Methods for estimation and inference for high-dimensional models

Lina Lin

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

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Mathias Drton, Chair

Ali Shojaie, Chair

Thomas S. Richardson

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Methods for estimation and inference for high-dimensional models

Lina Lin

Co-Chairs of the Supervisory Committee:
Professor Mathias Drton
Statistics

Associate Professor Ali Shojaie
Biostatistics

This thesis tackles three different problems in high-dimensional statistics. The first two parts of the thesis focus on estimation of sparse high-dimensional undirected graphical models under non-standard conditions, specifically, non-Gaussianity and missingness, when observations are continuous. To address estimation under non-Gaussianity, we propose a general framework involving augmenting the score matching losses introduced in Hyvärinen [2005, 2007] with an ℓ_1-regularizing penalty. This method, which we refer to as *regularized score matching*, allows for computationally efficient treatment of Gaussian and non-Gaussian continuous exponential family models because the considered loss becomes a penalized quadratic and thus yields piecewise linear solution paths. Under suitable irrepresentability conditions and distributional assumptions, we show that regularized score matching generates consistent graph estimates in sparse high-dimensional settings. Through numerical experiments and an application to RNAseq data, we confirm that regularized score matching achieves state-of-the-art performance in the Gaussian case and provides a valuable tool for computationally efficient estimation in non-Gaussian graphical models.

To address estimation of sparse high-dimensional undirected graphical models with missing observations, we propose adapting the regularized score matching framework by substi-

tuting in surrogates of relevant statistics to accommodate these circumstances, as in Loh and Wainwright [2012] and Kolar and Xing [2012]. For Gaussian and non-Gaussian continuous exponential family models, the use of these surrogates may result in a loss of semi-definiteness, and thus nonconvexity, in the objective. Nevertheless, under suitable distributional assumptions, the global optimum is close to the truth in matrix ℓ_1 norm with high probability in sparse high-dimensional settings. Furthermore, under the same set of assumptions, we show that the composite gradient descent algorithm we propose for minimizing the modified objective converges at a geometric rate to a solution close to the global optimum with high probability.

The last part of the thesis moves away from undirected graphical models, and is instead concerned with inference in high-dimensional regression models. Specifically, we investigate how to construct asymptotically valid confidence intervals and p-values for the fixed effects in a high-dimensional linear mixed effect model. The framework we propose, largely founded on a recent work [Bühlmann, 2013], entails de-biasing a 'naive' ridge estimator. We show via numerical experiments that the method controls for Type I error in hypothesis testing and generates confidence intervals that achieve target coverage, outperforming competitors that assume observations are homogeneous when observations are, in fact, correlated within group.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

The author would like to deeply thank the people who have provided me with guidance over the course of her studies: these include but are not limited to Mathias Drton and Ali Shojaie, her co-advisors, Thomas Richardson, Jon Wellner, and Marina Meila.

The author would also like to thank her parents and friends, from both Seattle and Toronto, for their support.

DEDICATION

to my parents

Chapter 1

INTRODUCTION

1.1 An overview

In many applications, the goal is to identify the set of covariates that influence the response. Define n to be the number of observations and p to be the number of covariates in consideration. Provided a response vector $y \in \mathbb{R}^n$ and a random or fixed covariate matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, the simplest relation that can be assumed is given by the linear model

$$y = \mathbf{X}\beta^* + \epsilon, \tag{1.1}$$

with $\beta^* \in \mathbb{R}^p$ the vector of (fixed) regression coefficients and ϵ_i, $i = 1, \ldots, n$, i.i.d. with $\mathbb{E}[\epsilon_1] = 0$ and $\mathrm{Var}[\epsilon_1] = \sigma^{*2} < \infty$. In the low-dimensional setting ($p < n$), an estimator for β^* can be derived by minimizing the squared-error loss (also known as *least squares*), i.e.

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \|y - \mathbf{X}\beta^*\|_2^2. \tag{1.2}$$

It is straightforward to show that $\hat{\beta}$ satisfies many desirable properties: specifically, it is consistent for β^* and normally distributed.

However, with modern datasets, p may be on the order of n or vastly exceed it, in which case, $\hat{\beta}$ is ill-defined because the matrix $\mathbf{X}^T\mathbf{X}$ is no longer invertible. In fact, most classical methods fail, and consistent estimation is usually only achievable under additional structural constraints. In particular, β^* is often assumed to be *sparse*; that is, the number of non-zero elements in β^*, d, is assumed to be small compared to p. The common strategy then is to augment the squared error loss being minimized in (1.2) with a sparsity-inducing penalty. As an example, the lasso estimator [Tibshirani, 1996] arises from modifying the loss with an ℓ_1 penalty:

$$\hat{\beta}^{(lasso)} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \|y - \mathbf{X}\beta^*\|_2^2 + \lambda \|\beta\|_1. \tag{1.3}$$

The solution $\hat{\beta}^{(lasso)}$ will have some entries exactly zero, with larger values for the tuning parameter λ generating sparser estimates. Several theoretical results have been proven for lasso. In particular, supposing $\mathbf{X}$ satisfies what is known as an *incoherence* condition, λ scaling at least $\sqrt{\log p/n}$, and n being at least on the order of $d \log p$, $\hat{\beta}^{(Lasso)}$ is sparsistent (i.e. $\hat{\beta}^{(Lasso)}$ correctly identifies the true set of non-zero elements in β^*) with probability converging to 1 as $p \to \infty$.

In place of lasso, one may also consider nonconvex penalties: see, for example, Fan and Li [2001] (smoothly clipped absolute deviation or *SCAD*) and Zhang [2010] (minimum concave penalty or *MCP*). The adaptive lasso, first proposed in Zou [2006], can be interpreted as an approximation to a nonconvex penalization approach [van de Geer et al., 2011, Huang et al., 2008]. A related procedure, the relaxed Lasso, was proposed in Meinshausen [2007]. The Dantzig selector [Candes and Tao, 2007] forms an alternative approach in which an ℓ_1 norm is minimized under a constraint on $\|X^T(y - X\beta)\|_\infty$. It was shown to have similar statistical properties to the lasso estimator in Bickel et al. [2009]. Other algorithms for model selection include orthogonal matching pursuit (which is more or less forward selection) [Tropp, 2004] or L2Boosting (or matching pursuit) [Bühlmann, 2006].

In other applications, the goal is to characterize the dependencies between a possibly large set of variables. For these purposes, *undirected graphical models*, also known as *Markov random fields*, provide a convenient framework. Each such model is associated to an undirected graph $G = (V, E)$, with vertex set V and edge set $E \subset V \times V$. For a random vector $X = (X_j : j \in V)$ indexed by the nodes of G, the graphical model given by G requires that X_j and X_k be conditionally independent given all other variables whenever nodes j and k are not joined by an edge in G [Lauritzen, 1996] (Note: equivalently, we can say that the distribution of X satisfies the *pairwise Markov property* of G). If G is the smallest graph such that X satisfies this requirement, we term G the *conditional independence graph* of X. In this case, X_j and X_k are conditionally independent given all other variables if and only if

j and k are non-adjacent in G. We will always take the vertex set to be $V = \{1, \ldots, p\}$, so p is the number of observed variables in X.

Specific models are obtained from additional distributional assumptions. Particularly, an assumption of multivariate normality gives Gaussian graphical models, for which estimation of conditional independence graphs is equivalent to *covariance selection* [Dempster, 1972]. If X is jointly multivariate normal with mean vector μ and covariance matrix $\mathbf{\Sigma}$—in symbols, $X \sim N(\mu, \mathbf{\Sigma})$—then the conditional independences among the random variables, and hence edges between nodes in the graph, are determined by the entries of the inverse covariance, or concentration matrix $\mathbf{K} = (\kappa_{jk}) = \mathbf{\Sigma}^{-1}$. More precisely, $\kappa_{jk} = 0$ for $j \neq k$ if and only if X_j and X_k are independent given all other variables.

As in regression, when $p > n$, estimation of undirected graphical models is only tractable in the presence of structural constraints such as sparsity: here, sparsity implies that the maximum number of edges incident to any node, d (reusing notation), is small. Also as was the case with regression, the general strategy involves augmenting a well-known loss function with a sparsity-inducing penalty, such as an ℓ_1 penalty.

Largely due to convenience, Gaussian models have been the primary tool for graphical modeling of data comprising continuous variables, such as gene expression data, and a large number of methods have been proposed for statistical estimation in high-dimensional Gaussian graphical models. Two widely-used approaches are the *graphical lasso* or *glasso* [Yuan and Lin, 2007] and *neighborhood selection* [Meinshausen and Bühlmann, 2006]. In glasso, an ℓ_1 penalty on the entries of the inverse covariance matrix is added to the negative Gaussian log-likelihood. Neighborhood selection, on the other hand, relates graphical model selection back to lasso regression (1.3): the approach leverages the fact that the node-wise full conditional distributions from a Gaussian graphical model form p linear regression models

The thesis addresses three selected problems concerning estimation and inference in high-dimensional linear regression and undirected graphical models under non-standard condi-

tions/assumptions.

In Chapter 2, we focus on the estimation of *non-Gaussian* graphical models. We do note that some variations of this problem have been previously addressed. See Miyamura and Kano [2006], Finegold and Drton [2011], Vogel and Fried [2011] and Sun and Li [2012], who consider outliers. Liu et al. [2009], Liu et al. [2012b] and Dobra and Lenkoski [2011], who focus on Gaussian copula models. Neighborhood selection/pseudo-likelihood procedures have been proposed for categorical models where the node-wise regression is logistic or multinomial [Lee et al., 2007, Höfling and Tibshirani, 2009, Ravikumar et al., 2010, Jalali et al., 2011]. Allen and Liu [2013] and Yang et al. [2012] discuss extensions using node-wise generalized linear models, and semi-/nonparametric methods were proposed by Fellinghauer et al. [2013] and Voorman et al. [2014]. The flexible framework we propose, called *regularized score matching*, is intended to serve as an alternative to these existing approaches. Under certain circumstances, there are clearcut advantages to using *regularized score matching*: particularly, when working with continuous exponential families, we show that the target objective becomes structurally analogous to that optimized in lasso regression (1.2), and is thus computationally very straightforward to optimize.

In Chapter 3, we extend the methodology from Chapter 2 to accommodate missing observations. Missingness is a commonly occurring trait in modern datasets, due to faulty machinery, inability to collect data in longitudinal studies, human error and limits of experimental design. Unlike with conventional methods such as Expectation-Maximization (EM), we can derive theoretical guarantees on statistical and optimization errors for this modified regularized score matching approach under assumptions on the missingness process.

In Chapter 4, we no longer consider graphical models but return to the more basic problem of high-dimensional regression, except here, the errors, ϵ_i, $i = 1, \ldots, n$, are now correlated. Specifically, Chapter 4 concerns the development of an inferential framework for the linear mixed effect model, which is characterized by

$$y_m = \mathbf{X}_m \beta^* + \mathbf{Z}_m \upsilon_m + \epsilon_m, \quad m = 1, \ldots, M \tag{1.4}$$

with

1. $\beta^* \in \mathbb{R}^p$ an unknown vector of fixed regression coefficients,

2. $\upsilon_m \in \mathbb{R}^q$, $m = 1, \ldots, M$ unknown vectors of group-specific random effects, with $\upsilon_m \underset{i.i.d.}{\sim} \mathcal{N}(0, \mathbf{\Psi}^*)$, $\mathbf{\Psi}^* \in \mathcal{S}_+^q$,

3. errors $\epsilon_m \underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^{*2}\mathbf{I}_{n\times n})$, which are generated independently of $\upsilon_1, \ldots, \upsilon_M$, and

4. $\mathbf{X}_m$ and $\mathbf{Z}_m$ known design matrices of dimensions $n \times p$ and $n \times q$, respectively,

for groups $m = 1, \ldots, M$. This is a suitable model for when, as an example, we have repeated n measurements for M subjects.

While the bulk of literature on high-dimensional linear regression has emphasized estimation, made evident above, there has been a recent flood of literature on assigning statistical significance to these high-dimensional estimates. Strategies are varied, ranging from stability selection [Meinshausen and Bühlmann, 2010, Shah and Samworth, 2013] to sample splitting [Wasserman and Roeder, 2009, Meinshausen et al., 2009] to debiasing high-dimensional estimators [Zhang and Zhang, 2014a, van de Geer et al., 2014b, Javanmard and Montanari, 2014, Bühlmann, 2013], (see Chapter 4 for a more in-depth review). However, the cited works all operate on the standard assumption that the observations are homogeneous and the errors ϵ_i, $i = 1, \ldots, n$ i.i.d. Chapter 4 aims to address this gap.

1.2 Notation

The following notation will be used consistently throughout this thesis.

Matrix/vector notations

1. Symbols representing matrix values are bold-faced. In addition, let $\mathbf{A}$ denote some generic matrix; then a_j and a_{jk} denote the jth column and (j,k)th entry of $\mathbf{A}$, respectively.

2. Let $r \in [1, \infty]$. We denote the ℓ_r norm of a vector $a \in \mathbb{R}^n$ by

$$\|a\|_r = \left(\sum_{i=1}^{p} |a_i|^r \right)^{1/r}.$$

   Likewise, for matrix $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$, we write $|||\mathbf{A}|||_r$ to represent induced norms, i.e.

$$|||\mathbf{A}|||_r = \sup \{ \|\mathbf{A}x\|_r : x \in \mathbb{R}^{n_2}, \|x\|_r = 1 \}.$$

   For example, we write $|||\mathbf{A}|||_2$ to represent the spectral norm, $|||\mathbf{A}|||_1$ the maximum absolute column sum of the matrix, and $|||\mathbf{A}|||_\infty$, the maximum absolute row sum of the matrix. We use $\|\mathbf{A}\|_r$ to denote the ℓ_r norm of the vectorized matrix $\mathbf{A}$.

3. For a matrix $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$, we use a_j to denote the jth column of $\mathbf{A}$ (so $a_j \in \mathbb{R}^{n_1}$), and $\mathbf{A}_J$ to denote the column-wise concatenation of columns indexed by the set J. On the other hand, we use $a^{(i)}$ to represent the ith row in $\mathbf{A}$.

4. For a matrix $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$, we use $\mathbf{P}_\mathbf{A}$ to denote the projection of $\mathbb{R}^{n_2}$ onto the linear space generated by the rows of $\mathbf{A}$, i.e.,

$$\mathbf{P}_\mathbf{A} = \mathbf{A}^T (\mathbf{A}\mathbf{A}^T)^- \mathbf{A}$$

   where $-$ superscript represents taking Moore-Penrose inverse (matrix needs to be square).

Relational notations

For functions $g_1(x)$ and $g_2(x)$, we write

1. $g_1(x) \lesssim g_2(x)$ if there exists a universal constant $c \in (0, \infty)$ such that $g_1(x) \leqslant c g_2(x)$,

2. $g_1(x) \gtrsim g_2(x)$ if there exists a universal constant $c \in (0, \infty)$ such that $g_1(x) \geqslant c g_2(x)$,

3. $g_1(x) \asymp g_2(x)$ if $f(n) \lesssim g(n)$ and $g_1(x) \gtrsim g_2(x)$ simultaneously,

4. $g_1(x) = o(g_2(x))$ if $g_1(x)/g_2(x) \to 0$ as $x \to \infty$ (also applies when x is a vector value, and $x \to \infty$ refers to some or all of its components approaching ∞),

5. $g_1(x) = O(g_2(x))$ if $g_1(x)/g_2(x) \leqslant c$ as $x \to \infty$ for some $c \in (0, \infty)$ (also applies when x is a vector value, and $x \to \infty$ refers to some or all of its components approaching ∞).

Additionally,

1. For square matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ of the same dimension

$$\mathbf{A}_1 \lesssim \mathbf{A}_2$$

   if the matrix difference $\mathbf{A}_2 - \mathbf{A}_1$ is positive semi-definite.

2. If $X \in \mathbb{R}$ is a random variable and $a \in \mathbb{R}$ is some constant, we write

$$|X - a| = o_P(1)$$

   if X converges to a in probability, i.e. $X \to_p a$.

Chapter 2

ESTIMATION OF HIGH-DIMENSIONAL GRAPHICAL MODELS USING REGULARIZED SCORE MATCHING

2.1 Introduction

In this chapter, we propose a novel approach to high-dimensional (undirected) graphical model selection. Addressing the case of *continuous* but not necessarily Gaussian observations, the proposed method is based on the *score matching* loss, first introduced by Hyvärinen [2005] in the setting of image analysis. This work differs from that of Forbes and Lauritzen [2015], who studied score matching in Gaussian graphical models with symmetry constraints, and demonstrated that, when the number of variables p is fixed, the estimators derived from the score matching loss are asymptotically efficient in some special cases, but not in general. Our focus is instead on the use of score matching in high-dimensional problems, for which we consider regularization with an ℓ_1 penalty. We will refer to this graphical model selection technique as *regularized score matching*.

We motivate regularized score matching by noting the fact that it is computationally very convenient for any exponential family comprising continuous distributions. Indeed, the score matching loss is a positive semi-definite quadratic function. It follows that the solution path for the regularized score matching problem is piecewise linear and can be computed in entirety. Moreover, theoretical analysis can be based on familiar techniques. Most importantly, as we demonstrate for Gaussian graphical models, regularized score matching exhibits state-of-the-art statistical efficiency in high-dimensional settings. The method also performs well in our applications to non-Gaussian models, which include models that seem rather difficult to handle via other methods.

In the Gaussian setting, regularized score matching is structurally closest to pseudo-

likelihood methods with symmetry constraints, such as *SPACE* [Peng et al., 2009], *symmetric lasso* [Friedman et al., 2010] and *SPLICE* [Rocha et al., 2008]. A thorough discussion of these different methods is given by Khare et al. [2015] who also reformulate the SPACE objective function to ensure convergence of coordinate descent algorithms. They abbreviate their method as *CONCORD*. For brevity, we refer to these algorithms collectively as SPACE. We note that in contrast to regularized score matching, the SPACE methods do not have piecewise linear solution paths. Furthermore, as remarked before, the computational convenience of regularized score matching carries over to non-Gaussian settings.

A limitation of the original score matching framework introduced by Hyvärinen [2005] is that it requires the data to be generated from a distribution whose density is twice differentiable on $\mathbb{R}^p$. Hyvärinen [2007] proposed a generalization of the approach to the important case of non-negative data. For exponential families, the non-negative score matching loss is again a semidefinite quadratic function. We explore regularization of the non-negative score matching loss as a tool for estimating conditional independence graphs from high-dimensional non-negative data, and we establish consistency results on this method.

