lemma 2.1 in van de Geer and Bühlmann (2009), with (**C1**) in the proof of Lemma C.1.1, we have $\|\boldsymbol{\beta}^j - \check{\boldsymbol{\beta}}^j\|_1 = \mathcal{O}(\lambda_j q_j)$. Thus,

$$\begin{aligned}
\left\|\left(\boldsymbol{G}_{(\backslash j,\backslash j)} - \boldsymbol{\Sigma}_{(\backslash j,\backslash j)}\right)\left(\boldsymbol{\beta}^j - \check{\boldsymbol{\beta}}^j\right)\right\|_\infty &\leq \left\|\boldsymbol{G}_{(\backslash j,\backslash j)} - \boldsymbol{\Sigma}_{(\backslash j,\backslash j)}\right\|_{\max} \left\|\boldsymbol{\beta}^j - \check{\boldsymbol{\beta}}^j\right\|_1 \\
&= \mathcal{O}_p\left(\lambda_j q_j \sqrt{\frac{\log(p)}{n}}\right).
\end{aligned} \tag{C.10}$$

Since $\lambda_j q_j \to l < \infty$ in (**A3**), we have $\|(\boldsymbol{G}_{(\backslash j,\backslash j)} - \boldsymbol{\Sigma}_{(\backslash j,\backslash j)})(\boldsymbol{\beta}^j - \check{\boldsymbol{\beta}}^j)\|_\infty = \mathcal{O}_p(\sqrt{\log(p)/n})$. Based on a well-known result on Gaussian random variables, we also have $\|\boldsymbol{X}_{\backslash j}^\top \boldsymbol{\epsilon}^j\|_\infty / n = \mathcal{O}_p(\sqrt{\log(p)/n})$. Thus, we obtain

$$\frac{1}{mn}\boldsymbol{w}^\top \left[\boldsymbol{X}_{\backslash j}^\top \left(\boldsymbol{X}_j - \boldsymbol{X}_{\backslash j}\check{\boldsymbol{\beta}}^j\right) - \lambda_j n \check{\boldsymbol{\tau}}^j\right] = \mathcal{O}_p\left(\sqrt{\frac{\log(p)}{n}}\right). \tag{C.11}$$

To bound the other term $\log(p)\boldsymbol{w}^\top \boldsymbol{G}_{(\backslash j,\backslash j)}\boldsymbol{w}/2n + \lambda_j \delta \sqrt{\log(p)}\|\boldsymbol{w}_{\backslash \check{n}e_j}\|_1/(2\sqrt{n})$ in (C.8), consider two cases: 1) $\check{n}e_j = \emptyset$ and 2) $\check{n}e_j \neq \emptyset$.

If 1) $\check{n}e_j = \emptyset$,

$$\frac{\log(p)}{2n}\boldsymbol{w}^\top \boldsymbol{G}_{(\backslash j,\backslash j)}\boldsymbol{w} + \frac{\lambda_j \delta}{2}\sqrt{\frac{\log(p)}{n}}\left\|\boldsymbol{w}_{\backslash \check{n}e_j}\right\|_1 \geq \frac{\lambda_j \delta}{2}\sqrt{\frac{\log(p)}{n}}\left\|\boldsymbol{w}\right\|_1 = \frac{\lambda_j \delta}{2}\sqrt{\frac{\log(p)}{n}} m.$$

The first inequality is due to the positive semi-definitiveness of $\boldsymbol{G}_{(\backslash j,\backslash j)}$. Hence,

$$Q\left(\check{\boldsymbol{\beta}}^j + \boldsymbol{w}\sqrt{\frac{\log(p)}{n}}\right) - Q\left(\check{\boldsymbol{\beta}}^j\right) \geq m\sqrt{\frac{\log(p)}{n}}\left(\lambda_j\frac{\delta}{2} - \frac{1}{mn}\boldsymbol{w}^\top\left[\boldsymbol{X}_{\backslash j}^\top\left(\boldsymbol{X}_j - \boldsymbol{X}_{\backslash j}\check{\boldsymbol{\beta}}^j\right) - \lambda_j n\check{\boldsymbol{\tau}}^j\right]\right). \tag{C.12}$$

By (C.11), and given that $\sqrt{\log(p)/n}/(\lambda_j\delta) \to 0$ by (**A3**), for any $m > 0$

$$Q\left(\check{\boldsymbol{\beta}}^j + \boldsymbol{w}\sqrt{\frac{\log(p)}{n}}\right) > Q\left(\check{\boldsymbol{\beta}}^j\right)$$

with high probability, which implies that (C.5) holds.