The remainder of the chapter is organized as follows. Section 2.2 provides the needed background on score matching and its applications. In Section 2.3, we describe the proposed method, *regularized* score matching. Implementation details are given in the Appendix. Section 2.4 provides sparsistency theory for both basic and non-negative regularized score matching. In Section 2.5, we present results of numerical experiments to compare the performance of the procedure with existing approaches. An application to RNAseq data is given in Section 2.6. We end with a discussion in Section 2.7. Proofs are deferred to the Appendix.

2.2 A review on score matching

Suppose X is a continuous random vector taking values in $\mathbb{R}^p$, with joint distribution F^*. Suppose further that F^* belongs to the family $\mathcal{F}$ that comprises all probability distributions with support equal to $\mathbb{R}^p$ and a twice differentiable density with respect to Lebesgue measure. We emphasize that in a statistical context the differentiability requirement is with respect to

data. We write f^* to denote the density of F^* and adopt the usual notation for the gradient (∇f) and Laplacian ($\Delta f = \sum_{j=1}^{p} \frac{\nabla^2 f}{\nabla x_j^2}$).

For a distribution $F \in \mathcal{F}$ with density f, define the divergence function

$$J(F) = \int_{\mathbb{R}^p} f^*(x) \left[\|\nabla \log f(x) - \nabla \log f^*(x)\|_2^2 \right] dx \tag{2.1}$$

as the expected squared distance between the gradients of the log-densities of the two distributions F and F^*. By choosing F to minimize (2.1), we are matching 'scores' with respect to the data vector x. Hence, (2.1) has been referred to as the *score matching loss*. It is evident from (2.1) that the score matching loss is uniquely minimized when $F = F^*$.

Upon initial inspection, optimization of $J(F)$ seems to require knowledge of F^* in an important way. However, Hyvärinen [2005] showed that, under mild regularity conditions, the score matching loss (2.1) can be rewritten as:

$$J(F) = \int_{\mathbb{R}^p} f^*(x) \left[\Delta \log f(x) + \frac{1}{2} \|\nabla \log f(x)\|_2^2 \right] dx \ + \ \text{const}, \tag{2.2}$$

where 'const' refers to a term independent of F. The key term in the integrand in (2.2) is the so-called Hyvärinen scoring rule

$$S(x, F) = \Delta \log f(x) + \frac{1}{2} \|\nabla \log f(x)\|_2^2.$$

The integral in (2.2) admits an empirical version in which the integration with respect to F is replaced by an average over an observed sample, which we arrange into a data matrix $\mathbf{x} \in \mathbb{R}^{n \times p}$. This leads to the *empirical score matching loss*

$$\hat{J}(\mathbf{x}, F) = \frac{1}{n} \sum_{i=1}^{n} S(x^{(i)}, F), \tag{2.3}$$

and the *score matching estimator* (SME)

$$\hat{F} = \arg\min_{F} \ \hat{J}(\mathbf{x}, F).$$

The score matching loss $J(F)$ was motivated by problems involving models whose distributions have an intractable normalization constant. Indeed, evaluating (2.2) and computing

the SME $\hat{F}$ requires no knowledge of the normalization constant, which is eliminated upon taking logarithmic derivatives with respect to x. Besides the imaging problems considered by Hyvärinen [2005], score matching has been applied to spatial statistics [Dawid and Musio, 2013] and neural networks [Köster and Hyvärinen, 2007, Vincent, 2011, Le et al., 2011].

The statistical properties of SMEs in classical large sample settings have been investigated by Hyvärinen [2005, 2007] and Forbes and Lauritzen [2015]. In particular, it has been shown that, under the usual regularity conditions, SMEs are asymptotically consistent and normal in large-sample theory. However, SMEs are not necessarily asymptotically efficient.

2.2.1 Extension to non-negative data

The partial integration arguments underlying (2.2) may fail to apply when considering distributions Q that are not supported on all of $\mathbb{R}^p$. In particular, when F is taken to be from $\mathcal{F}_+$, i.e. the family of distributions that are supported on $\mathbb{R}^p_+ = [0,\infty)^p$ with Lebesgue densities that are twice differentiable on $(0,\infty)^p$, then partial integration may not be possible due to discontinuities at points with zero coordinates. We thus consider the non-negative score matching loss,

$$J_+(F) = \int_{\mathbb{R}^p_+} f^*(x) \left[\left\| \nabla \log f(x) \circ x - \nabla \log f^*(x) \circ x \right\|_2^2 \right] dx, \tag{2.4}$$

as proposed in Hyvärinen [2007]. Here, '$\circ$' stands for the Hadamard product, that is, element-wise multiplication.

The score matching loss (2.1) can be thought of as a function of the Euclidean distance between the gradients of the model density f and true density f^* with respect to a hypothetical location parameter μ, evaluated at 0. That is, we may write (2.1) as

$$J(F) = \int_{\mathbb{R}^p} f^*(\mathbf{x}) \left[\left\| \nabla_{\mu=0} \log f(x+\mu) - \nabla_{\mu=0} \log f^*(x+\mu) \right\|_2^2 \right] dx.$$

Likewise, the non-negative score matching loss compares the gradient of the model density f and true density f^* with respect to a hypothetical scale parameter σ evaluated at 1,

$$J_+(F) = \int_{\mathbb{R}^p_+} f^*(\mathbf{x}) \left[\left\| \nabla_{\sigma=1} \log f(x \circ \sigma) - \nabla_{\sigma=1} \log f^*(x \circ \sigma) \right\|_2^2 \right] dx.$$

Under suitably adjusted regularity conditions, Hyvärinen [2007] showed that the non-negative score matching loss from (2.4) can be simplified into

$$J_+(F) = \int_{\mathbb{R}^p_+} p(x) S_+(x, F)\, dx \;+\; \text{const} \tag{2.5}$$

with scoring rule

$$S_+(x, F) = \sum_{j=1}^{p} \left[2x_j \frac{\partial \log f(x)}{\partial x_j} + x_j^2 \frac{\partial^2 \log f(x)}{\partial x_j^2} + \frac{1}{2} x_j^2 \left(\frac{\partial \log f(x)}{\partial x_j} \right)^2 \right]. \tag{2.6}$$

For a data matrix $\mathbf{x} \in \mathbb{R}^{n \times p}$, one obtains the *empirical non-negative score matching loss*

$$\hat{J}_+(\mathbf{x}, F) = \frac{1}{n} \sum_{i=1}^{n} S_+(x^{(i)}, F), \tag{2.7}$$

and the *non-negative score matching estimator* (SME_+)

$$\hat{F}_+ = \arg\min_F \; \hat{J}_+(\mathbf{x}, F).$$

Again, under the usual regularity conditions, the estimator $\hat{F}_+$ is asymptotically consistent and normal in traditional large-sample theory.

2.2.2 Score matching in exponential families

Hyvärinen [2007] and Forbes and Lauritzen [2015] have shown that the SME has a convenient closed form as a rational function of the data when $\mathcal{P}$ is an exponential family. Hyvärinen [2007] showed the same for SME_+ for the example of truncated normal distributions. As they provide the basis for our later work, we revisit these results for both SME and SME_+.

Let $\mathcal{F} = (F_\theta : \theta \in \mathcal{T})$ be an exponential family with natural parameter space $\mathcal{T}$. Suppose that the distributions F_θ have their common support equal to either $\mathcal{X} = \mathbb{R}^p$ or $\mathcal{X} = \mathbb{R}^p_+$, and that $\mathcal{F}$ is dominated by Lebesgue measure on $\mathbb{R}^p$. Assuming that the sufficient statistics $t(x)$ take values in $\mathbb{R}^s$, the log-densities of the distributions F_θ have the form

$$\log f_\theta(x) = \theta^T t(x) - \psi(\theta) + b(x), \quad x \in \mathcal{X}, \tag{2.8}$$

and

$$\mathcal{T} = \left\{ \theta \in \mathbb{R}^s \;:\; \psi(\theta) = \log \int_{\mathcal{X}} e^{\theta^T t(x)} dx < \infty \right\}. \tag{2.9}$$

Lemma 1. *Let $\mathbf{x} \in \mathbb{R}^{n \times p}$ be a data matrix, and suppose $\mathcal{F} = (F_\theta : \theta \in \mathcal{T})$ is an exponential family characterized by (2.8) and (2.9). If $\mathcal{F}$ has support $\mathcal{X} = \mathbb{R}^p$, then the empirical score matching loss $\hat{J}(\mathbf{x}, F_\theta)$ is a quadratic function in θ with*

$$\hat{J}(\mathbf{x}, F_\theta) = \frac{1}{2}\, \theta^T \mathbf{\Gamma}(\mathbf{x}) \theta + \gamma(\mathbf{x})^T \theta + c(\mathbf{x}), \tag{2.10}$$

where $\mathbf{\Gamma}(\mathbf{x})$ is a positive semidefinite $s \times s$ matrix, and $\gamma(\mathbf{x})$ is an s-vector. The same is true for $\hat{J}_+(\mathbf{x}, F_\theta)$ when $\mathcal{F}$ has support $\mathcal{X} = \mathbb{R}^p_+$.

Proof. For $j = 1, \ldots, m$ and $x \in \mathbb{R}^p$, define the s-vectors

$$h_j(x) = \frac{\partial}{\partial x_j} t(x), \qquad\qquad h_{jj}(x) = \frac{\partial^2}{\partial x_j^2} t(x).$$

It then follows from (2.8) that $\hat{J}(\mathbf{x}, F_\theta)$ can be expressed in the claimed form with

$$\mathbf{\Gamma}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} h_j(x^{(i)}) h_j(x^{(i)})^T, \tag{2.11}$$

$$\gamma(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(\frac{\partial}{\partial x_j} b(x^{(i)}) \right) h_j(x^{(i)})^T + \Delta t(x^{(i)}), \tag{2.12}$$

$$c(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \left\| \nabla b(x^{(i)}) \right\|_2^2 + \Delta b(x^{(i)}). \tag{2.13}$$

For non-negative score matching, $\hat{J}_+(\mathbf{x}, F_\theta)$ admits the claimed form with

$$\mathbf{\Gamma}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij}^2 h_j(x^{(i)}) h_j(x^{(i)})^T, \tag{2.14}$$

$$\gamma(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(\frac{\partial}{\partial x_j} b(x^{(i)}) \right) h_j(x^{(i)})^T + x_{ij}^2 h_{jj}(x^{(i)})^T + 2 x_j^{(i)} h_j(x^{(i)})^T, \tag{2.15}$$

$$c(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{1}{2} x_{ij}^2 \left(\frac{\partial}{\partial x_j} b(x^{(i)}) \right)^2 + x_{ij}^2 \frac{\partial^2}{\partial x_j^2} b(x^{(i)}) + 2 x_{ij} \frac{\partial}{\partial x_j} b(x^{(i)}), \tag{2.16}$$

where the x_{ij} are the entries of the $n \times p$ data matrix $\mathbf{x}$. □

Lemma 1 implies that, when working with exponential families, both score matching objectives are quadratic functions of the unknown parameter vector θ. A score matching estimator $\hat{\theta}$ thus satisfies a set of *linear* estimating equations

$$\hat{\theta}^T \mathbf{\Gamma}(\mathbf{x}) + \gamma(\mathbf{x}) = 0. \tag{2.17}$$

2.2.3 Pairwise interaction models

The most basic class of exponential families that appear in graphical modeling are pairwise interaction models with log-densities

$$\log f_\theta(x) \;=\; \sum_{1 \leqslant j \leqslant k \leqslant p} \theta_{jk} t_{jk}(x_j, x_k) - \psi(\theta) + \phi(x), \quad x \in \mathcal{X} \subseteq \mathbb{R}^p. \tag{2.18}$$

Here, the t_{jk} are sufficient statistics that depend only on the jth and kth coordinate of x, and the θ_{jk} are interaction parameters. If F_θ denotes the distribution with density given by (2.18), then the Hammersley-Clifford Theorem implies that an edge between nodes j and k exists in the conditional independence graph of F_θ if and only if θ_{jk} is nonzero. The specific models we consider later either exactly have the form in (2.18) or are closely related extensions with log-densities

$$\log f_\theta(x) \;=\; \sum_{g_1=1}^{G_1} \sum_{j \leqslant k} \theta_{jk}^{(g_1)} t_{jk}^{(g_1)}(x_j, x_k) + \sum_{g_2=1}^{G_2} \sum_{j=1}^{p} \theta_j^{(g_2)} t_j^{(g_2)}(x_j) - \psi(\theta) + \phi(x), \tag{2.19}$$

where pairwise interactions may be of G_1 different types and we also include G_2 sets of sufficient statistics $t_j^{(g_2)}$ depending on the individual coordinates. The latter appear, for instance, when allowing distributions to vary in location. The distribution F_θ defined by (2.19) has no edge between j and k in its conditional independence graph if and only if $\theta_{jk}^{(1)} = \cdots = \theta_{jk}^{(G_1)} = 0$.

In our study of score matching methods for models of the type (2.18) or (2.19), it will be convenient to introduce the symmetric $p \times p$ interaction matrix $\mathbf{\Theta}$ with entries

$$\mathbf{\Theta}_{jk} = \begin{cases} \theta_{jk} & \text{if } \; j \leqslant k, \\ \theta_{kj} & \text{if } \; j > k. \end{cases}$$

Lemma 2. *Let $\mathcal{F}$ to be the pairwise interaction model given by (2.18) with symmetric $p \times p$ interaction matrix $\mathbf{\Theta}$. If $\mathcal{F}$ has support $\mathcal{X} = \mathbb{R}^p$, then the empirical score matching loss $\hat{J}(\mathbf{x}, F_\theta)$ equals*

$$\frac{1}{2} vec(\mathbf{\Theta})^T \mathbf{\Gamma}(\mathbf{x}) vec(\mathbf{\Theta}) + \gamma(\mathbf{x})^T vec(\mathbf{\Theta}) + c(\mathbf{x}) \tag{2.20}$$

for a symmetric $p^2 \times p^2$ matrix $\mathbf{\Gamma}(\mathbf{x})$ that is block-diagonal, with all blocks of size $p \times p$. The same is true for $\hat{J}_+(\mathbf{x}, F_\theta)$ when $\mathcal{F}$ has support $\mathcal{X} = \mathbb{R}^p_+$.

Proof. By (2.11) and (2.14), it suffices to show that there exists a block-diagonal matrix $\mathbf{\Gamma}_j(x)$ such that

$$\theta^T h_j(x) h_j(x)^T \theta \;=\; \mathrm{vec}(\mathbf{\Theta})^T \mathbf{\Gamma}_j(x) \mathrm{vec}(\mathbf{\Theta}), \tag{2.21}$$

where $\theta = (\theta_{jk} : j \leqslant k)$.

Now,

$$\begin{aligned} h_j(x)^T \theta &= \sum_{k \geqslant j} \frac{\partial}{\partial x_j} t_{jk}(x_j, x_k) \theta_{jk} + \sum_{k<j} \frac{\partial}{\partial x_j} t_{kj}(x_k, x_j) \theta_{kj} \\ &= \sum_{k \geqslant j} \frac{\partial}{\partial x_j} t_{jk}(x_j, x_k) \mathbf{\Theta}_{kj} + \sum_{k<j} \frac{\partial}{\partial x_j} t_{kj}(x_k, x_j) \mathbf{\Theta}_{kj}. \end{aligned}$$

Define a vector $\bar{h}_j(x) \in \mathbb{R}^{p^2}$, indexed by pairs (k, l) with $1 \leqslant k, l \leqslant m$, by setting the entries to

$$\bar{h}_j(x)_{kl} = \begin{cases} \frac{\partial}{\partial x_k} t_{kl}(x_k, x_l) & \text{if } \; j = k \leqslant l, \\ \frac{\partial}{\partial x_k} t_{lk}(x_k, x_j) & \text{if } \; j = k > l, \\ 0 & \text{if } \; j \neq k. \end{cases} \tag{2.22}$$

Then $h_j(x)^T \theta = \bar{h}_j(x) \mathrm{vec}(\mathbf{\Theta})$ and (2.21) holds with $\mathbf{\Gamma}_j(x) = \bar{h}_j(x) \bar{h}_j(x)^T$, which is block-diagonal as it is zero with the exception of the $p \times p$ block indexed by pairs (k, l) with $k = j$. □

Remark 1. When $\mathcal{F}$ is a model as specified in (2.19), then the empirical (non-negative) score matching loss may still be represented as an explicit quadratic form with a block-diagonal symmetric matrix $\mathbf{\Gamma}(\mathbf{x})$ as in (2.20). However, $\mathbf{\Gamma}(\mathbf{x})$ is then of size $(G_1 p^2 + G_2 p) \times (G_1 p^2 + G_2 p)$, and its p diagonal blocks are of size $(G_1 p + G_2) \times (G_1 p + G_2)$. The jth block has its rows and columns corresponding to the jth columns of each of $\mathbf{\Theta}^{(1)}, \ldots, \mathbf{\Theta}^{(A)}$ as well $(\theta_j^{(1)}, \ldots, \theta_j^{(L)})$.

Example 1. If the exponential family is taken to be the family of centered multivariate normal distributions with precision matrix $\mathbf{K} = (\kappa_{jk})$, then the support is $\mathcal{X} = \mathbb{R}^p$ and

$$f_{\mathbf{K}}(x) \propto \exp\left\{-\frac{1}{2}x^T\mathbf{K}x\right\}, \quad x \in \mathbb{R}^p. \tag{2.23}$$

With

$$\nabla \log f_{\mathbf{K}}(x) = -\mathbf{K}x, \qquad \Delta \log f_{\mathbf{K}}(x) = -\sum_{j=1}^{p} \kappa_{jj},$$

and dropping a term that is constant in $\mathbf{K}$, the empirical score matching loss from (2.2) takes the form

$$-\operatorname{tr}(\mathbf{K}) + \frac{1}{2}\operatorname{tr}(\mathbf{K}\mathbf{K}\mathbf{W}), \tag{2.24}$$

where

$$\mathbf{W} = \frac{\mathbf{x}^T\mathbf{x}}{n}$$

is the empirical covariance matrix (under knowledge of zero mean). Lemma 2 applies with $t_{jk}(x_j, x_k) = x_j x_k$, in which case the matrix $\mathbf{\Gamma}_j(x)$ constructed in the proof of the lemma does not depend on j, other than through the location of the nonzero block. Indeed, (2.20) holds with $\mathbf{\Gamma}(\mathbf{x}) = \mathbf{I}_{p\times p} \otimes \mathbf{W}$ and $\gamma(\mathbf{x}) = \operatorname{vec}(\mathbf{I}_{p\times p})$, where $\mathbf{I}_{p\times p}$ is the $p \times p$ identity matrix. Clearly, $\mathbf{\Gamma}(\mathbf{x})$ is positive definite if and only if $\mathbf{W}$ is as well. If $\mathbf{W}$ is invertible then SME of $\mathbf{K}$ is $\hat{\mathbf{K}} = \mathbf{W}^{-1}$ and coincides with the maximum likelihood estimator.