If 2) $\check{ne}_j \neq \emptyset$, i.e., $q_j \geq 1$, we further consider two cases: i) $\|\boldsymbol{w}_{\backslash\check{ne}_j}\|_1 > 3\|\boldsymbol{w}_{\check{ne}_j}\|_1$, and ii) $\|\boldsymbol{w}_{\backslash k}\|_1 \leq 3|\boldsymbol{w}_k|$, where k is any index such that $k \in ne_j$. Note that these two cases are not mutually exclusive. However, we proved in Lemma C.2.2 that if $b^j_{\min} > 3\lambda_j\sqrt{q_j}/\phi^2$, which happens when n is sufficiently large, then $\check{ne}_j \supseteq ne_j$. Thus, for any $\boldsymbol{w}$ and sufficiently large n, because $k \in ne_j$, $\|\boldsymbol{w}_{\backslash\check{ne}_j}\|_1 > 3\|\boldsymbol{w}_{\check{ne}_j}\|_1$ implies

$$\left\|\boldsymbol{w}_{\backslash k}\right\|_1 \geq \left\|\boldsymbol{w}_{\backslash ne_j}\right\|_1 \geq \left\|\boldsymbol{w}_{\backslash\check{ne}_j}\right\|_1 > 3\left\|\boldsymbol{w}_{\check{ne}_j}\right\|_1 \geq 3\left\|\boldsymbol{w}_{ne_j}\right\|_1 \geq 3\left|\boldsymbol{w}_k\right|.$$

Therefore, although the two cases i) $\|\boldsymbol{w}_{\backslash\check{ne}_j}\|_1 > 3\|\boldsymbol{w}_{\check{ne}_j}\|_1$ and ii) $\|\boldsymbol{w}_{\backslash k}\|_1 \leq 3|\boldsymbol{w}_k|$ are not mutually exclusive, they cover all possibilities.

If i) $\|\boldsymbol{w}_{\backslash\check{ne}_j}\|_1 > 3\|\boldsymbol{w}_{\check{ne}_j}\|_1$, because $\|\boldsymbol{w}\|_1 = \|\boldsymbol{w}_{\backslash\check{ne}_j}\|_1 + \|\boldsymbol{w}_{\check{ne}_j}\|_1 = m$, $\|\boldsymbol{w}_{\backslash\check{ne}_j}\|_1 > 3m/4$, and

$$\begin{aligned}\frac{\log(p)}{2n}\boldsymbol{w}^\top\boldsymbol{G}_{(\backslash j,\backslash j)}\boldsymbol{w} + \lambda_j\frac{\delta}{2}\sqrt{\frac{\log(p)}{n}}\left\|\boldsymbol{w}_{\backslash\check{ne}_j}\right\|_1 &\geq \lambda_j\frac{\delta}{2}\sqrt{\frac{\log(p)}{n}}\left\|\boldsymbol{w}_{\backslash\check{ne}_j}\right\|_1 \\ &> \lambda_j\frac{\delta}{2}\sqrt{\frac{\log(p)}{n}}\frac{3m}{4}.\end{aligned} \tag{C.13}$$

Combining (C.8), (C.9) and (C.13), we get

$$Q\left(\check{\boldsymbol{\beta}}^j + \boldsymbol{w}\sqrt{\frac{\log(p)}{n}}\right) - Q\left(\check{\boldsymbol{\beta}}^j\right) \geq m\sqrt{\frac{\log(p)}{n}}\left(\lambda_j\frac{3\delta}{8} - \frac{1}{mn}\boldsymbol{w}^\top\left[\boldsymbol{X}_{\backslash j}^\top\left(\boldsymbol{X}_j - \boldsymbol{X}_{\backslash j}\check{\boldsymbol{\beta}}^j\right) - \lambda_j n\check{\boldsymbol{\tau}}^j\right]\right). \tag{C.14}$$

Because $\sqrt{\log(p)/n}/(\lambda_j\delta) \to 0$ by (**A3**) and $\boldsymbol{w}^\top[\boldsymbol{X}_{\setminus j}^\top(\boldsymbol{X}_j - \boldsymbol{X}_{\setminus j}\check{\boldsymbol{\beta}}^j) - \lambda_j n\check{\boldsymbol{\tau}}^j]/(mn) = \mathcal{O}_p\left(\sqrt{\log(p)/n}\right)$ by (C.11), with any $m > 0$ and n sufficiently large, we have

$$Q\left(\check{\boldsymbol{\beta}}^j + \boldsymbol{w}\sqrt{\frac{\log(p)}{n}}\right) > Q\left(\check{\boldsymbol{\beta}}^j\right),$$

and hence (C.5) holds.

On the other hand, if ii) $\|\boldsymbol{w}_{\setminus k}\|_1 \le 3|\boldsymbol{w}_k|$, because $\|\boldsymbol{w}\|_1 = \|\boldsymbol{w}_{\setminus k}\|_1 + |\boldsymbol{w}_k| = m$, we have $|\boldsymbol{w}_k| \ge m/4$. Let $\mathcal{S} = ne_j \cup \{i\}$ where $i = \arg\max_{j:j\notin ne_j} |w_j|$. Then $k \in \mathcal{S}$, $|\mathcal{S}| = q_j + 1 \le 2q_j$ and $\|\boldsymbol{w}_{\setminus\mathcal{S}}\|_\infty \le \min_{j\in\mathcal{S}\setminus ne_j} |w_j|$. Hence, the $(ne_j, q_j, 3)$-restricted eigenvalue condition in Lemma C.1.1 implies that, with probability tending to one, as n approaches infinity,

$$\begin{aligned}
\frac{\log(p)}{2n}\boldsymbol{w}^\top\boldsymbol{G}_{(\setminus j,\setminus j)}\boldsymbol{w} + \lambda_j\frac{\delta}{2}\sqrt{\frac{\log(p)}{n}}\left\|\boldsymbol{w}_{\setminus \check{n}e_j}\right\|_1 &\ge \frac{\log(p)}{2n}\boldsymbol{w}^\top\boldsymbol{G}_{(\setminus j,\setminus j)}\boldsymbol{w} \\
&\ge \frac{\log(p)\phi^2}{2n}\|\boldsymbol{w}_{\mathcal{S}}\|_2^2 \\
&\ge \frac{\log(p)\phi^2}{2n}\boldsymbol{w}_k^2 \\
&\ge \frac{\log(p)\phi^2}{2n}\frac{m^2}{16}. \qquad \text{(C.15)}
\end{aligned}$$

Thus, combining (C.8), (C.9) and (C.15), we find that for n sufficiently large, $\phi^2 \ge \kappa^2/2$, and

$$Q\left(\check{\boldsymbol{\beta}}^j + \boldsymbol{w}\sqrt{\frac{\log(p)}{n}}\right) - Q\left(\check{\boldsymbol{\beta}}^j\right) \ge m\sqrt{\frac{\log(p)}{n}}\left(\frac{\kappa^2}{128}\sqrt{\frac{\log(p)}{n}}m - \frac{1}{mn}\boldsymbol{w}^\top\left[\boldsymbol{X}_{\setminus j}^\top\left(\boldsymbol{X}_j - \boldsymbol{X}_{\setminus j}\check{\boldsymbol{\beta}}^j\right) - \lambda_j n\check{\boldsymbol{\tau}}^j\right]\right).$$

Since $\boldsymbol{w}^\top[\boldsymbol{X}_{\setminus j}^\top(\boldsymbol{X}_j - \boldsymbol{X}_{\setminus j}\check{\boldsymbol{\beta}}^j) - \lambda_j n\check{\boldsymbol{\tau}}^j]/(mn) = \mathcal{O}_p(\sqrt{\log(p)/n})$, we can choose m to be sufficiently large, not depending on n, such that (C.5) holds. □

Lemma C.2.4. *Suppose (**A2**) and (**A3**) hold. For $\check{n}e_j \neq \emptyset$, we have*

$$\sqrt{\frac{\log(p)}{n}}\frac{1}{\check{b}_{\min}^j} \to 0, \qquad \text{(C.16)}$$

where $\check{b}_{\min}^j \equiv \min\{|\check{\beta}_k^j| : \check{\beta}_k^j \neq 0\}$.

Proof. We show that $\sqrt{\log(p)/n}/|\check{\beta}_k^j| \to 0$ for any $k \in \check{ne}_j$ by considering two cases: 1) for $k \in ne_j$, and 2) for $k \in \check{ne}_j \backslash ne_j$.