Example 2. Consider truncated normal densities of the form

$$f_{\mathbf{K}}(x) \propto \exp\left\{-\frac{1}{2}x^T\mathbf{K}x\right\}, \quad x \in \mathbb{R}^p_+. \tag{2.25}$$

Using κ_j to denote the jth column of $\mathbf{K}$, it can be shown that the empirical non-negative score matching objective is given by

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} 2x_{ij}x^{(i)T}\kappa_j - x_{ij}^2\kappa_{jj} + \frac{1}{2}\kappa_j^T\left(x_{ij}^2 x^{(i)}x^{(i)T}\right)\kappa_j. \tag{2.26}$$

The loss can be written as in (2.10) with $\mathbf{\Gamma}(\mathbf{x})$ a block diagonal $p^2 \times p^2$ matrix, whose jth block is given by

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 x^{(i)}x^{(i)T}.$$

Moreover, $\gamma(\mathbf{x}) = 2w + w_{\text{diag}}$, where $w = \text{vec}(\mathbf{W})$ and $w_{\text{diag}} = \text{vec}(\text{diag}(\mathbf{W}))$. The maximum likelihood estimator for $\mathbf{K}$ has no closed form due to intractable normalizing constants.

Example 3. Finally, consider the family of distributions with densities of the form

$$f_{\mathbf{B}^{(2)},\mathbf{B},\mathbf{b}}(x) \propto \exp\left\{ \sum_{1 \leqslant j \neq k \leqslant p} \beta^{(2)}_{jk} x_j^2 x_k^2 + \sum_{j,k=1}^{p} \beta_{jk} x_j x_k + \sum_{j=1}^{p} \beta_j x_j \right\}, \qquad x \in \mathbb{R}^p. \tag{2.27}$$

Here, $\mathbf{b} = (\beta_1, \ldots, \beta_p)^T$ is an p-vector, and $\mathbf{B} = (\beta_{jk})$ and $\mathbf{B}^{(2)} = (\beta^{(2)}_{jk})$ are symmetric $p \times p$ interaction matrices, the latter having a zero diagonal. This family is a class of distributions with normal conditionals, with densities that need not be unimodal [Arnold et al., 1999, Gelman and Meng, 1991]. This family is intriguing from the perspective of graphical modeling as, in contrast to the Gaussian case, conditional dependence may also express itself in the variances. For conditional independence of X_j and X_k both β_{jk} and $\beta^{(2)}_{jk}$ need to vanish.

By Remark 1, the empirical score matching loss for the family from (2.27) can be written as a quadratic function with the quadratic term given by block-diagonal matrix $\mathbf{\Gamma}(\mathbf{x})$ of size $(2p^2 + p) \times (2p^2 + p)$. The blocks are of size $(2p + 1) \times (2p + 1)$, and the jth block has its rows and columns corresponding to the jth columns of $\mathbf{B}$ and $\mathbf{B}^{(2)}$ and the jth entry in $\mathbf{b}$.

2.3 Regularized score matching

Building on the ideas underlying methods such as glasso, neighborhood selection and SPACE, we augment the score matching loss with a sparsity-promoting penalty. Our focus is on the most basic case of an ℓ_1 penalty but other regularization schemes could be considered instead; see also Example 3 below.

Using the generic representation given in Lemma 1, for an exponential family, the proposed method is based on minimizing the objective

$$\hat{J}_\lambda(\theta) = \frac{1}{2}\, \theta^T \mathbf{\Gamma}(\mathbf{x}) \theta - \gamma(\mathbf{x})^T \theta + c(\mathbf{x}) + \lambda \|\theta\|_1, \quad \theta \in \mathbb{R}^s, \tag{2.28}$$

where $\mathbf{\Gamma}(\mathbf{x})$ is positive semidefinite and $\lambda \geqslant 0$ is a tuning parameter that controls the sparsity level. Larger values of λ yield sparser solutions, and $\lambda = 0$ gives the unregularized SME.

Since $\mathbf{\Gamma}(\mathbf{x})$ is positive semidefinite, the function $\hat{J}_\lambda(\theta)$ is convex but in the settings of interest here $\mathbf{\Gamma}(\mathbf{x})$ will be singular and $\hat{J}_\lambda(\theta)$ will not be strictly convex.

The regularized score matching objective from (2.28) is similar to the lasso objective in linear regression [Tibshirani, 1996], where the function to be minimized takes the special form

$$\frac{1}{2}\|y - X\theta\|_2^2 + \|\theta\|_1, \tag{2.29}$$

for a 'response vector' y and a 'design matrix' X. In the applications we have in mind (2.28) cannot be written exactly as in (2.29) because the vector $\gamma(\mathbf{x})$ is generally not in the column span of $\mathbf{\Gamma}(\mathbf{x})$. However, we may adapt existing optimization methods for lasso to solve the regularized score matching problem.

If the considered exponential family is supported on $\mathcal{X} = \mathbb{R}^p$ and we use the loss function (2.3), then we call the minimizer of (2.28) the regularized score matching estimator (rSME). If $\mathcal{X} = \mathbb{R}^p_+$ and we use the loss function (2.7), then we abbreviate with rSME$_+$. In specific instances of graphical models, we may apply the ℓ_1 penalty only to those coordinates of θ whose vanishing corresponds to absence of edges in a conditional independence graph. If the subset $\mathcal{E} \subseteq \{1, \ldots, s\}$ holds the relevant coordinates then we use the penalty

$$\|\theta\|_{1,\mathcal{E}} \equiv \sum_{j \in \mathcal{E}} |\theta_j|.$$

Example 1 (cont.). For the (centered) Gaussian case considered in Example 1, the target of estimation is the symmetric precision matrix $\mathbf{K}$. The conditional independence graph corresponds to the pattern of zeros in the off-diagonal entries of $\mathbf{K}$ and the rSME is

$$\hat{\mathbf{K}} = \arg \min_{\mathbf{K} \in \mathrm{Sym}_p} \left\{ -\operatorname{tr}(\mathbf{K}) + \frac{1}{2}\operatorname{tr}(\mathbf{K}\mathbf{K}\mathbf{W}) + \lambda \|\mathbf{K}\|_{1,\mathrm{off}} \right\}, \tag{2.30}$$

where W is the empirical covariance matrix and $\|\mathbf{K}\|_{1,\mathrm{off}} = \|\mathbf{K}\|_{1,\mathcal{E}}$ penalizes only the off-diagonal entries indexed by $\mathcal{E} = \{(j,k) : j \neq k\}$. We emphasize that while in this example the natural parameter space is the positive definite cone, we propose minimizing simply over the entire space of symmetric $p \times p$ matrices, denoted by Sym_p. As our interest is primarily

in graph selection, we do not enforce positive definiteness of $\hat{\mathbf{K}}$, which is in line with methods such as SPACE or neighborhood selection; compare Khare et al. [2015].

We remark that evaluating the function from (2.30) at an asymmetric matrix $\mathbf{K}$ as well as its transpose $\mathbf{K}^T$ gives the same value. By convexity, minimizing over all $p \times p$ matrices gives a solution in Sym_p, which then must equal $\hat{\mathbf{K}}$.

Example 2 (cont.). In the truncated normal family from Example 2, the conditional independence graph corresponds again to the zero pattern in the off-diagonal entries of the positive definite interaction matrix $\mathbf{K}$. Proceeding in analogy to the Gaussian case, we define the rSME$_+$ as the minimizer $\hat{\mathbf{K}}_+$ of the objective given by (2.26) with the penalty $\lambda\|\mathbf{K}\|_{1,\mathrm{off}}$ added on. Again, we ignore the positive definiteness requirement and minimize the penalized non-negative score matching loss with respect to $\mathbf{K} \in \mathrm{Sym}_p$.

Example 3 (cont.). For the family of distributions with normal conditionals from Example 3, we would like a penalty to induce joint sparsity in the two symmetric interaction matrices $\mathbf{B}$ and $\mathbf{B}^{(2)}$, because an edge between nodes j and k is absent from the conditional independence graph if and only both $\mathbf{B}$ and $\mathbf{B}^{(2)}$ have their (j,k) entries zero. For this purpose, it is natural to adopt the group lasso penalty [Yuan and Lin, 2006]. The rSME is then obtained by minimizing the empirical score matching loss augmented by the penalty

$$\lambda \sum_{j \neq k} \sqrt{(\beta_{jk})^2 + (\beta^{(2)}_{jk})^2}.$$

Ignoring again any refined constraints from the natural parameter space of the family, we propose minimizing the penalized loss with respect to $\mathbf{b} \in \mathbb{R}^p$ and $\mathbf{B}, \mathbf{B}^{(2)} \in \mathrm{Sym}_p$. Since the group lasso is applied with small groups (of size 2), the problem would be suitable for application of exact block-coordinate descent as discussed in Foygel and Drton [2010b].

2.3.1 Uniqueness of rSME

In the setup from Lemma 1, we may write

$$\mathbf{\Gamma}(\mathbf{x}) = \mathbf{H}(\mathbf{x})^T \mathbf{H}(\mathbf{x}) \tag{2.31}$$

for an $np \times s$ matrix $\mathbf{H}(\mathbf{x})$; recall (2.11) and (2.14). Based on the arguments leading to Lemmas 3 and 5 in Tibshirani [2013], the function $\hat{J}_\lambda(\theta)$ from (2.28) has a unique minimizer $\hat{\theta}$ as long as $\lambda > 0$ and the columns of $\mathbf{H}(\mathbf{x})$ are in *general position*. To clarify, suppose that $\mathcal{U} \subset \mathbb{R}^{np}$ is a collection of $|\mathcal{U}| = N$ vectors. Then $\mathcal{U}$ is in general position if for all $N' < \min\{np, s\}$, all choices of vectors $u_1, \ldots, u_{N'+1} \in \mathcal{U}$ and signs $\sigma_1, \ldots, \sigma_{N'+1} \in \{-1, 1\}$, the affine span of $\sigma_1 u_1, \ldots, \sigma_{N'+1} u_{k+1}$ does not contain any vector u or $-u$ for $u \in \mathcal{U} \backslash \{u_1, \ldots, u_{N'+1}\}$.

The graphical models we are interested in are pairwise interaction models that have additional special structure in that the matrix $\mathbf{\Gamma}(\mathbf{x})$ is block-diagonal with p blocks of equal size; recall Lemma 2 and Remark 1. Denote the diagonal blocks by $\mathbf{\Gamma}_1(\mathbf{x}), \ldots, \mathbf{\Gamma}_p(\mathbf{x})$, which in the setup from (2.19) are of size $(G_1 p^2 + G_2 p) \times (G_1 p^2 + G_2 p)$. Each block is the sum of n symmetric rank one matrices and we have the decomposition

$$\mathbf{\Gamma}_j(\mathbf{x}) = \mathbf{H}_j(\mathbf{x})^T \mathbf{H}_j(\mathbf{x}), \quad j = 1, \ldots, p. \tag{2.32}$$

The n columns of each of the matrices $\mathbf{H}_j(\mathbf{x})$ were specified in (2.22). It now holds that the regularized score matching problem from (2.28) has a unique minimizer provided each one of the $n \times (G_1 p + G_2)$ blocks $\mathbf{H}_1(\mathbf{x}), \ldots, \mathbf{H}_p(\mathbf{x})$ defined in (2.32) has its columns in general position.

Example 1 (cont.). In the Gaussian case, $\mathbf{H}_1(\mathbf{x}) = \cdots = \mathbf{H}_p(\mathbf{x}) = \mathbf{x}$. By the Lemma in Okamoto [1973], the set of matrices $\mathbf{x}$ that fail to be in general position has measure zero. The rSME $\hat{\mathbf{K}}$ is unique almost surely when data are generated from a continuous joint distribution.

Example 2 (cont.). In the truncated normal case, $\mathbf{H}_j(\mathbf{x})$ is equal to the matrix obtained from $\mathbf{x}$ by multiplying each column element-wise with x_j, the jth column of $\mathbf{x}$. The Lemma in Okamoto [1973] implies that the rSME$_+$ is unique almost surely.

For the normal conditionals model from Example 3, almost sure uniqueness would have to be derived by appealing to results on uniqueness of group lasso [Roth and Fischer, 2008].

2.3.2 Piecewise linear paths

The rSME depends on the regularization parameter λ. In this section we make this explicit and denote it by $\hat{\theta}^\lambda$. Adopting standard language, we refer to the set of $\hat{\theta}^\lambda$ obtained by varying λ as the *solution path* and call this path *piecewise linear* if there exists $0 = \lambda_0 < \lambda_1 < \ldots < \lambda_M = \infty$ and $\xi_0, \ldots, \xi_{M-1} \in \mathbb{R}^p$ such that $\hat{\theta}^\lambda = \hat{\theta}^{\lambda_m} + (\lambda - \lambda_m)\xi_r$ for $\lambda \in [\lambda_m, \lambda_{m+1}]$. Piecewise linear solution paths have the appeal that the entire solution path can be found by calculating the change points λ_m and associated slopes ξ_m.

The next lemma is a consequence of the quadratic nature of the score matching objective for exponential families, and holds for the lasso problem (1.3) as well.

Lemma 3. *The solution path $\hat{\theta}^\lambda$ for the regularized score matching problem from (2.28) is piecewise linear.*

Proof. An s-vector z belongs to $\partial\|\theta\|_1$, the subdifferential of the ℓ_1 norm, if

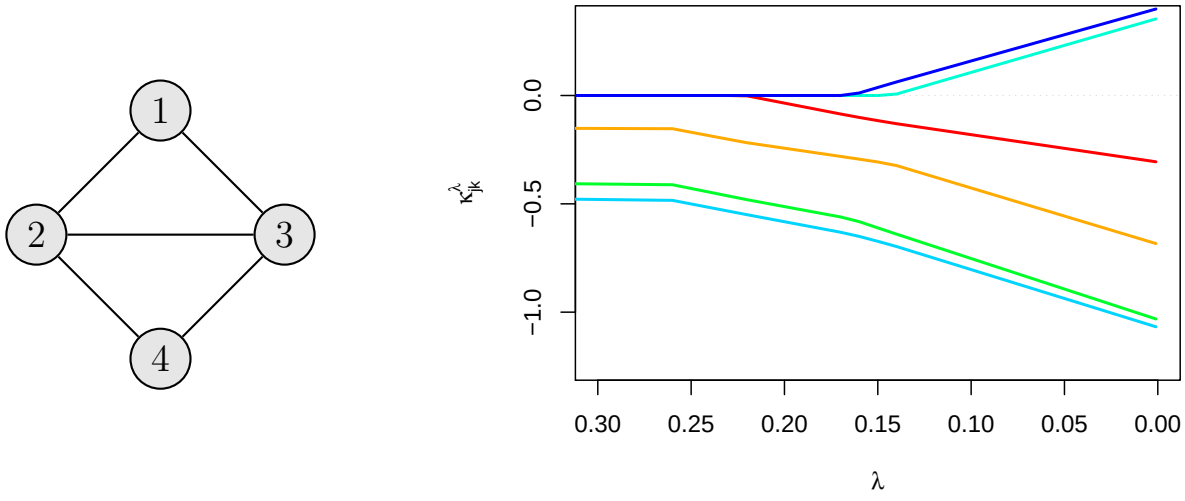
$$z_j = \begin{cases} \operatorname{sign}(\theta_j) & \text{if } \theta_j \neq 0, \\ \in [-1, 1] & \text{if } \theta_j = 0. \end{cases} \tag{2.33}$$

The Karush-Kuhn-Tucker (KKT) conditions characterizing optimality in (2.28) are

$$\mathbf{\Gamma}(\mathbf{x})\hat{\theta} - \gamma(\mathbf{x}) + \lambda\hat{z} = 0, \quad \hat{z} \in \partial\|\hat{\theta}\|_1. \tag{2.34}$$

The linear relationship between $\hat{\theta}$ and λ (for "fixed" $\hat{z}$) implies the claim. □

While straightforward to show, the property of piecewise linear paths is special to the score matching method we propose. Other methods that give symmetric estimates of precision matrices in Gaussian graphical models, such as glasso or the SPACE-type methods discussed in Khare et al. [2015] do not have piecewise linear solution paths. This said, piecewise linear paths also arise in neighborhood selection [Meinshausen and Bühlmann, 2006], which, however, is a formulation without symmetry. Note also that when using a group lasso penalty as suggested for Example 3, rSME solution paths are no longer piecewise linear.



(a) (b)

Figure 2.1: (a) A conditional independence graph with $p = 4$ nodes. (b) rSME solution path for Gaussian graphical modeling ($p = 4$, $n = 12$).

Example 1 (cont.). In the Gaussian model, the KKT conditions state that $\hat{\mathbf{K}}$ is a solution to (2.28) if and only if

$$(\mathbf{I}_{p\times p} \otimes \mathbf{W})\,\mathrm{vec}(\hat{\mathbf{K}}) - \mathrm{vec}\,(\mathbf{I}_{p\times p}) + \lambda\hat{z} = 0 \tag{2.35}$$

for $\hat{z} \in \partial\|\hat{\mathbf{K}}\|_{1,\mathrm{off}}$, which in slight abuse of notation, we take to mean that

$$\hat{z}_{jk} = \begin{cases} 0 & \text{if } j = k, \\ \mathrm{sign}(\hat{\kappa}_{jk}) & \text{if } \hat{\kappa}_{jk} \neq 0 \text{ and } j \neq k, \\ \in [-1, 1] & \text{if } \hat{\kappa}_{jk} = 0 \text{ and } j \neq k. \end{cases} \tag{2.36}$$

The first case accounts for the fact that the objective is smooth in the diagonal entries of

the precision matrix, which are not penalized. Combining (2.35) and (2.36), we have that

$$-1 + \sum_{k=1}^{p} w_{jk}\hat{\kappa}_{jk} = 0, \quad j = 1, \ldots, p, \tag{2.37}$$

$$\sum_{\ell=1}^{p} w_{j\ell}\hat{\kappa}_{\ell k} + \sum_{\ell=1}^{p} w_{k\ell}\hat{\kappa}_{\ell j} + \lambda\hat{z}_{jk} = 0, \quad 1 \leqslant j \neq k \leqslant p. \tag{2.38}$$

A Gaussian solution path is shown in Figure 2.1b, with the horizontal axis transformed to $t(\lambda) = \sum_{j \neq k} |\hat{\kappa}^{\lambda}_{jk}|$. The data were drawn from a multivariate normal distribution with the conditional independence graph from Figure 2.1a, with sample size $n = 12$. We note that, as one would hope, the coefficient that last enters the solution corresponds to the absent edge $(1, 4)$.

2.3.3 Implementation

The piecewise linear solution path for regularized score matching can be computed using Algorithm 1, which is an adaptation of the LARS-Lasso algorithm for linear regression [Efron et al., 2004]. It is also a special case of the algorithm found in Rosset and Zhu [2007]. In our pseudocode, $\hat{S}$ is the current active set, i.e. $\hat{S} = \{j : \theta_j^{\lambda} \neq 0\}$ for the currently relevant value of the regularization parameter λ.

In the Gaussian and truncated Gaussian case, the algorithm stops when the active set has size $|\hat{S}| = \min\{n, p\}p$. For larger active sets the matrix $\mathbf{\Gamma}_{\hat{S}\hat{S}}$ is not invertible. Finding the step size in Algorithm 1 requires $\mathcal{O}(\min\{n, p\}p)$ operations, while the inversion step is at its worst $\mathcal{O}(|\hat{S}|^2) = \mathcal{O}(\min\{n, p\}^2 p^2)$. Overall, the complexity of Algorithm 1 can be found to be $\mathcal{O}(\min\{n, p\}^3 p^2)$; the heaviest cost comes from the matrix inversion step.

For large-scale problems, LARS-type algorithms may be slow and coordinate-descent methods are popular alternatives [see e.g. Friedman et al., 2007]. Algorithm 2 describes a coordinate-descent algorithm to minimize the regularized score matching objective from (2.28). It entails updating one coordinate, or one element in the parameter vector/matrix, such that it minimizes the objective function while holding all others as constant, until a convergence criterion is satisfied. Results in Tseng [2001] ensure convergence of Algorithm 2.

Algorithm 1

Initialize $\hat{\theta} = 0$

Initialize $\hat{S} = \arg\max_j \left| \left(\mathbf{\Gamma}(\mathbf{x})\hat{\theta} + \gamma(\mathbf{x}) \right)_j \right|$

Initialize $\xi_{\hat{S}} = -\text{sign}\left[\left(\mathbf{\Gamma}(\mathbf{x})\hat{\theta} + \gamma(\mathbf{x}) \right)_{\hat{S}} \right]$

Initialize $\xi_{\hat{S}^c} = 0$

while $\left\| \mathbf{\Gamma}(\mathbf{x})\hat{\theta} + \gamma(\mathbf{x}) \right\|_\infty > 0$ and $\mathbf{\Gamma}(\mathbf{x})_{\hat{S}\hat{S}}$ is invertible **do**

  $\eta_1 \leftarrow \min\{\eta > 0 : |(\mathbf{\Gamma}(\mathbf{x})\hat{\theta} + \gamma(\mathbf{x})|_j = |(\mathbf{\Gamma}(\mathbf{x})\hat{\theta} + \gamma(\mathbf{x})|_{\hat{S}},\ j \notin \hat{S}\}$.

  $\eta_2 \leftarrow \min\{\eta > 0 : (\hat{\theta} + \eta\xi)_j = 0,\ j \in \hat{S}\}$.

  $\eta \leftarrow \min\{\eta_1, \eta_2\}$.

  $\hat{\theta} \leftarrow \hat{\theta} + \eta\xi$

  if $\eta = \eta_1$ **then**

    Add variable that attains equality to $\hat{S}$.

  else

    Remove variable that attains 0 from $\hat{S}$.

  end if

  $\xi_{\hat{S}} \leftarrow (\mathbf{\Gamma}(\mathbf{x})_{\hat{S}\hat{S}})^{-1} \text{sign}(\hat{\theta}_{\hat{S}})$

end while

Example 1 (cont.). For the Gaussian case, the coordinate descent procedure alternates between updating the diagonal entries and off-diagonal entries, by manipulating the estimating equations (2.37) and (2.38) accordingly. The updates are of the form

$$\kappa_{jj}^{(t+1)} \leftarrow \frac{1 - \sum_{j' \neq j} w_{jj'} \kappa_{jj'}^{(t)}}{w_{jj}},$$

$$\kappa_{jk}^{(t+1)}, \kappa_{kj}^{(t+1)} \leftarrow \text{Soft}\left(\frac{-\sum_{j' \neq j} w_{jj'} \kappa_{j'k}^{(t)} - \sum_{k' \neq k} w_{jk'} \kappa_{k'k}^{(t)}}{w_{jj} + w_{kk}}, \frac{2\lambda}{w_{jj} + w_{kk}} \right),$$

Algorithm 2

Input: Initial estimate $\hat{\theta}^{(0)}$

Input: t_{max}, maximum number of iterations

Input: ϵ, the maximal tolerance level

Initialize $t \leftarrow 1$

Initialize crit $\leftarrow \epsilon + 1$

while crit $> \epsilon$ or $t < t_{max}$ **do**

  $\hat{\theta}^{(t)} \leftarrow \hat{\theta}^{(t-1)}$

  for $j \leftarrow 1, 2, \ldots, s$ **do**

    $\hat{\theta}_j^{(t)} \leftarrow \text{Soft}\left(\frac{-(\mathbf{\Gamma}(\mathbf{x})_{-j,j})^T \hat{\theta}_{-j}^{(t)} - \gamma(\mathbf{x})_j}{\mathbf{\Gamma}(\mathbf{x})_{jj}}, \frac{\lambda}{\mathbf{\Gamma}(\mathbf{x})_{jj}}\right)$.

  end for

  crit $\leftarrow \|\hat{\theta}^{(t)} - \hat{\theta}^{(t-1)}\|_1$

  $t \leftarrow t + 1$

end while

for $j, k \in \{1, \ldots, p\}$, and 'Soft' stands for the soft-thresholding operator, i.e.

$$\text{Soft}(a, \lambda) = \begin{cases} a - \lambda & \text{if } a > \lambda \\ 0 & \text{if } -\lambda \leqslant a \leqslant \lambda \\ a + \lambda & \text{if } a < -\lambda. \end{cases} \tag{2.39}$$

The computational complexity of this scheme can be shown to be $\min(\mathcal{O}(np^2), \mathcal{O}(p^3))$, which is the same as for the methods classified under SPACE; the complexity of glasso is $\mathcal{O}(p^3)$. We do not prove this fact, as it follows directly from reasoning elaborated on in Khare et al. [2015].

2.3.4 Tuning

A question that remains is how λ in (2.28) should be chosen for model selection purposes. Fortunately, a number of methods have been proposed for selecting the regularization pa-

rameter λ in ℓ_1 penalization methods and can be applied in our context. On the one hand, a predictive assessment as in cross-validation can be considered, but the selected graphs are typically too dense. Other possibilities include generalized cross validation (GCV) [Tibshirani, 1996], Akaike's Information Criterion (AIC), approaches based on stability under resampling [Meinshausen and Bühlmann, 2010, Shah and Samworth, 2013, Liu et al., 2010], the Bayesian Information Criterion (BIC) [Schwarz, 1978] as well as extensions of BIC proposed to cope with large model spaces [Chen and Chen, 2008, Gao et al., 2012, Foygel and Drton, 2010a, Barber and Drton, 2015]. The latter come with some consistency guarantees.

As a demonstration, we may consider the BIC criterion based on the basic score matching loss (2.2), i.e.

$$\mathrm{BIC}(\lambda) = (\hat{\theta}^\lambda)^T \mathbf{\Gamma}(\mathbf{x}) \hat{\theta}^\lambda - 2\gamma(\mathbf{x})^T \hat{\theta}^\lambda + |\hat{E}^\lambda| \log n, \tag{2.40}$$

where $\hat{E}^\lambda = \{(j,k) : \hat{\mathbf{\Theta}}^\lambda_{jk} \neq 0, j < k\}$ and we have momentarily used the λ superscript to indicate dependence of estimated quantities on λ. Alternatively, we could refit, that is, replace $\mathbf{\Theta}^\lambda$ by an unregularized SME computed in the submodel given by constraining all θ_{jk} with $(j,k) \notin \hat{E}^\lambda$ to be zero in (2.40). In either case, we choose λ to minimize (2.40).

2.4 Results on model selection consistency

This section establishes high-dimensional model selection consistency (sparsistency) of regularized score matching. While the results in this section are clearly generalizable, we consider the continuous pairwise interaction model as given by (2.18) with symmetric $p \times p$ interaction matrix $\mathbf{\Theta} = (\theta_{jk})$. We let $\theta = \mathrm{vec}(\mathbf{\Theta})$. Then the regularized score matching estimator, in its basic or non-negative version, is

$$\hat{\theta} = \arg\min_\theta \frac{1}{2} \theta^T \mathbf{\Gamma}(\mathbf{x}) \theta + \gamma(\mathbf{x})^T \theta + c(\mathbf{x}) + \lambda \|\theta\|_1. \tag{2.41}$$

By Lemma 2, $\mathbf{\Gamma}(\mathbf{x})$ is a symmetric $p^2 \times p^2$ matrix that is block-diagonal, with blocks of size $p \times p$. For notational convenience, we drop the explicit reference to the data matrix $\mathbf{x}$ and denote $\mathbf{\Gamma}(\mathbf{x})$ and $\gamma(\mathbf{x})$ as $\mathbf{\Gamma}$ and γ. In addition, the true data-generating distribution

is assumed to belong to the considered model. We denote the true interaction matrix by $\mathbf{\Theta}^* = (\theta^*_{jk})$ and its vectorization by θ^*. We define $\mathbf{\Gamma}^*$ and γ^* to be the expected values of $\mathbf{\Gamma}$ and γ. The support of θ^*, that is,

$$S \equiv S(\theta^*) = \{(j,k) : \theta^*_{jk} \neq 0\}$$

is the edge set of the true conditional independence graph. Similarly,

$$\hat{S} \equiv S(\hat{\theta}) = \{(j,k) : \hat{\theta}_{jk} \neq 0\}$$

determines the graph inferred by regularized score matching. Finally, we write d for the maximum degree of the p nodes of the conditional independence graph. In other words, d is the maximum number of nonzero off-diagonal entries in any row (or column) of $\mathbf{\Theta}^*$.

2.4.1 Irrepresentability

We say that the irrepresentability (or mutual incoherence) condition holds with incoherence parameter α if the following assumption holds.

Assumption 1. *There exists an* $\alpha \in (0,1]$ *such that*

$$\left|\left|\left|\mathbf{\Gamma}^*_{S^cS}(\mathbf{\Gamma}^*_{SS})^{-1}\right|\right|\right|_\infty \leqslant (1-\alpha). \tag{2.42}$$

Irrepresentability conditions play a key role in the analysis of ℓ_1 regularization techniques [Bühlmann and van de Geer, 2011]. For neighborhood selection in Gaussian graphical models, it has been formulated in terms of the covariance matrix $\mathbf{\Sigma}^*$ [Meinshausen and Bühlmann, 2006].I n the theoretical analysis of the glasso, the constraint is placed on the Hessian of the log-determinant of the precision matrix $\mathbf{K}^*$, i.e. $(\mathbf{K}^*)^{-1} \otimes (\mathbf{K}^*)^{-1}$ [Ravikumar et al., 2011].

In order to highlight the differences in conditions required for sparsistency of glasso, neighborhood selection, SPACE and regularized score matching, we revisit the Gaussian graphical model example in Meinshausen [2008]. Let $\rho \in (0, 1/\sqrt{2})$, and let $\mathbf{\Sigma} = \left(\sigma_{ij}\right)$ be the 4×4 covariance matrix with ones along the diagonal, $\sigma_{23} = \sigma_{32} = 0$, $\sigma_{14} = \sigma_{41} = 2\rho^2$

and all other off-diagonal entries equal to ρ. The precision matrix $\mathbf{K} = (\mathbf{\Sigma})^{-1}$ then has $\kappa_{14} = \kappa_{41} = 0$. The conditional independence graph G is as in Figure 2.1a.

Meinshausen showed that for samples drawn from $N(0, \mathbf{\Sigma})$, glasso can consistently recover G only if $\rho \leqslant \sqrt{3/2} - 1 \approx 0.23$. For neighborhood selection, the corresponding necessary condition is $\rho \leqslant 0.5$. If these conditions fail, then for large sample size, the probability of erroneously including the edge $(1, 4)$, i.e. $P(\hat{\kappa}_{14} \neq 0)$ can be shown to be at least 0.5. It turns out that for regularized score matching, the analogous necessary condition gives a bound that falls in between 0.23 and 0.5, specifically, $\rho \leqslant \sqrt{2} - 1 \approx 0.41$.

We observe that glasso, which yields positive definite estimates, requires the most stringent condition. When working with symmetric matrices as in regularized score matching, the condition is markedly relaxed. Allowing non-symmetric matrices in neighborhood selection leads to further relaxation of the condition. Interestingly, the pseudo-likelihood methods classified under SPACE have the same necessary condition as score matching.

Assumption 1 should be seen as sufficient for consistency of regularized score matching. For Meinshausen's example, it can be shown to amount to $\rho < \frac{1}{2}(\sqrt{3} - 1) \approx 0.37$. The analogous sufficient condition for glasso from Ravikumar et al. [2011] requires that $\rho < \frac{1}{2}(\sqrt{2} - 1) \approx 0.21$. For neighborhood selection, the condition is $\rho < 0.5$.

2.4.2 Main Results

We define

$$c_{\mathbf{\Gamma}^*} = \left|\left|\left|(\mathbf{\Gamma}^*_{SS})^{-1}\right|\right|\right|_\infty, \text{ and } c_{\mathbf{\Theta}^*} = |||\mathbf{\Theta}^*|||_\infty. \tag{2.43}$$

Moreover, let

$$\mathbf{R}_1 = (\mathbf{\Gamma} - \mathbf{\Gamma}^*), \qquad r_2 = \gamma^* - \gamma, \qquad r_3 = \mathbf{\Gamma}^*\theta^* - \gamma^*, \tag{2.44}$$

such that the KKT conditions from (2.34) can be written as

$$\mathbf{\Gamma}^*(\hat{\theta} - \theta^*) + R_1\hat{\theta} + r_2 + r_3 + \lambda\hat{z} = 0, \quad \hat{z} \in \partial\|\hat{\theta}\|_1. \tag{2.45}$$

Theorem 1. *Assume that $\mathbf{\Gamma}^*_{SS}$ is invertible and the irrepresentability condition holds with incoherence parameter $\alpha \in (0, 1]$ (Assumption 1). Furthermore, assume that*

$$\|\mathbf{R}_1\|_\infty < \epsilon_1, \qquad \|r_2\|_\infty < \epsilon_2, \tag{2.46}$$

with $d\epsilon_1 \leqslant \alpha/6c_{\mathbf{\Gamma}^*}$. If

$$\lambda > \frac{3(2-\alpha)}{\alpha} \max\{c_{\mathbf{\Theta}^*}\epsilon_1, \epsilon_2\}, \tag{2.47}$$

then the following statements hold:

(a) The rSME $\hat{\theta}$ is unique, has its support included in the true support ($\hat{S} \subseteq S$), and satisfies

$$\|\hat{\theta} - \theta^*\|_\infty < \frac{c_{\mathbf{\Gamma}^*}}{2-\alpha}\,\lambda.$$

(b) If

$$\min_{1 \leqslant j < k \leqslant p} |\theta^*_{jk}| > \frac{c_{\mathbf{\Gamma}^*}}{2-\alpha}\,\lambda,$$

then $\hat{S} = S$ and $\operatorname{sign}(\hat{\theta}_{jk}) = \operatorname{sign}(\theta^*_{jk})$ for all $(j, k) \in S$.

Theorem 1 imposes deterministic conditions on the data, namely, the bounds in (2.46). In the following corollaries, we will consider specific distributional assumptions and impose population conditions that imply bounds of the form (2.46) with high probability.

First, we provide a result for regularized score matching for the Gaussian case (Example 1), which has $\mathbf{\Gamma} = \mathbf{I}_{p\times p} \otimes \mathbf{W}$ with $\mathbf{W}$ being the sample covariance matrix, and $\gamma = \operatorname{vec}(\mathbf{I}_{p\times p})$. When the data is generated from a normal distribution with covariance matrix $\mathbf{\Sigma}^*$ then $\mathbf{\Gamma} = \mathbf{I}_{p\times p} \otimes \mathbf{\Sigma}^*$ and, of course, $\gamma^* = \gamma = \operatorname{vec}(\mathbf{I}_{p\times p})$.

Corollary 1. *Suppose the data is generated from a normal distribution $N(0, \mathbf{\Sigma}^*)$ such that $\mathbf{\Gamma}^*_{SS}$ is invertible and irrepresentability holds for $\alpha \in (0, 1]$. Let $\mathbf{K}^* = (\kappa^*_{jk}) = (\mathbf{\Sigma}^*)^{-1}$,*

$$c^* = 3200 \max_{j=1,\ldots,p} \sigma^{*2}_{jj} \quad \textit{and} \quad c_1 = \frac{4}{\alpha} c_{\mathbf{\Gamma}^*}.$$

Take any $\tau_1 > 2$. If the sample size satisfies

$$n > c^* c_1^2 d^2 (\log p^\tau + \log 4), \tag{2.48}$$

and the regularization parameter is

$$\lambda > \frac{2c_{\mathbf{K}^*}(2-\alpha)}{\alpha}\sqrt{\frac{c^*(\log p^\tau + \log 4)}{n}}, \tag{2.49}$$

then the following statements hold with probability $1 - 1/p^{\tau-2}$*:*

(a) The rSME $\hat{\mathbf{K}}$ *from (2.30) is unique, has its support included in the true support (*$\hat{S} \subseteq S$*), and satisfies*

$$\|\hat{\mathbf{K}} - \mathbf{K}^*\|_\infty < \frac{c_{\mathbf{\Gamma}^*}}{2-\alpha}\,\lambda.$$

(b) If

$$\min_{1 \leqslant j < k \leqslant p} |\kappa^*_{jk}| > \frac{c_{\mathbf{\Gamma}^*}}{2-\alpha}\,\lambda,$$

then $\hat{S} = S$ *and* $\operatorname{sign}(\hat{\mathbf{K}}_{jk}) = \operatorname{sign}(\kappa^*_{jk})$ *for all* $(j,k) \in S$*.*

The corollary is proven in Appendix A.2.2. Numerical experiments reported in Section 2.5 suggest that the sample size n indeed needs to satisfy $n \gtrsim d^2 \log p$ for sparsistency.