1) For any $k \in ne_j$, and for n sufficiently large, Lemma C.2.2 indicates that

$$\left\|\check{\boldsymbol{\beta}}^j - \boldsymbol{\beta}^j\right\|_\infty \leq \left\|\check{\boldsymbol{\beta}}^j - \boldsymbol{\beta}^j\right\|_2 \leq \frac{\lambda_j\sqrt{8q_j}}{\phi^2}.$$

Now, by (**A2**), for any $k \in ne_j$, $|\beta_k^j| > 3\lambda_j\sqrt{q_j}/\phi^2$, with a sufficiently large n. Hence, for any $k \in ne_j$, $|\check{\beta}_k^j| > (3-2\sqrt{2})\lambda_j\sqrt{q_j}/\phi^2$. Therefore, given the rates in (**A2**),

$$0 < \sqrt{\frac{\log(p)}{n}}\frac{1}{\left|\check{\beta}_k^j\right|} < \sqrt{\frac{\log(p)}{n}}\frac{1}{\lambda_j}\cdot\frac{\phi^2}{\left(3-2\sqrt{2}\right)\sqrt{q_j}} \to 0.$$

If $\check{ne}_j = ne_j$, then our proof is complete. Otherwise, 2) for $k \in \check{ne}_j \backslash ne_j$, consider the stationary condition of (4.10),

$$n\lambda_j\check{\boldsymbol{\tau}}_{ne_j}^j = \mathrm{E}\left[\boldsymbol{X}_{\check{ne}_j}^\top\boldsymbol{X}\right]\left(\boldsymbol{\beta}^j - \check{\boldsymbol{\beta}}^j\right) = \mathrm{E}\left[\boldsymbol{X}_{\check{ne}_j}^\top\boldsymbol{X}_{\check{ne}_j}\right]\left(\boldsymbol{\beta}_{\check{ne}_j}^j - \check{\boldsymbol{\beta}}_{\check{ne}_j}^j\right).$$

The second equality holds because based on Lemma C.2.2, $\check{ne}_j \supseteq ne_j$, i.e., $\boldsymbol{\beta}_{\backslash\check{ne}_j}^j = \check{\boldsymbol{\beta}}_{\backslash\check{ne}_j}^j = \mathbf{0}$. Rearranging terms,

$$\check{\boldsymbol{\beta}}_{\check{ne}_j}^j = \boldsymbol{\beta}_{\check{ne}_j}^j - n\lambda_j\mathrm{E}\left[\boldsymbol{X}_{\check{ne}_j}^\top\boldsymbol{X}_{\check{ne}_j}\right]^{-1}\check{\boldsymbol{\tau}}_{\check{ne}_j}^j. \tag{C.17}$$

Recall that for any $k \in \check{ne}_j \backslash ne_j$, $\beta_k^j = 0$. Thus, for any $k \in \check{ne}_j \backslash ne_j$,

$$\left|\check{\beta}_k^j\right| = \left|n\lambda_j\mathrm{E}\left[\boldsymbol{X}_{\check{ne}_j}^\top\boldsymbol{X}_{\check{ne}_j}\right]^{-1}\check{\boldsymbol{\tau}}_{\check{ne}_j}^j\right|_k = \lambda_j\left|\left[\boldsymbol{\Sigma}_{(\check{ne}_j,\check{ne}_j)}\right]^{-1}\check{\boldsymbol{\tau}}_{\check{ne}_j}^j\right|_k. \tag{C.18}$$

By (**A3**), we have

$$\sqrt{\frac{\log(p)}{n}}\frac{1}{\lambda_j}\left(\min_{k\in\check{ne}_j\backslash ne_j}\left|\left[\boldsymbol{\Sigma}_{(\check{ne}_j,\check{ne}_j)}\right]^{-1}\check{\boldsymbol{\tau}}_{\check{ne}_j}^j\right|_k\right)^{-1} \to 0,$$

which means for any $k \in \check{ne}_j \backslash ne_j$, we have

$$\sqrt{\frac{\log(p)}{n}}\frac{1}{\left|\check{\beta}_k^j\right|} \to 0.$$

□

Now we proceed to prove Proposition 4.4.2.