Corollary 2. *Suppose the data is generated from a non-negative Gaussian distribution with parameter* $\mathbf{K}^*$*, i.e.* $N(0, (\mathbf{K}^*)^{-1})$ *is truncated to* $\mathbb{R}^p_+$*. Suppose further that* $\mathbf{\Gamma}^*_{SS}$ *is invertible and irrepresentability holds for* $\alpha \in (0,1]$*. Define* $v_1 = \max_{j,k,l} Var[X_j^2 X_k X_l]$ *and* $v_2 = \max_{j,k} Var[X_j X_k]$*, and let*

$$c^{**} = \max\left\{\left(\frac{L}{2}\right)^4 \sqrt{v_1}, \left(\frac{L}{2}\right)^2 \sqrt{v_2}\right\} \quad \textit{and} \quad c_2 = \frac{6}{\alpha} c_{\mathbf{\Gamma}^*}$$

where $L > 0$ *is an absolute constant. Take any* $\tau > 3$*. If the sample size satisfies*

$$n > c^{**} c_2^2 d^2 (\log p^\tau + \log 2)^8, \tag{2.50}$$

and the regularization parameter is

$$\lambda > \frac{3(2-\alpha)}{\alpha} \max\{c_{\mathbf{K}^*}, 1\} \sqrt{\frac{c^{**}(\log p^\tau + \log 2)^8}{n}}, \tag{2.51}$$

then the following statements hold with probability $1 - \frac{1}{p^{\tau-3}}$*:*

(a) The rSME $\hat{\mathbf{K}}_+$ *based on penalizing (2.26) with* $\lambda\|\mathbf{K}\|_{1,off}$ *is unique, has its support included in the true support (*$\hat{S} \subseteq S$*), and satisfies*

$$\|\hat{\mathbf{K}}_+ - \mathbf{K}^*\|_\infty < \frac{c_{\mathbf{\Gamma}^*}}{2-\alpha}\,\lambda.$$

(b) If

$$\min_{1 \leqslant j < k \leqslant p} |\kappa^*_{jk}| > \frac{c_{\mathbf{\Gamma}^*}}{2-\alpha}\,\lambda,$$

then $\hat{S} = S$ *and* $\operatorname{sign}(\hat{\mathbf{K}}_{+,jk}) = \operatorname{sign}(\kappa^*_{jk})$ *for all* $(j,k) \in S$.

The proof of the corollary, which is given in Section A.2.3, uses general tail bounds that apply to log-concave measures. The lower bound for n given in (2.50) could well be suboptimal and a lower power of $\log p$ may be sufficient for sparsistency. However, the experiments in Section 2.5 suggest that the exponent for $\log p$ cannot be taken too much smaller than 8.

We also compared the lower bound we obtained for the non-negative Gaussian case to a result implied by the work of Yang et al. [2013] who treat consistency of neighborhood selection in a general framework that allows node-wise conditional distributions to arise from exponential families. Interestingly, when working out what their general theorem would say about the above non-negative Gaussian model we found that the sample size n would also be required to be at least $d^2(\log p)^8$. Our result from Corollary 2 is thus at least comparable to existing results in the literature.

2.5 Numerical experiments

2.5.1 Empirical verification of theoretical results

We perform experiments to provide empirical support for Corollary 1. This corollary treats Gaussian graphical models for which the sample size n ought to be of order $d^2 \log p$. We experiment by varying the number of variables p and the degree d. In addition, we investigate how the sample size n required for sparsistency for non-negative Gaussian graphical models

needs to depend on p, and how this scaling differs from that obtained in Corollary 2. All reported results are based on averaging over 100 trials.

Dependence on number of nodes p in the Gaussian setting

Consider first the case where the underlying conditional independence graph is a chain of length $p \in \{64, 100, 225, 375\}$. The degree d is always 2, and we choose the tridiagonal precision matrix $\mathbf{K}^*$ to have entries $\kappa^*_{jk} = 0.3$ if $(j,k) \in E$ and $\kappa^*_{jj} = 1$ for $j = 1, \ldots p$. Here, α, $c_{\mathbf{K}^*}$ and $c_{\mathbf{\Gamma}^*}$ are constant across all p. We let the regularization parameter λ scale with $\sqrt{\log p/n}$, a choice corroborated by Corollary 1.

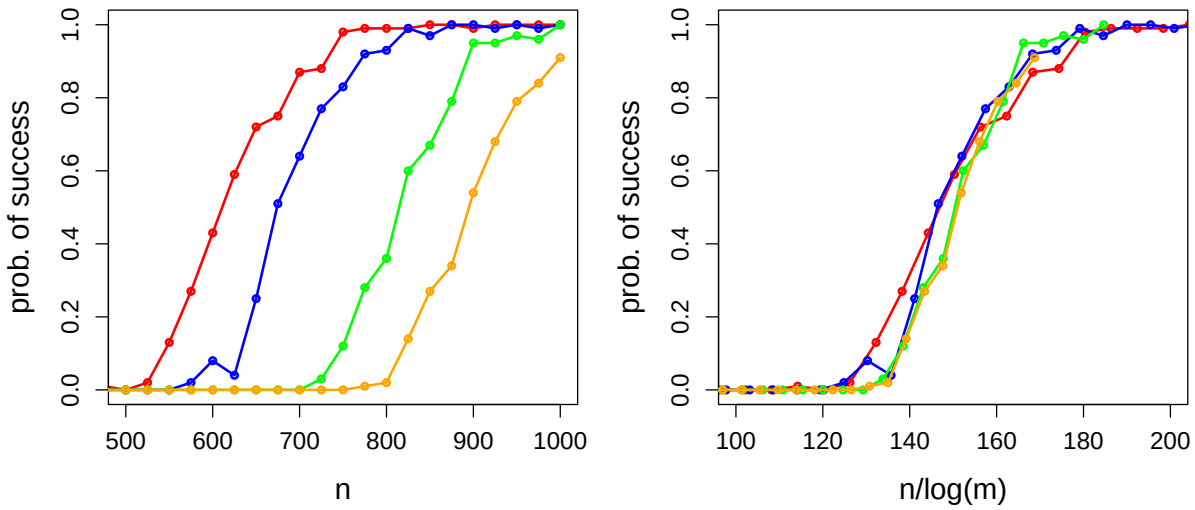
Figure 2.2 shows the probability of correct signed support recovery plotted against the sample size n, with different curves corresponding to different p. As expected, we see from Figure 2.2a that successful support recovery requires n to grow with p. However, upon rescaling n by $1/\log(p)$, the curves overlap as seen in Figure 2.2b.

We conclude that with C and d held constant, the sample size n needs to scale with $\log p$ for consistent signed support recovery. This is consistent with Corollary 1.

Dependence on the node degree d in the Gaussian setting

We now fix the number of nodes to $p = 200$ and vary d. We consider a star graphs with varying hub node degree $d \in \{15, 20, 25\}$. The precision matrix $\mathbf{K}^*$ is chosen such that $\sigma^*_{jk} = 2.5/d$ for $(j,k) \in E$, and $\sigma^*_{jj} = 1$ for $j = 1, \ldots, p$. Now, α, $c_{\mathbf{K}^*}$ and $c_{\mathbf{\Gamma}^*}$ are constant across all d.

Figure 2.3 shows the probability of correct signed support recovery plotted against n. The left panel demonstrates that correct recovery is more difficult with increasing d. Larger n is needed to attain the same success rate. Upon rescaling n by $1/d^2$ in the right panel, the three curves align. This validates Corollary 1 in that for fixed p, α, $c_{\mathbf{K}^*}$ and $c_{\mathbf{\Gamma}^*}$, the sample size n needs to scale with d^2 to ensure sign consistency.



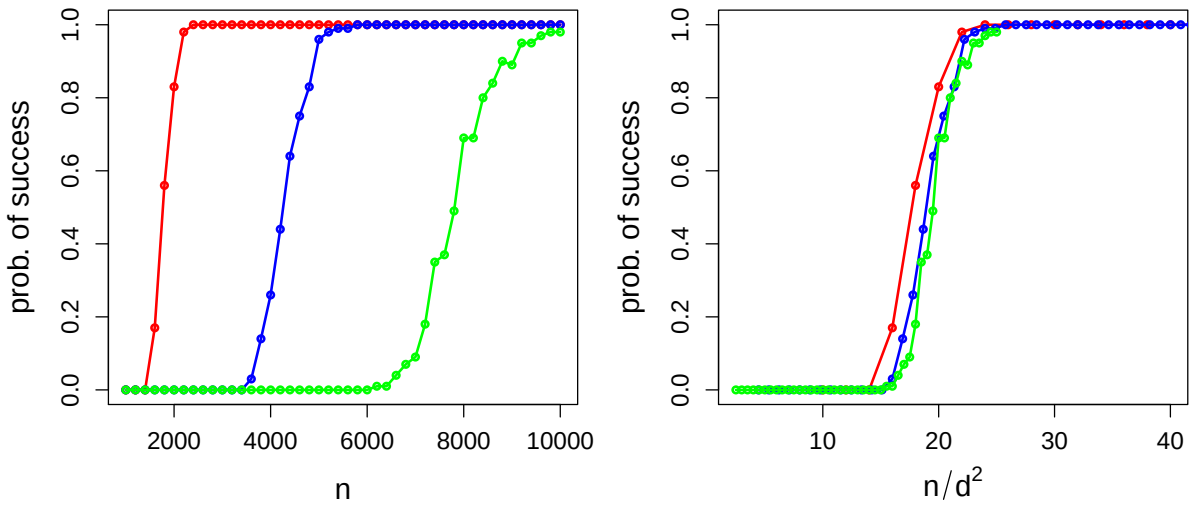
(a) (b)

Figure 2.2: Relative frequencies of signed support recovery for Gaussian observations with a conditional independence graph that is a chain of varying length p. Panels (a) and (b) differ only in the scaling of the x-axis. The colored lines correspond to $m = 64$ (—), $m = 100$ (—), $m = 225$ (—) and $m = 375$ (—).

Non-negative Gaussian experiments

Finally, we experiment with regularized non-negative score matching for normal observations truncated to the positive orthant. According to Corollary 2, a sample size of $n \gtrsim d^2(\log p)^8$ is sufficient for signed support recovery. The aim of our experiments is to explore to what extent this scaling is necessary. Specifically, we will consider exponents other than 8 for $\log p$.

For our experiments, we revisit the chain-structured graphs, and choose a triangular matrix $\mathbf{K}^*$ with $\kappa^*_{jk} = 0.3$ if $(j,k) \in E$ and and $\kappa^*_{jj} = 1$ for $j = 1, \ldots p$. The degree d is fixed at 2 and we only vary $p \in \{20, 25, 30\}$. We let the regularization parameter λ to scale with $\sqrt{(\log p)^8/n}$. Figure 2.4 plots the probability of correct signed support recovery against n,



(a) (b)

Figure 2.3: Relative frequencies of signed support recovery for Gaussian observations whose conditional independence graph is a star with varying degree d. Panels (a) and (b) differ only in the scaling of the x-axis. The colored lines correspond to $d = 10$ (—), $d = 15$ (—), and $d = 20$ (—).

with different curves for the different values of p.

Panel (a) in Figure 2.4 illustrates that, larger n is needed account for larger p. The other three panels have the x-axis rescaled to $n/(\log p)^a$ for exponents $a \in \{6, 7, 8\}$. Panel (b) suggests that n scaling with $(\log p)^6$ is not sufficient for support recovery. Comparing panels (c) and (d), $(\log p)^8$ seems more than what is necessary. It thus appears that the scaling of the sample size we assumed in Corollary 2 is suboptimal but not drastically so.

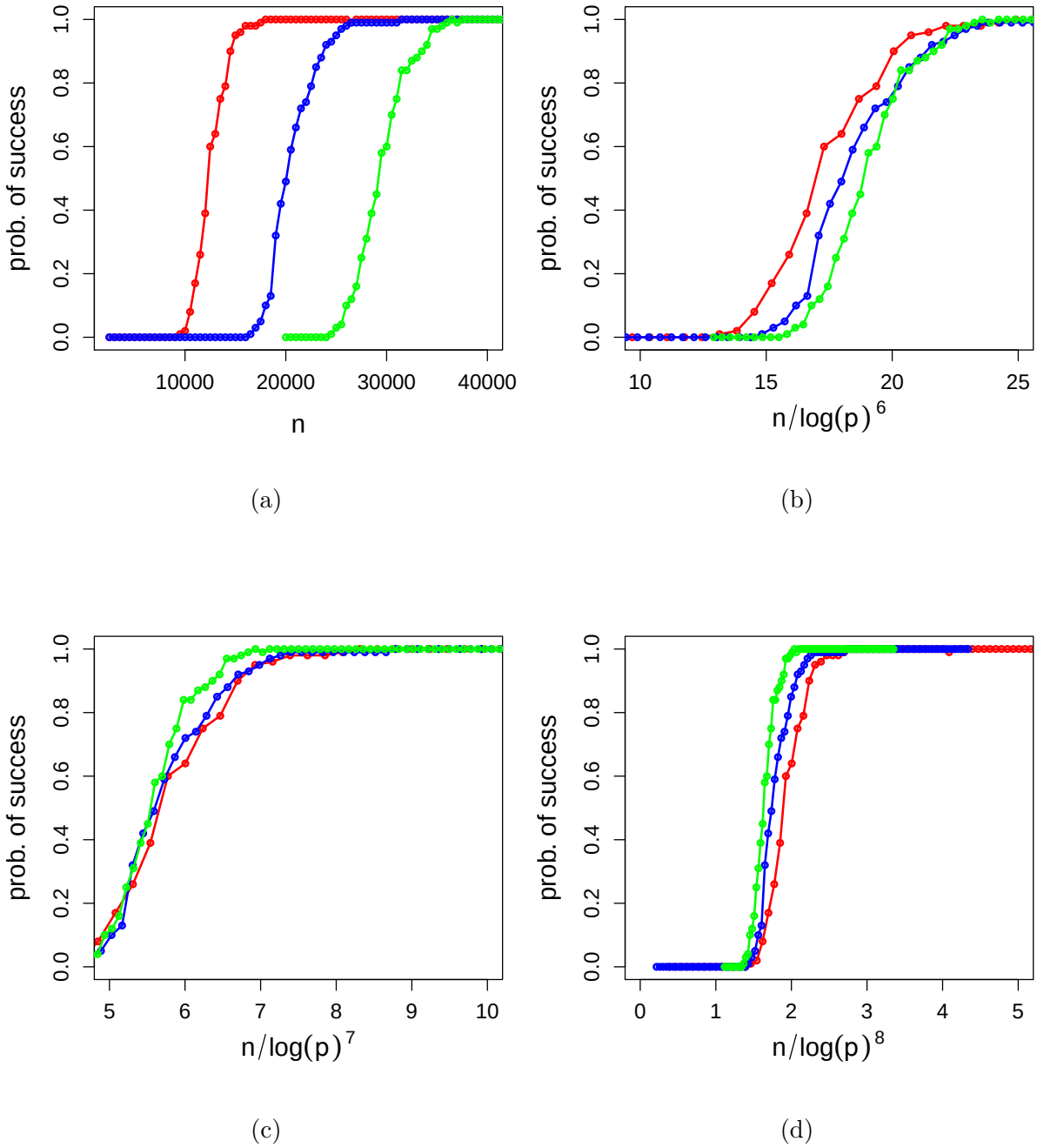


Figure 2.4: Relative frequencies of signed support recovery for truncated Gaussian observations whose conditional independence graph is a chain of varying length p. The four panels differ only in the scaling of the x-axis. The colored lines correspond to $p = 20$ (——), $p = 25$ (——), and $p = 30$ (——).

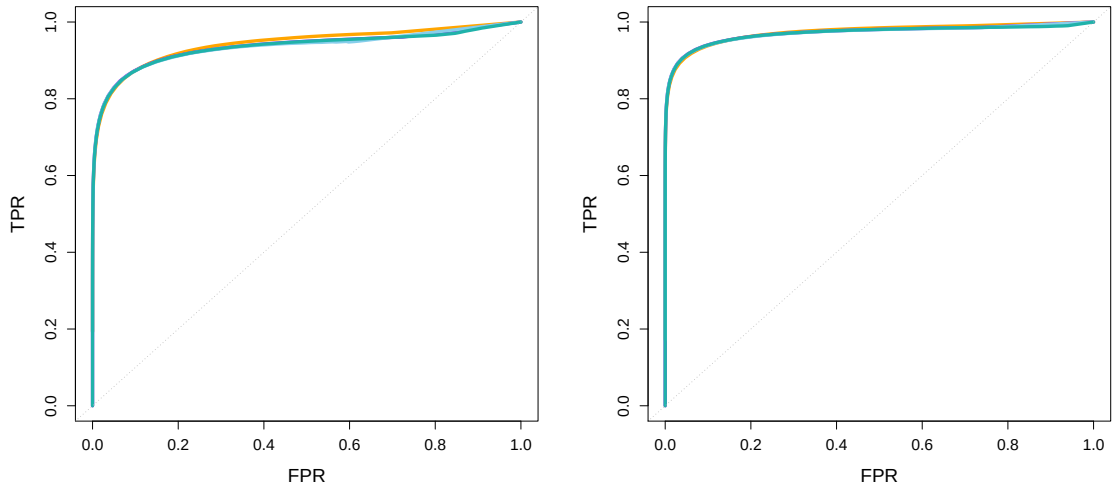
2.5.2 Comparisons to other methods

We perform numerical experiments comparing regularized score matching to existing methods when data is simulated from (i) a multivariate normal distribution, (ii) a multivariate truncated normal distribution, and (iii) a distribution with normal conditionals. The comparison is made against three methods for estimation of Gaussian graphical models, namely, glasso, neighborhood selection (both implemented in the `R` packages `huge`) and SPACE (in its CONCORD formulation, with `R` package `gconcord`). In addition, we consider the *nonparanormal SKEPTIC*, which applies glasso to a matrix of rank correlations (Kendall's τ or Spearman's ρ) and can be motivated by a Gaussian copula model [Liu et al., 2012b]. We utilize the version based on Kendall's τ. Finally, we compare to *SPACEJAM* [Voorman et al., 2014], which is based on additive modeling of conditional means and implemented in the `R` package `spacejam`. We conclude this section with brief investigations on the robustness of regularized score matching when data is not generated under the assumed model. All results in this section are based on averaging over 100 independently generated datasets.