Proof of Proposition 4.4.2. We first note that according to the stationary conditions of (4.10) and (4.9), respectively, we have

$$\check{\boldsymbol{\tau}}^j = \frac{1}{\lambda_j}\boldsymbol{\Sigma}_{(\backslash j,\backslash j)}\left(\boldsymbol{\beta}^j - \check{\boldsymbol{\beta}}^j\right), \tag{C.19}$$

$$\hat{\boldsymbol{\tau}}^j = \frac{1}{n\lambda_j}\boldsymbol{X}_{\backslash j}^\top\left(\boldsymbol{X}_j - \boldsymbol{X}_{\backslash j}\hat{\boldsymbol{\beta}}^j\right). \tag{C.20}$$

Writing $\boldsymbol{X}_j = \boldsymbol{X}_{\backslash j}\boldsymbol{\beta}^j + \boldsymbol{\epsilon}^j$, (C.20) gives

$$\hat{\boldsymbol{\tau}}^j = \frac{1}{\lambda_j}\boldsymbol{G}_{(\backslash j,\backslash j)}\left(\boldsymbol{\beta}^j - \hat{\boldsymbol{\beta}}^j\right) + \frac{1}{n\lambda_j}\boldsymbol{X}_{\backslash j}^\top\boldsymbol{\epsilon}^j \tag{C.21}$$

where $\boldsymbol{G}_{(\backslash j,\backslash j)} = \boldsymbol{X}_{\backslash j}^\top\boldsymbol{X}_{\backslash j}/n$. Combining (C.19) and (C.21),

$$\begin{aligned}
\hat{\boldsymbol{\tau}}^j - \check{\boldsymbol{\tau}}^j &= \frac{1}{\lambda_j}\left(\boldsymbol{G}_{(\backslash j,\backslash j)}\left[\boldsymbol{\beta}^j - \hat{\boldsymbol{\beta}}^j\right] - \boldsymbol{\Sigma}_{(\backslash j,\backslash j)}\left[\boldsymbol{\beta}^j - \check{\boldsymbol{\beta}}^j\right]\right) + \frac{1}{n\lambda_j}\boldsymbol{X}_{\backslash j}^\top\boldsymbol{\epsilon}^j \\
&= \frac{1}{\lambda_j}\left(\boldsymbol{G}_{(\backslash j,\backslash j)}\left[\boldsymbol{\beta}^j - \hat{\boldsymbol{\beta}}^j\right] - \boldsymbol{\Sigma}_{(\backslash j,\backslash j)}\left[\boldsymbol{\beta}^j - \hat{\boldsymbol{\beta}}^j\right]\right. \\
&\quad \left. + \boldsymbol{\Sigma}_{(\backslash j,\backslash j)}\left[\boldsymbol{\beta}^j - \hat{\boldsymbol{\beta}}^j\right] - \boldsymbol{\Sigma}_{(\backslash j,\backslash j)}\left[\boldsymbol{\beta}^j - \check{\boldsymbol{\beta}}^j\right]\right) + \frac{1}{n\lambda_j}\boldsymbol{X}_{\backslash j}^\top\boldsymbol{\epsilon}^j \\
&= \frac{1}{\lambda_j}\left(\boldsymbol{G}_{(\backslash j,\backslash j)} - \boldsymbol{\Sigma}_{(\backslash j,\backslash j)}\right)\left(\boldsymbol{\beta}^j - \hat{\boldsymbol{\beta}}^j\right) + \frac{1}{\lambda_j}\boldsymbol{\Sigma}_{(\backslash j,\backslash j)}\left(\check{\boldsymbol{\beta}}^j - \hat{\boldsymbol{\beta}}^j\right) + \frac{1}{n\lambda_j}\boldsymbol{X}_{\backslash j}^\top\boldsymbol{\epsilon}^j.
\end{aligned} \tag{C.22}$$

We now bound all three terms on the right hand side of (C.22). First, by the Gaussianity of the data,

$$\frac{1}{n\lambda_j}\left\|\boldsymbol{X}_{\backslash j}^\top\boldsymbol{\epsilon}^j\right\|_\infty = \mathcal{O}_p\left(\frac{1}{\lambda_j}\sqrt{\frac{\log(p)}{n}}\right).$$

In addition, we proved in Lemma C.2.3 that $\|\hat{\boldsymbol{\beta}}^j - \check{\boldsymbol{\beta}}^j\|_1 = \mathcal{O}_p\left(\sqrt{\log(p)/n}\right)$. Thus, because $\|\boldsymbol{\Sigma}_{(\backslash j,\backslash j)}\|_{\max} = 1$,

$$\frac{1}{\lambda_j}\left\|\boldsymbol{\Sigma}_{(\backslash j,\backslash j)}\left(\hat{\boldsymbol{\beta}}^j - \check{\boldsymbol{\beta}}^j\right)\right\|_\infty \le \frac{1}{\lambda_j}\left\|\boldsymbol{\Sigma}_{(\backslash j,\backslash j)}\right\|_{\max}\left\|\hat{\boldsymbol{\beta}}^j - \check{\boldsymbol{\beta}}^j\right\|_1 = \mathcal{O}_p\left(\frac{1}{\lambda_j}\sqrt{\frac{\log(p)}{n}}\right).$$

Finally, based on Theorem 7.2 in Bickel et al. (2009), Lemma C.1.1 imply that $\|\boldsymbol{\beta}^j - \hat{\boldsymbol{\beta}}^j\|_1 =$

$\mathcal{O}_p(\lambda_j q_j)$. Thus,

$$\begin{aligned}&\left\|\left(\boldsymbol{G}_{(\backslash j,\backslash j)}-\boldsymbol{\Sigma}_{(\backslash j,\backslash j)}\right)\left(\boldsymbol{\beta}^j-\hat{\boldsymbol{\beta}}^j\right)\right\|_\infty \le \left\|\boldsymbol{G}_{(\backslash j,\backslash j)}-\boldsymbol{\Sigma}_{(\backslash j,\backslash j)}\right\|_{\max}\left\|\boldsymbol{\beta}^j-\hat{\boldsymbol{\beta}}^j\right\|_1\\ &=\mathcal{O}_p\left(\sqrt{\frac{\log(p)}{n}}\lambda_j q_j\right)=\mathcal{O}_p\left(\sqrt{\frac{\log(p)}{n}}\right)=o_p\left(\frac{1}{\lambda_j}\sqrt{\frac{\log(p)}{n}}\right),\end{aligned}$$

where the last inequality holds because $\lambda_j q_j \to l < \infty$ by (**A2**) and hence $\lambda_j \to 0$. Thus, (C.22) shows that $\|\check{\boldsymbol{\tau}}^j - \hat{\boldsymbol{\tau}}^j\|_\infty = \mathcal{O}_p(\sqrt{\log(p)/n}/\lambda_j)$. By (**A3**), we have that $\limsup_{n\to\infty}\|\check{\boldsymbol{\tau}}^j_{\backslash \check{n}e_j}\|_\infty \le 1-\delta$ for $\sqrt{\log(p)/n}/(\lambda_j\delta)\to 0$, and hence $\lim_{n\to\infty}\Pr[\|\hat{\boldsymbol{\tau}}^j_{\backslash \check{n}e_j}\|_\infty < 1] = 1$. Thus,

$$\lim_{n\to\infty}\Pr\left[\check{n}e_j \supseteq \hat{n}e_j\right] = 1. \tag{C.23}$$

On the other hand, if $\check{n}e_j = \emptyset$, $\check{n}e_j \subseteq \hat{n}e_j$. If $\check{n}e_j \neq \emptyset$, by Lemma C.2.4,

$$\sqrt{\frac{\log(p)}{n}}\frac{1}{\check{b}^j_{\min}} \to 0. \tag{C.24}$$

Lemma C.2.3 shows that $\|\hat{\boldsymbol{\beta}}^j - \check{\boldsymbol{\beta}}^j\|_1 = \mathcal{O}_p\left(\sqrt{\log(p)/n}\right)$, i.e., there exists a constant $C>0$ such that

$$\lim_{n\to\infty}\Pr\left[\left\|\hat{\boldsymbol{\beta}}^j-\check{\boldsymbol{\beta}}^j\right\|_\infty > C\sqrt{\frac{\log(p)}{n}}\right] = 0. \tag{C.25}$$

Based on (C.24), for n sufficiently large, $\check{b}^j_{\min} > 2C\sqrt{\log(p)/n}$. Thus, combining (C.24) and (C.25), for n sufficiently large, whenever $\check{\beta}^j_k \neq 0$, we have $|\check{\beta}^j_k| > 2C\sqrt{\log(p)/n}$ and hence $\lim_{n\to\infty}\Pr\left[|\hat{\beta}^j_k| > 0\right] = 1$. Therefore

$$\lim_{n\to\infty}\Pr\left[\check{n}e_j \subseteq \hat{n}e_j\right] = 1, \tag{C.26}$$

which completes the proof. □

VITA

Sen Zhao was born and raised in Beijing, China. He obtained a Bachelor of Arts in Economics and Mathematics from Carleton College in Northfield, MN, and a Doctor of Philosophy in Biostatistics from the University of Washington in Seattle, WA under the advisory of Dr. Ali Shojaie. Sen joined Google Research in April 2017 as a statistician. His personal website is `www.sen-zhao.com`, and he welcomes your comments to `sen-zhao@sen-zhao.com`.