Gaussian data

We consider a graph with $p = 1000$ nodes, composed of 10 connected components, each 100 nodes in size and structured as a 10×10 2-D lattice (4 nearest neighbors). Each connected component also features three hubs with node degree 20, randomly selected from the subset of nodes in the component.

We follow a procedure similar to the one from Peng et al. [2009] to convert the adjacency matrix of the graph into a sparse diagonally dominant partial correlation matrix. For each non-zero element of the adjacency matrix, we sample a draw from a uniform distribution on $[0.5, 1]$. Each row of this new matrix is then rescaled by 1.5 times the sum of the absolute values of the off-diagonal entries in the row. We average this matrix with its transpose to ensure symmetry, and set its diagonal elements to 1. This matrix is inverted and converted into a correlation matrix to form $\mathbf{\Sigma}^*$.



(a) $n = 600$ (b) $n = 1000$

Figure 2.5: ROC curves for the Gaussian case. The dashed grey line represents random selection of edges. The color to method correspondence is as follows: regularized score matching (——), neighborhood selection (——), glasso (——), and SPACE (——). The curves are almost perfectly aligned.

Data is then generated from a multivariate normal distribution with mean zero and a covariance matrix $\mathbf{\Sigma}^*$. We choose sample size $n = 600$ and 1000. The setup agrees with that in Peng et al. [2009], except that the number of nodes has been scaled up.

Figure 2.5 shows the ROC curves obtained under both sample sizes. Since the truth is Gaussian, we do not report results for SKEPTIC or SPACEJAM. For both sample sizes, the curve for regularized score matching almost perfectly aligns with those for neighborhood selection, SPACE, and glasso. The results indicate that regularized score matching estimators achieves state-of-the-art statistical efficiency in Gaussian models.

Glasso, SPACE, neighborhood selection and SKEPTIC all presume some form of underlying Gaussianity. In the next set of experiments, we demonstrate the application of

regularized score matching in scenarios where these assumptions do not hold to highlight the versatility of the proposed approach.

Non-negative Gaussian data

Similar to the Gaussian setting, we consider a graph with $p = 100$ nodes, composed of 10 disconnected subgraphs with equal number of nodes. Using the lower triangular elements adjacency matrix of each 10 node subgraph, we construct ten matrices, where in each matrix, the element is drawn independently to be 0 with probability 0.2, and from a uniform distribution on $[0.5, 1]$ with probability 0.8. The matrices, after symmetrization, are combined into a 100×100 block matrix. The diagonal elements are set to a common positive number such that the minimum eigenvalue is 0.1 to form the precision matrix of the pre-truncated normal, $\mathbf{K}^*$.

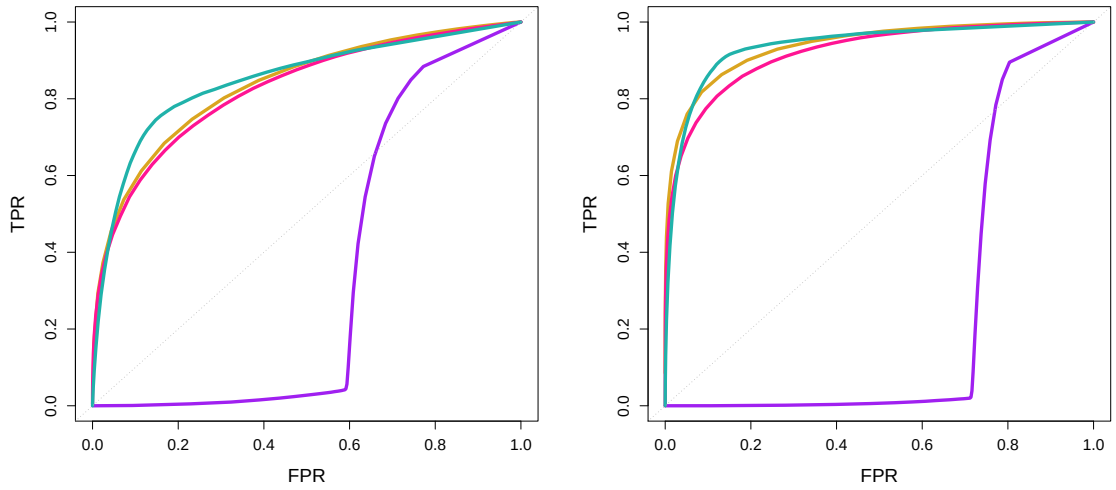
Data was then generated from a truncated centered multivariate normal, left-truncated at 0 and with $\mathbf{\Sigma}^* = (\mathbf{K}^*)^{-1}$ as normal covariance. We used the Gibbs sampler from the `tmvtnorm` package in `R` with a burnin period of 100 samples. We thinned out the remaining samples, keeping one in ten. The sample size n is taken to be either 2500 or 5000. The need for a larger sample size is explained by our theoretical findings in Section 2.4, specifically Corollary 2.

The ROC curves are shown in Figure 2.6, where regularized score matching outperforms all competitors considered. The closest competitor to regularized score matching are SKEPTIC and SPACEJAM, both of which, objectively, perform well, being capable of capturing some of the non-Gaussianity in the data.

Normal conditionals

Next, we take the data-generating distribution to have a density from the class

$$f_{\mathbf{B},\mathbf{b},\mathbf{b}^{(2)}}(x) \propto \exp\left\{\sum_{j\neq k}\beta_{jk}x_j^2x_k^2 + \sum_{j=1}^{p}\beta_j^{(2)}x_j^2 + \sum_{j=1}^{p}\beta_j x_j\right\}, \qquad x \in \mathbb{R}^p, \tag{2.52}$$



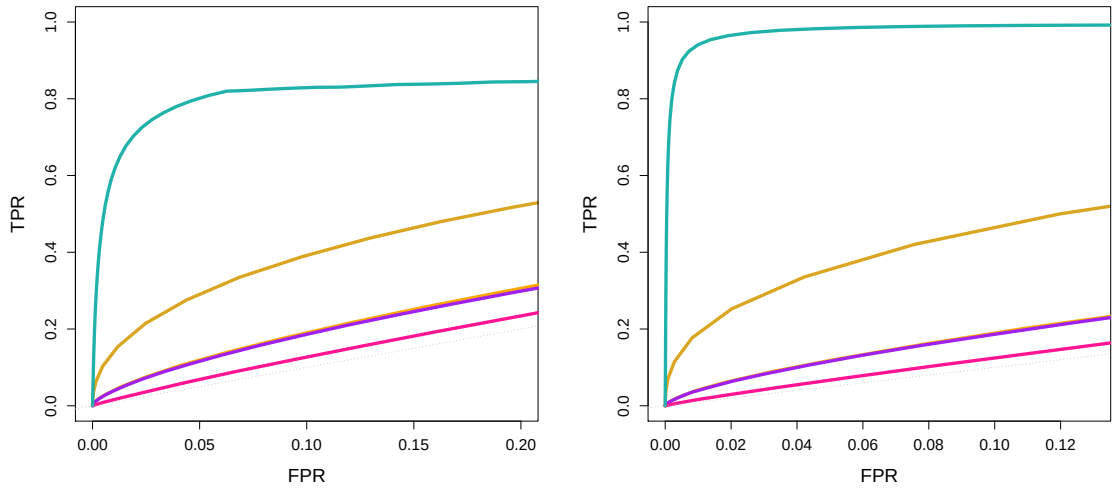
(a) $n = 2500$ (b) $n = 5000$

Figure 2.6: ROC curves for the non-negative Gaussian case. The dashed line represents random selection of edges. The color to method correspondence is as follows: regularized score matching (——), glasso (——), SPACE (——), SKEPTIC (——), and SPACEJAM (——).

where $\mathbf{B} = \{\beta_{jk}\}$ is a symmetric matrix with diagonal entries 0. This family is a special case of the distributions with normal conditionals from Example 3.

We consider the case $p = 625$, with the graph being a 25×25 2-D lattice (4 nearest neighbors). The true interaction matrix $\mathbf{B}^*$ is constructed by multiplying the adjacency matrix by $-1/25$. The coefficients for the terms x_j^2 are all set equal to -1 and those for the x_j all equal to $8/50$, which makes the marginal distributions deviate noticeably from Gaussianity. Data can be generated by Gibbs sampling using the Gaussian full conditionals. We discard the first 100 samples and thin out the remaining samples, keeping one in ten, as in Section 2.5.2.

We plot the ROC curves for conditional normal data in Figure 2.7. Regularized score matching outperforms its competitors by a clear margin. This is not surprising, as both glasso



(a) $n = 450$ (b) $n = 1000$

Figure 2.7: ROC curves for the normal conditionals case. The dashed line represents random selection of edges. The color to method correspondence is as follows: regularized score matching (——), glasso (——), SPACE (——), SKEPTIC (——), and SPACEJAM (——). The curve for glasso overlaps with the curve for SPACE.

and SPACE are derived under normality. A Gaussian copula model as underlying SKEPTIC is of little help. SPACEJAM does best among the competitors but cannot fully extract the available signal about the edge structure as the conditional means are non-additive and the conditional variances are not constant.

2.5.3 A robustness check

It is of interest to study robustness of regularized score matching: in particular, we ask how efficiently regularized score matching recovers the CIG encoded in the 'original' data from the'distorted' data. We consider two scenarios. First, we apply the Gaussian score matching to a contaminated Gaussian setting similar to that explored in Finegold and Drton [2011].

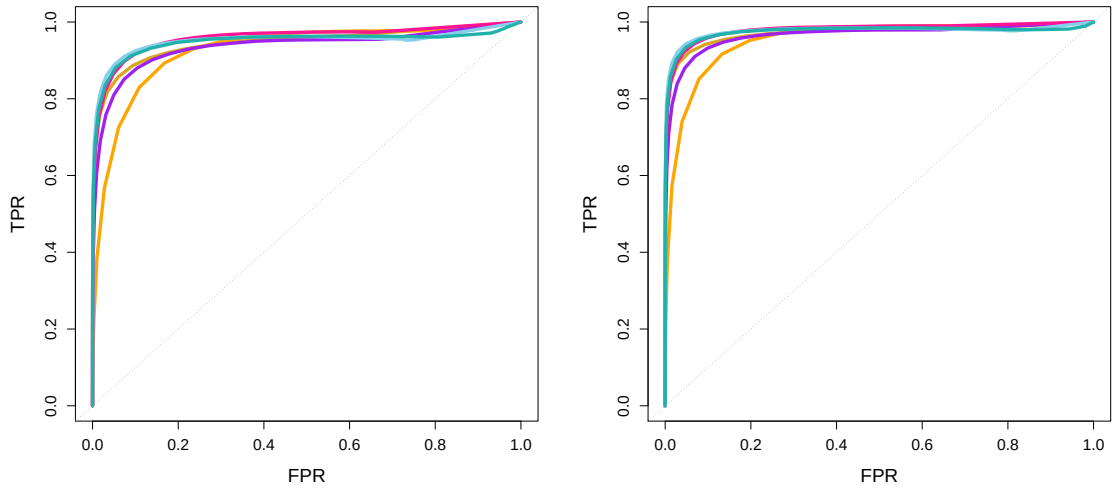
That is, a random subset of Gaussian observations is replaced with Gaussian noise. In the second example, we investigate the performance of the regularized Gaussian score matching when the observations are not Gaussian but rather drawn from a multivariate t-distribution.

Contaminated Gaussians

We mimic the setup used in the numerical experiments in Finegold and Drton [2011], who consider these settings to test the robustness of their *tlasso*. Fixing $p = 200$, we construct a sparse precision matrix $\mathbf{K}^*$ according to the following steps: (1) choose each (strictly) lower triangular element of $\mathbf{K}^*$ to be independently -1, 0, 1 with probability 0.01, 0.98 and 0.01 respectively, (2) symmetrize the matrix (3) for each row, i.e. for $j = 1, \ldots, p$, set $\kappa^*_{jj} = 1 + \|\kappa^*_{j,-j}\|_0$ where $\kappa^*_{j,-j}$ refers to the jth row of $\mathbf{K}^*$ with the diagonal element in that row removed. To strengthen partial correlations, the diagonal elements are scaled down by a common positive factor such that the minimum eigenvalue of the resulting matrix is approximately 0.6 (close to 0.62 in our setup). The covariance matrix $\mathbf{\Sigma}^*$ is obtained by inverting $\mathbf{K}^*$.

We generate either $n = 150$ or $n = 200$ observations from a multivariate normal distribution with mean zero and a covariance matrix $\mathbf{\Sigma}^*$. We then corrupt 2% of the observations, substituting them with i.i.d. $N(0, 0.2)$ draws. The corrupted observations cannot easily be differentiated from normal observations, and this elevates the difficulty of the estimation problem.

We present the ROC curves in Figure 2.8. Interestingly, score matching performs reasonably well, on par with SKEPTIC and neighborhood selection. For both sample sizes, the differences, which are subtle, are most apparent in the regime where the number of false positives detected is small: score matching falls slightly short of neighborhood selection, but it also appears to slightly outperform SKEPTIC. Surprisingly, there is a clear margin of difference between the performances of regularized score matching and SPACE, the former outperforming the latter, despite their noted structural similarities. Glasso, which utilizes the full Gaussian likelihood, performs the worst. Overall, we conclude that regularized score



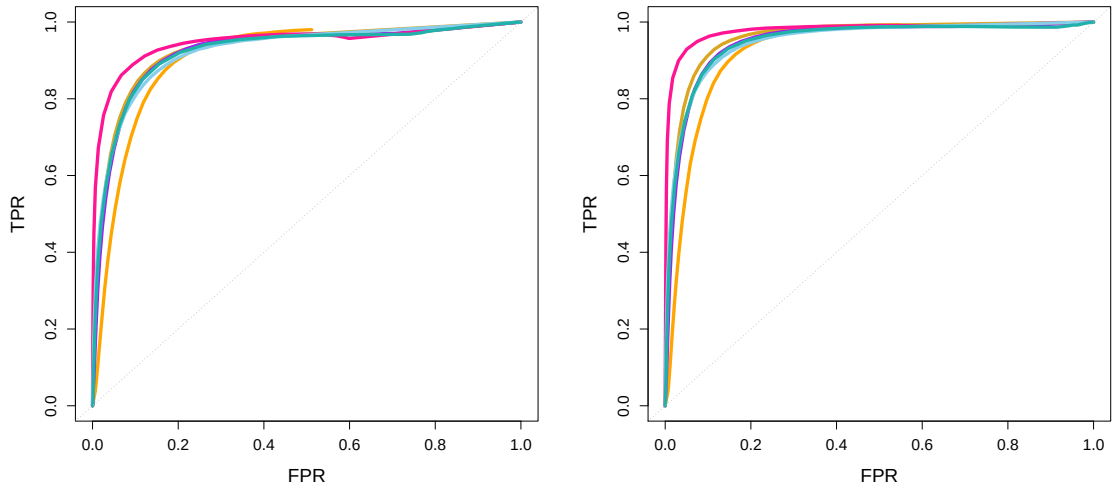
(a) $n = 150$ (b) $n = 200$

Figure 2.8: ROC curves for the contaminated Gaussian case. The dashed line represents random selection of edges. The color to method correspondence is as follows: regularized score matching (—), neighborhood selection (—), glasso (—), SPACE (—), SKEPTIC (—), and SPACE-JAM (—).

matching is competitively robust when compared to its alternatives in the contaminated Gaussian setting.

Multivariate t-distributed observations

In this section, we apply regularized Gaussian score matching to observations arising from a multivariate t-distribution with mean 0 and covariance matrix $\mathbf{\Sigma}^*$. This corresponds to testing the robustness of regularized score matching under model misspecification. Like in the previous section, we consider the case when $p = 200$. To set up $\mathbf{\Sigma}^*$, we construct a $p \times p$ adjacency matrix based on an Erdős-Rényi graph with the probability of drawing an edge between any two arbitrary nodes set to 0.01. We then convert the adjacency matrix into



(a) $n = 100$ (b) $n = 150$

Figure 2.9: ROC curves for the t-distributed case. The dashed line represents random selection of edges. The color to method correspondence is as follows: regularized score matching (—), neighborhood selection (—), glasso (—), SPACE (—), SKEPTIC (—), and SPACEJAM (—).

$\mathbf{\Sigma}^*$ using the same procedure as in Section 2.5.2. Samples were drawn from a multivariate t-distribution with covariance matrix $\mathbf{\Sigma}^*$ and three degrees of freedom.

The ROC curves are plotted in Figure 2.9 for $n = 100$ and $n = 150$. As expected, SKEPTIC outperforms all others, owing to its flexibility to accommodate outliers, as previously demonstrated in Liu et al. [2012b]. In fact, for elliptical distributions, such as the multivariate t-distribution, Kendall's τ allows for consistent estimation of $\mathbf{\Sigma}^*$, so SKEPTIC should perform optimally [Liu et al., 2012c]. Nonetheless, regularized score matching is reasonably robust under this setting: its performance is comparable to that of SPACEJAM – only falling slightly short – SPACE, and neighborhood selection. Again, glasso yields the poorest results.

2.6 Application to RNAseq data

The American Cancer Society estimates that in 2015 there will be 220,800 new cases of prostate cancer and 27,540 deaths. To understand how the cancer develops, as well as how it may be treated, it is necessary to decipher the genetic machinery which drives it. Since cancer is such a complex disease, it is insufficient to study a single gene at a time, as genes may interact with one another in many ways. Graphical modeling of gene expression data has the potential to aid in discovery of such interactions.

RNAseq data from next-generation sequencing technology can be used to identify genes that are activated/transcribed or suppressed at the time of measurement. However, RNAseq data are non-negative and have skewed marginals, which presents a challenge for existing methodologies. Graphical models based on truncated Gaussian models are interesting alternatives to existing approaches that primarily consist of applying Gaussian methods after transformations. Whether truncation models are truly useful scientifically deserves a fuller exploration; here we simply illustrate how different estimates can be obtained from the proposed methodology.

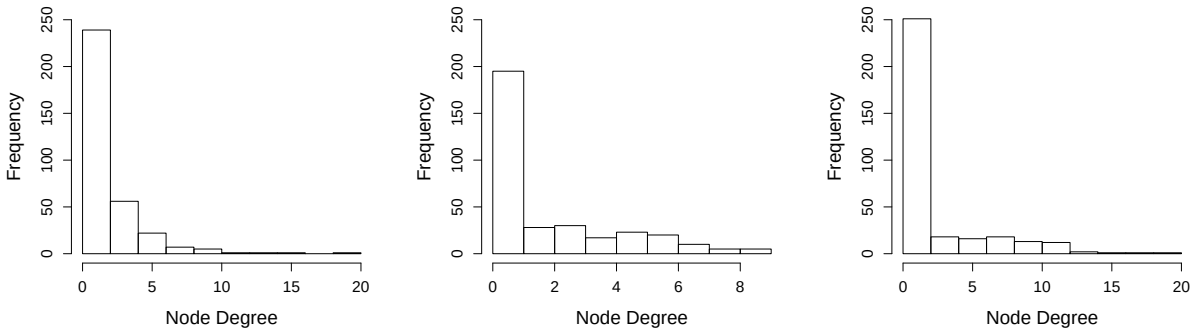
Our case study is based on the RNAseq data from 487 prostate adenocarcinoma samples available in The Cancer Genome Atlas dataset. We focus on 350 genes that belong to 'known' cancer pathways in the Kyoto Encyclopedia of Genes and Genomes. Removing genes with more than 10% missing values, we obtained a dataset with $p = 333$ genes. Remaining missing values were simply set to zero, adding to the challenge. We will properly treat missingness in Chapter 3. We consider an exponential family of truncated normal distributions with density

$$f_{\mu,\mathbf{K}}(x) \propto \exp\left\{\frac{1}{2}(x-\mu)^T\mathbf{K}(x-\mu)\right\}, \quad x \in \mathbb{R}^p_+.$$

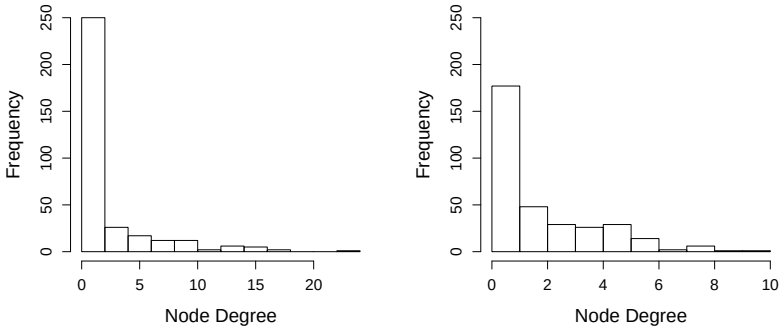
This generalizes the family of distributions considered in Example 2 by allowing the truncated normal distribution to have nonzero mean.

We compare regularized non-negative score matching, SPACE (using CONCORD formulation), glasso, SKEPTIC and SPACEJAM. We apply SPACE and glasso directly to the

standardized data. We do not consider any marginal transformations as they are naturally accounted for when comparing to the rank correlation-based SKEPTIC. For each method, we tune the regularization parameter λ in order to obtain $|E| = 333$ (or 334) edges. We show the estimated networks in fixed layouts in Figure 2.11. Node degree distributions are plotted in Figure 2.10.



(a) Reg. score matching (b) SPACE (c) Glasso



(d) SKEPTIC (e) SPACEJAM

Figure 2.10: Node degree distributions for inferred networks of $|E| = 333$ or 334 edges for all considered methods.

By visual inspection, glasso and SKEPTIC give similar topologies, which can be explained by the fact that both are derived from the full Gaussian likelihood. Interestingly, we observe that SPACEJAM and SPACE likewise yield similar graphs, which reinforces findings from Shojaie and Sedaghat [2016]. Regularized non-negative score matching yields a graph that is fairly different from the rest.

While the usefulness of these models remains to be further explored, our case study demonstrates that regularized score matching can provide estimates that differ in interesting ways to the estimates generated by other methods. We compile a list of the top ten most highly connected genes in each of the estimated graphs in Table 2.1 (some lists have more than ten genes due to ties), as there is strong evidence that highly connected nodes play important roles in biological networks [Carter et al., 2004b, Jeong et al., 2001, Han et al., 2004]. There are slight overlaps between the lists. Upon further inspection, we observe that six of the ten genes listed under regularized score matching have been previously linked to prostate cancer, five of which have not been identified by the competing methods:

- *CCNE2* (cyclin E2): a protein which is required for transition of the a_1 to S phase of the cell cycle, which determines cell division. Regulated by PTEN, a tumor suppressor, it is over-expressed in metastatic prostate tumor cells [Wu et al., 2009].

- *BRCA2* (breast cancer 2): mutations in the BRCA2 gene have been associated with early-onset prostate cancer in men; men carrying mutations have a predisposition to more aggressive phenotypes [Gayther et al., 2000, Mitra et al., 2008, Tryggvadóttir et al., 2007, Fan et al., 2006].

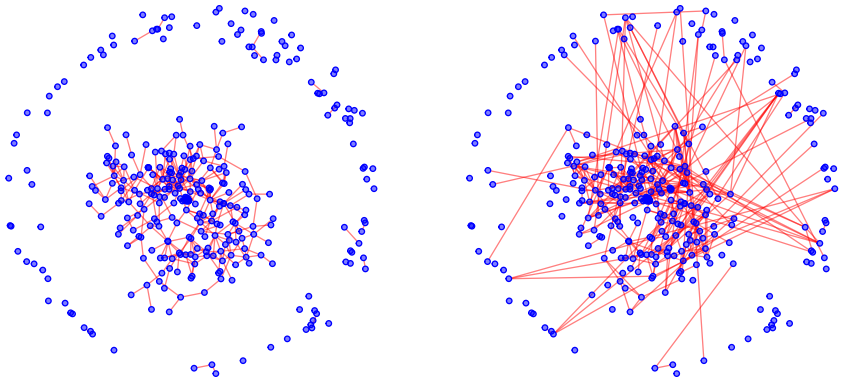
- *BIRC5* (survivin): a protein which prevents cell death, or apoptosis, and regulates cell division. Heightened expression has been found to be associated with higher final Gleason score, i.e. more aggressive cancer and worse prognosis [Kishi et al., 2004, Shariat et al., 2004].

- *SKP2* (S-phase kinase-associated protein 2, E3 ubiquitin protein ligase): a positive regulator of the a_1 to S phase of the cell cycle, which determines cell division. SKP2 labelling frequency in cancer was positively correlated with the Gleason score, and shown to be a significant predictor of reduced recurrence-free survival time after radical prostatectomy [Yang et al., 2002, Wang et al., 2008]. It has been proposed elsewhere as a promising therapeutic target for prostate cancer [Wang et al., 2012].

- *STAT5B* (signal transducer and activator of transcription 5B): a transcription factor that encourages metastatic behavior of human prostate cancer cells. Its inhibition has been shown to induce apoptosis in human prostate cancer cells [Gu et al., 2010b, Ahonen et al., 2003, Moser et al., 2012].

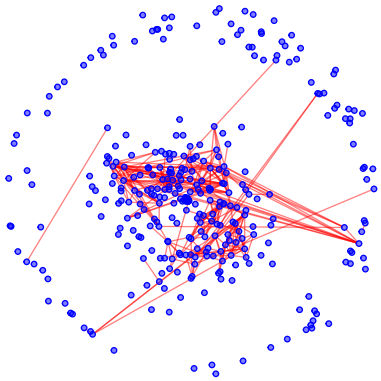
Furthermore, via the Kolmogorov-Smirnov test, we fail to reject the hypothesis that the degrees of the nodes for the regularized score matching graph estimate follow a power law distribution, with significance level of 0.05. On the other hand, we reject this hypothesis for all other generated estimates at the same significance level. There is evidence that genetic networks are 'scale-free', which implies that their degree distribution can be approximated by a power law distribution [Albert, 2005, Barabási and Albert, 1999, Jeong et al., 2001]. In this aspect, the topology of regularized score matching estimate is most similar to the hypothesized structure of gene networks.

Finally, we would like to emphasize that we do not intend to claim that regularized score matching provides the *best* estimate of the underlying gene network, as the truth is unknown to us. What we can posit is that truncated Gaussian may be a useful model that provides potentially valid targets for therapy which may be missed by other methods.

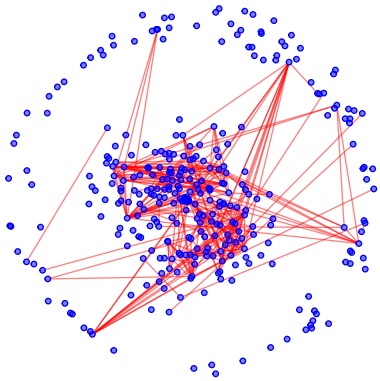


(a) Reg. score matching

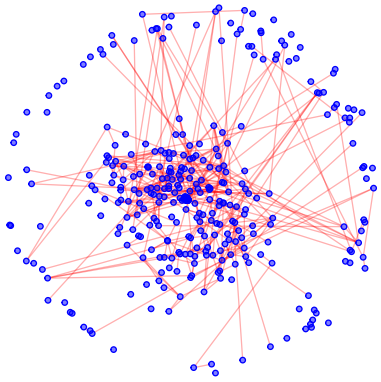
(b) SPACE



(c) Glasso



(d) SKEPTIC



(e) SPACEJAM

Figure 2.11: Topology of inferred networks of $|E| = 333$ or 334 edges for all considered methods. Layout of nodes is fixed across graph estimates.

2.7 Discussion

This chapter proposes the use of regularized score matching for estimation of conditional independence graphs in high dimensions. The focus is on modifying the score matching loss of Hyvärinen [2005] with an ℓ_1 penalty to accommodate underlying sparsity, which is in the spirit of popular existing methods such as glasso and neighborhood selection. This said, any other regularization scheme can be considered instead. For instance, the method from Defazio and Caetano [2012] can be applied to encourage hub structure in the inferred graph.

Our study of the Gaussian example of Meinshausen [2008] suggests that ℓ_1-regularized score matching falls in between neighborhood selection and glasso in terms of conditions for required for graph selection consistency. Here, the glasso requires the most stringent conditions, and the score matching approach appears to be similar to pseudo-likelihood methods that work with symmetric estimates of precision matrices, such as SPACE [Peng et al., 2009] and subsequent reformulations such as CONCORD [Khare et al., 2015]. However, as demonstrated, regularized score matching is particularly convenient in that the score matching loss is a quadratic function, even for non-Gaussian exponential families. This brings about piecewise linear solution paths and allows for a simple theoretical analysis.

Regularized score matching is an interesting method for Gaussian models, as we showed empirically and theoretically. In particular, for consistency (under the usual irrepresentability conditions), the sample n must be on the order $d^2 \log p$, which matches the conditions for the existing methods mentioned above. However, as our numerical experiments show, regularized score matching really shines in the context of non-Gaussian models, where it eliminates the need to deal with computationally intractable normalization constants in a way that the loss continues to be a quadratic function of parameters.

From a practical point of view, regularized score matching opens a lot of new possibilities for graphical modeling such as the truncated normal model we applied to RNAseq data. Additionally, we anticipate that the simple structure of score matching will lead to further advances in graphical modeling methodology-wise. For instance, we hypothesize this sim-

plicity can be leveraged to develop new methods for tuning regularization parameters, as in Chichignoud et al. [2014]. In the next chapter, we address how this framework can be readily adapted to accommodate missing data.

Reg. score matching	Glasso	SKEPTIC	SPACE	SPACEJAM
CCNE2 (19)	EP300 (20)	PIK3CA (23)	TRAF6 (9)	BHX (10)
PIK3CG (16)	SOS1 (17)	FZD7 (18)	TPR (9)	SOS2 (9)
BRCA2 (13)	BAD (16)	PDGFRB (17)	SOS1 (9)	TRAF6 (8)
BIRC5 (12)	TPR (13)	TGFBR2 (16)	JAK1 (9)	TGFBR2 (8)
SKP2 (10)	RBX1 (13)	TCEB2 (16)	EP300 (9)	SOS1 (8)
PIK3CD(10)	PIK3CD (12)	MMP2 (16)	SOS2 (8)	RRM2 (8)
LAMB3 (10)	LAMA4 (12)	LAMA4 (16)	EGFR (8)	PDGFRB (8)
STAT5B (9)	HRAS (12)	GLI2 (15)	CBL (8)	EP300 (8)
HRAS (9)	GLI2 (12)	SOS1 (14)	BAX (8)	PIK3CA (7)
PDGFRB (8)	TRAF6 (11)	PDGFRA (14)	APPL1 (8)	ARNT (7)
GSTP1 (8)	TGFBR2 (11)	MITF (14)		
	TCEB2 (11)	EP300 (14)		
	SPI1 (11)			
	SOS2 (11)			
	PDGFRB (11)			
	MAP2K2 (11)			
	APPL1 (11)			

Table 2.1: The most densely connected genes according to the estimated graphs generated via nonnegative regularized score matching, glasso, SKEPTIC, SPACE and SPACEJAM. The number in parenthesis corresponds to the estimated degree of the gene.

Chapter 3

EXTENSIONS OF REGULARIZED SCORE MATCHING FOR MISSING DATA PROBLEMS

3.1 Introduction

As discussed in Section 2.2, pairwise interaction models form a flexible class of multivariate models that, under sparse parameterization, have intuitive graphical representations. When the data is fully observed, the problem of estimating pairwise interaction models is fairly well studied: see review presented in the introductory chapter. In the previous chapter, we proposed regularized score matching as a flexible framework for estimating Gaussian *and* more especially, non-Gaussian continuous pairwise interaction models in the high-dimensional setting, as it requires no knowledge of the typically intractable normalizing constant for the latter. The method was shown to be sparsistent with appropriate scaling of n, the sample size, with p, the dimension, and d, the nodal degree or maximum non-zero entries in any given row of the symmetric interaction matrix $\mathbf{\Theta}$ which parametrizes the model. For example, for Gaussian graphical models, which are a type of pairwise interaction model, n needs to scale at least with $d^2 \log p$.

However, data is rarely fully observed. To illustrate, pairwise interaction models have been broadly applied to analyze (often high-dimensional) data from biology, such as data from microarray or RNAseq experiments. These types of data frequently feature a considerable amount of missingness that arise for reasons such as faulty machinery, inability to collect data in longitudinal studies, human error, and limits of experimental designs.

To estimate graphical models from continuous data with missing-completely-at-random observations and arising from a pairwise interaction model in the high-dimensional setting, we propose, in this chapter, adapting the regularized score matching framework from the

previous chapter. Our motivation is simple. As discussed in Forbes and Lauritzen [2015] and Section 2.3 of this thesis, and as noted in concurrent works by Janofsky [2015] and Sun et al. [2015a], when working with continuous pairwise interaction models, the regularized score matching objective becomes a semidefinite quadratic function of the parameter vector, augmented with an ℓ_1 penalty. This makes it structurally analogous to *lasso* [Tibshirani, 1996], and as a result, we are able to leverage ideas from high-dimensional regression to treat our missing data problem.

The Expectation-Maximization (EM) algorithm forms a traditional approach to missing data, and its application to high-dimensional regression is explored in Städler and Bühlmann [2012]. However, we turn instead to an alternative framework proposed in Loh and Wainwright [2012], whose approach, in brief summary, entails plugging in surrogates derived solely from the available data. A similar idea is presented in Kolar and Xing [2012], who work with the glasso objective and target Gaussian graphical models more explicitly. Unlike with EM, the approach presented in Loh and Wainwright [2012] has theoretical guarantees that the Euclidean distance between the estimate obtained by composite gradient descent and the truth is small with high probability when n scales at least with $d \log p$ and n and p are large, where d is the number of non-zero elements in the true regression coefficient vector.

The remainder of this chapter is organized as follows. In Section 3.2, we precisely describe how the methods proposed in Loh and Wainwright [2012] and Kolar and Xing [2012] may be combined with regularized score matching to allow consistent estimation of sparse pairwise interaction models in the presence of missing data, and give a composite gradient descent algorithm for computing the estimate. In Section 3.3, we prove consistency of our method by providing non-asymptotic bounds on the statistical and optimization errors — in the spectral and matrix ℓ_1 norm — which hold with high probability. In Section 3.4, we compare the performance of our method against alternative imputation strategies such as marginal mean imputation. We follow up with a real data application, where we return to the RNAseq dataset from Section 2.6. We conclude with a discussion of our findings and possible avenues for future research in Section 3.6. Proofs are largely deferred to the supplement.

3.2 Accommodating missing data using plug-in surrogates

Reusing notation from the previous chapter, let $\mathbf{x} \in \mathbb{R}^{n \times p}$ be a data matrix whose rows $x^{(i)}$ represent n independent observations of the random vector $X = (X_1, \ldots, X_p)$. In addition, let $\mathbf{D} = (\delta_{ij})$ be an $n \times p$ matrix with i.i.d. Bernoulli$(1-\rho)$ entries, for $\rho \in [0, 1)$. Suppose that instead of the complete data $\mathbf{x}$, we merely have the data matrix

$$\mathbf{z} = \mathbf{x} \circ \mathbf{D} \in \mathbb{R}^{n \times p}, \tag{3.1}$$

with $\circ$ denoting entrywise multiplication, otherwise known as the Hadamard product. In other words, $z_{ij} = x_{ij}$ if $\delta_{ij} = 1$ and $z_{ij} = 0$ otherwise. Here, 0's are used to represent missing values, which is justified because the singleton set $\{0\}$ is a null set for the Lebesgue measure. Note that this implies that the data is missing-completely-at-random.

Lemma 1 in Chapter 2 informs us that the regularized score matching objective for continuous pairwise interaction models can be written as

$$\hat{\theta} \in \arg\min_{\theta \in \mathrm{Sym}_p} \frac{1}{2}\theta^T \mathbf{\Gamma}(\mathbf{x})\theta - \gamma(\mathbf{x})^T\theta + \lambda\|\theta\|_1. \tag{3.2}$$

Recall that Sym_p represents the entire space of symmetric $p \times p$ matrices, and note that the constraint on $\hat{\theta}$ refers to its interaction matrix form, i.e. $\hat{\mathbf{\Theta}}$ (where $\hat{\theta} = \mathrm{vec}(\hat{\mathbf{\Theta}})$) being symmetric.

Since we do not observe $\mathbf{x}$, we cannot directly estimate θ^* via (3.2). To accommodate missing data, we leverage the similarities between (3.2) and the lasso for linear regression, and turn to ideas presented in Loh and Wainwright [2012] for handling high-dimensional linear regression with missing values. We propose substituting $\mathbf{\Gamma}(\mathbf{x})$ and $\gamma(\mathbf{x})$ with suitable surrogates derived solely from the observed $\mathbf{z}$. A similar idea is presented by Kolar and Xing [2012] who work with the glasso objective of Yuan and Lin [2007]. Their ideas can be integrated into the framework of Loh and Wainwright [2012].

Denote the surrogates of $\mathbf{\Gamma}(\mathbf{x})$ and $\gamma(\mathbf{x})$ based on either $\mathbf{z}$ or $\mathbf{x}$ (depending on approach, see below) as $\check{\mathbf{\Gamma}}$ and $\check{\gamma}$, respectively. The general idea is to ensure that the resulting estimating equations are unbiased, in the sense that $\mathbb{E}_{\theta^*}[\check{\mathbf{\Gamma}}] = \mathbb{E}_{\theta^*}[\mathbf{\Gamma}(\mathbf{x})]$ and $\mathbb{E}_{\theta^*}[\check{\gamma}] = \mathbb{E}_{\theta^*}[\gamma(\mathbf{x})]$.

In Loh and Wainwright [2012], $\mathbf{z}$ is plugged in place of $\mathbf{x}$ into $\mathbf{\Gamma}(\mathbf{x})$ and $\gamma(\mathbf{x})$, and the expectations are corrected via a function of the probability of missingness, ρ, to form $\tilde{\mathbf{\Gamma}}$ and $\tilde{\gamma}$. Here ρ is treated as known; in practice we use the very accurate estimate of ρ obtained from the count of missing values (with no effect on later theory). In Kolar and Xing [2012], the surrogates are $\mathbf{\Gamma}(\mathbf{x})$ and $\gamma(\mathbf{x})$ computed based on observed instantiations of the statistics: e.g., if an element of $\mathbf{\Gamma}(\mathbf{x})$ requires the empirical average of $X_j^2 X_k^2$, we compute the empirical average of $X_j^2 X_k^2$ based on samples when X_j and X_k are both observed. We give the explicit forms of $\check{\mathbf{\Gamma}}$ and $\check{\gamma}$ for when $\mathbf{x}$ arises as i.i.d. Gaussian and non-negative Gaussian observations as continuations of Examples 1 and 2 from Chapter 2.

Example 1 (cont.). For the centered Gaussian case, $\mathbf{\Gamma}(\mathbf{x}) = \mathbf{I}_{p\times p} \otimes \mathbf{W}$, with $\mathbf{W} = \mathbf{x}^T\mathbf{x}/n$ as the sample covariance matrix, and $\gamma(\mathbf{x}) = \gamma = \text{vec}(\mathbf{I}_{p\times p})$ in (3.2). Extending the approach of Loh and Wainwright [2012], appropriate surrogates for $\mathbf{\Gamma}(\mathbf{x})$ and $\gamma(\mathbf{x})$ are

$$\check{\mathbf{\Gamma}} = \mathbf{\Gamma}(\mathbf{z}) \oslash (\mathbf{I}_{p\times p} \otimes \mathbf{M}), \qquad \check{\gamma} = \gamma, \tag{3.3}$$

with $\mathbf{M} = (m_{jk}) \in \mathbb{R}^{p\times p}$ and

$$m_{jk} = \begin{cases} 1-\rho & \text{if } j = k, \\ (1-\rho)^2 & \text{if } j \neq k. \end{cases} \tag{3.4}$$

The operator '$\oslash$' refers to element-wise division. In contrast, the approach of Kolar and Xing [2012] would lead to

$$\tilde{\mathbf{\Gamma}} = \mathbf{I}_{p\times p} \otimes \tilde{\mathbf{W}}, \qquad \tilde{\gamma}(\mathbf{z}) = \gamma, \tag{3.5}$$

with $\tilde{\mathbf{W}} = (\tilde{w}_{jk})$ given by

$$\tilde{w}_{jk} = \frac{\sum_{i=1}^n \delta_{ij}\delta_{ik}x_{ij}x_{ik}}{\sum_{i=1}^n \delta_{ij}\delta_{ik}}.$$

Example 2 (cont.). For the non-negative Gaussian case, $\mathbf{\Gamma}(\mathbf{x})$ is a $p^2 \times p^2$ block diagonal matrix with jth block given by

$$\frac{1}{n}\sum_{i=1}^n x_{ij}^2 x^{(i)} x^{(i)T},$$

and $\gamma = \text{vec}(\mathbf{W}) + \text{vec}(\mathbf{W}_{\text{diag}})$. In the presence of missing data, we may adapt the strategy of Loh and Wainwright [2012] and substitute $\mathbf{\Gamma}$ and γ with

$$\check{\mathbf{\Gamma}} = \mathbf{\Gamma}(\mathbf{z}) \oplusdot (\mathbf{I}_{p\times p} \otimes \mathbf{M}_+), \qquad \check{\gamma} = \gamma(\mathbf{z}) \oplusdot \text{vec}(\mathbf{M}), \tag{3.6}$$

respectively, where the elements of $\mathbf{M}$ are given by (3.4), and $\mathbf{M}_+ = (m_{+,jk})$ with

$$m_{+,jk} = \begin{cases} 1-\rho & \text{if } j = k = 1, \\ (1-\rho)^2 & \text{if } j = k \neq 1, \\ (1-\rho)^3 & \text{if } j \neq k. \end{cases} \tag{3.7}$$

Alternatively, in the spirit of Kolar and Xing [2012], we can consider a different set of surrogates with $\tilde{\mathbf{\Gamma}}$ a $p^2 \times p^2$ block diagonal matrix with the jth block given by

$$\sum_{i=1}^n \check{x}_{ij}^2 \check{x}^{(i)} \check{x}^{(i)T} \oplusdot \left(\sum_{i=1}^n \delta_{ij} \delta^{(i)} \delta^{(i)T} \right), \tag{3.8}$$

and

$$\check{\gamma} = \text{vec}(\tilde{\mathbf{W}}) \tag{3.9}$$

with $\check{\mathbf{x}} = \mathbf{x} \circ \mathbf{D}$ and $\tilde{\mathbf{W}}$ as defined in Example 1.

The main challenge is that plugging in the surrogates into the regularized score matching problem in (3.2) can render it possibly nonconvex due to the presence of negative eigenvalues in $\tilde{\mathbf{\Gamma}}$. In fact, if $n < p$, the problem is guaranteed to be nonconvex with probability 1. The same issue was observed for regression in Loh and Wainwright [2012]. Kolar and Xing [2012] did not experience this issue because their surrogate only appears in a linear term of the glasso problem. However, the glasso objective can be unbounded from below, upon plugging in the surrogate, despite still being strictly convex. To counter this, we use the following regularized estimator:

$$\hat{\theta} \in \arg \min_{\substack{\|\mathbf{\Theta}\|_1 \leqslant R \\ \mathbf{\Theta} \in \text{Sym}_p}} \frac{1}{2} \theta^T \check{\mathbf{\Gamma}} \theta - \check{\gamma}^T \theta + \lambda \|\theta\|_1. \tag{3.10}$$

Let $\mathbf{\Theta}^*$ be the true value of $\mathbf{\Theta}$. From a theoretical perspective, the radius bound R, needed to counter the non-convexity of the objective, should be set large enough such that $\mathbf{\Theta}^*$

remains feasible (i.e., $R \geqslant \|\mathbf{\Theta}^*\|_1$). In practice, $\|\mathbf{\Theta}^*\|_1$ is not known and the radius becomes an additional tuning parameter. We refer to (3.10) as the 'symmetric' problem.

Since $\check{\mathbf{\Gamma}}$ is a diagonal block matrix, the objective in (3.10) actually decouples over the p columns of the interaction matrix $\mathbf{\Theta}$, and it can be more computationally and theoretically convenient to consider an alternative problem where there is a constraint on ℓ_1 norm on each column:

$$\hat{\theta} \in \arg\min_{\|\theta_j\|_1 \leqslant R \ \forall j} \frac{1}{2}\theta^T \check{\mathbf{\Gamma}} \theta - \check{\gamma}^T \theta + \lambda \|\theta\|_1. \tag{3.11}$$

Here θ_j is the jth column of the interaction matrix $\mathbf{\Theta}$. We have dropped the symmetry constraint to ease the computation, so the resultant $\mathbf{\Theta}$ estimator need not be symmetric. To obtain a symmetric estimate, we can consider a two-step procedure. That is, write $\hat{\mathbf{\Theta}}^{(1)}$ as the preliminary estimate from solving (3.11), and obtain the final estimator $\hat{\mathbf{\Theta}}^{(2)}$ via

$$\hat{\mathbf{\Theta}}^{(2)} \in \arg\min_{\mathbf{\Theta} \in \mathrm{Sym}_p} \left\| \mathbf{\Theta} - \hat{\mathbf{\Theta}}^{(1)} \right\|_1. \tag{3.12}$$

Problem (3.12) can be readily solved via linear programming. We refer to (3.11) as the 'block' problem.

3.2.1 Implementation

We propose a composite gradient descent algorithm [Nesterov, 2007] for solving either (3.10) and (3.11). To avoid repetition, we discuss the algorithm for the 'block' problem (3.11) only. Adapting the algorithm to the 'symmetric' problem is straightforward.

The 'block' problem decomposes into p sub-problems, the jth of which is given by

$$\hat{\theta}_j \in \arg\min_{\|\theta_j\|_1 \leqslant R} \frac{1}{2}\theta_j^T \check{\mathbf{\Gamma}}_j \theta_j - \check{\gamma}_j^T \theta_j + \lambda \|\theta_j\|_1. \tag{3.13}$$

Now, $\check{\mathbf{\Gamma}}_j$ is the jth diagonal $p \times p$ block of $\check{\mathbf{\Gamma}}$ and $\check{\gamma}_j = \check{\gamma}_{(jp-p+1):jp}$ is the jth 'column' of $\check{\gamma}$. Let $\mathcal{L}(\theta_j) = \frac{1}{2}\theta_j^T \check{\mathbf{\Gamma}}_j \theta_j - \check{\gamma}_j^T \theta_j$. Then, the gradient of the quadratic loss function (with respect to θ_j now) in (3.13) is $\nabla \mathcal{L}_j(\theta_j) = \check{\mathbf{\Gamma}}_j \theta_j - \check{\gamma}_j$.

Algorithm 3

Input: Initial estimate $\hat{\theta}^{(0)}$

Input: t_{max}, maximum number of iterations

Input: ϵ, the maximal tolerance level

Initialize $t \leftarrow 1$

Initialize crit $\leftarrow \epsilon + 1$ (crit stands for convergence criteria)

while crit $> \epsilon$ or $t < t_{max}$ **do**

  for $j \leftarrow 1, 2, \ldots, s$ **do**

    $\theta_j^{(t)} \leftarrow \Pi\left[\text{Soft}\left(\theta_j^{(t-1)} - \frac{1}{\eta}\nabla\mathcal{L}(\theta_j^{(t-1)}), \lambda\right)\right]$.

  end for

  crit $\leftarrow \|\hat{\theta}^{(t)} - \hat{\theta}^{(t-1)}\|_1$

  $t \leftarrow t + 1$

end while

The composite gradient descent algorithm improves an initial point $\theta_j^{(0)}$ via the following iterative procedure,

$$\theta_j^{(t)} \leftarrow \arg\min_{\|\theta_j\|_1 \leqslant R} \mathcal{L}_j(\theta_j^{(t-1)}) + \langle \nabla L_j(\theta_j^{(t-1)}), \theta_j - \theta_j^{(t-1)} \rangle + \frac{\eta}{2}\|\theta_j - \theta_j^{(t-1)}\|_2^2 + \lambda\|\theta_j\|_1 \quad (3.14)$$

until a convergence criterion is satisfied.

The update (3.14) is equivalent to

$$\theta_j^t \leftarrow \Pi\left[\text{Soft}\left(\theta_j^{(t-1)} - \frac{1}{\eta}\nabla\mathcal{L}(\theta_j^{(t-1)}), \lambda\right)\right], \quad (3.15)$$

where 'Soft' is the standard soft-thresholding operator (2.39), and Π denotes the ℓ_2-projection onto the ℓ_1 ball of radius $\|\theta_j^*\|_1$ for all $j \in \{1, \ldots, p\}$. These projections can be done in $\mathcal{O}(p^2)$ operations via the algorithm provided in Duchi et al. [2008]. See Algorithm 3.

3.3 Theoretical results on consistency

We attempt a different perspective from Chapter 2 and focus on estimation error. As in the previous chapter, let

$$\mathbf{\Gamma}^* = \mathbb{E}[\mathbf{\Gamma}(\mathbf{x})] = \mathbb{E}[\tilde{\mathbf{\Gamma}}]$$
$$\gamma^* = \mathbb{E}[\gamma(\mathbf{x})] = \mathbb{E}[\tilde{\gamma}].$$

Theorem 2. *Suppose $\mathbf{\Gamma}^*$ is positive definite and the surrogates $(\tilde{\mathbf{\Gamma}}, \tilde{\gamma})$ satisfy the deviation bounds*

$$\|\check{\mathbf{\Gamma}} - \mathbf{\Gamma}^*\|_\infty \leqslant \epsilon_1, \text{ and} \tag{3.16}$$
$$\|\check{\gamma} - \mathbf{\Gamma}^*\theta^*\|_\infty \leqslant \epsilon_2. \tag{3.17}$$

Then with $\lambda = 4|||\Theta^*|||_1\epsilon_1 + 2\epsilon_2$ and $R = |||\mathbf{\Theta}^*|||_1$, the global optimum $\hat{\mathbf{\Theta}}^{(1)}$ of the 'block' problem (3.11) satisfies the bounds

$$\left|\left|\left|\hat{\mathbf{\Theta}}^{(1)} - \mathbf{\Theta}^*\right|\right|\right|_1 \leqslant \frac{48d}{\nu_{\min}(\mathbf{\Gamma}^*)}\lambda, \tag{3.18}$$

where $\nu_{\min}(\mathbf{\Gamma}^*)$ refers to the smallest eigenvalue of $\mathbf{\Gamma}^*$. For $\hat{\mathbf{\Theta}}^{(2)}$, the upper bound (3.18) is simply doubled.

Theorem 2 tells us that despite the non-convexity of the 'block' problem (3.11), the global optimum lies in a ball around the true parameter θ^* whose radius is determined by the deviation bound. Under additional distributional assumptions on the data-generating mechanism, we can show that if n is sufficiently large, the deviation conditions will hold with high probability. These results are provided in Corollaries 3 and 4. We note that the ℓ_1 matrix norm (3.18) bounds the matrix operator norm.

Theorem 2 characterizes the quality of the global optimum for the nonconvex problem (3.11) but not the estimates generated by the proposed composite gradient descent algorithm (3.14). In other words, while we know that the composite gradient descent algorithm is guaranteed to converge to a stationary point (or possibly multiple), we do not know how

closely such a point lies to the truth. Theorem 3 addresses this by showing that all stationary points are guaranteed to be equally 'good' if the same conditions from Theorem 2 hold.

Theorem 3. *Suppose the assumptions in Theorem 2 hold with* $d\epsilon_1 \lesssim 1$*, with* ϵ_1 *as defined in (3.16). Then, for any global optimum* $\hat{\theta}_j$ *of the* j*th block problem (3.13), there exist positive constants* c_1 *and* c_2 *depending only on* $\mathbf{\Theta}^*$*, the product* $d\epsilon_1$*, and a contraction coefficient* $\zeta \in (0,1)$ *such that the composite gradient descent updates (3.15) satisfy*

$$\|\theta_j^{(t)} - \hat{\theta}_j\|_2^2 \leqslant c_1 \|\theta_j^* - \hat{\theta}_j\|_2^2, \tag{3.19}$$

for all $t \geqslant T$*, with* $T \equiv c_2 \log\left(\frac{\phi(\theta_j^0) - \phi(\hat{\theta}_j)}{c_1 \|\theta_j^* - \hat{\theta}_j^t\|_2^2}\right) / \log(1/\zeta)$*, where* $\phi(\theta_j) = \mathcal{L}(\theta_j) + \lambda \|\theta_j\|_1$.

Theorem 3 is a consequence of Theorem 2 in Agarwal et al. [2012] and observing that under the added condition that $d\epsilon_1 \lesssim 1$, $\tilde{\mathbf{\Gamma}}$ satisfies what Loh and Wainwright [2012] refer to as restricted eigenvalue conditions (defined in (B.12) and (B.13) in the Appendix) for certain constants. In words, our Theorem 3 shows that once a certain iteration has been reached, the optimization error between our current estimate $\theta_j^{(t)}$ and $\hat{\theta}_j$ is controlled by the statistical error bounded in Theorem 2. Hence, if we run the composite gradient descent for enough iterations, all estimates generated by the composite gradient descent algorithm lie in a ball around the global optima, and all global optima lie in a ball of similar order magnitude about the truth. This result, is again, deterministic, and we defer to Corollaries 3 and 4 to show that under certain distributional assumptions, for n large enough, the assumptions needed for Theorem 3 hold with high probability.

Remark 2. *in the Appendix, where we address support recovery/sparsistency of the method for completeness, we also manage to show that the stationary points within the feasible region are in fact unique if more stringent conditions hold. In other words, there is a single stationary point, and it corresponds to the global optimum. This has been observed In all our numerical experiments: despite initializing the composite gradient descent algorithm at different starting points, the end output was always unique.*

Corollary 3. *Suppose* $\mathbf{x} \in \mathbb{R}^{n\times p}$ *is generated from a normal distribution* $N(0, \mathbf{\Sigma}^*)$, $\mathbf{\Sigma}^*$ *positive definite, and let* $\mathbf{z} \in \mathbb{R}^{n\times p}$ *be the observed data matrix generated according to (3.1) with parameter* $\rho \in [0, 1)$. *Furthermore, let* $\mathbf{K}^* = (\mathbf{\Sigma}^*)^{-1}$ *represent the true precision matrix. Then, for the set of surrogates given by (3.3),*

1. *the deviation conditions (3.16) and (3.17) are satisfied with*

$$\epsilon_1 \lesssim \frac{(1 + 4/(1-\rho)) \max_j \sigma^*_{jj}}{(1-\rho)^2} \sqrt{\frac{\log p}{n}},$$
$$\epsilon_2 = 0,$$

   respectively, with probability $1 - c_1 exp(-c_2 \log p)$ *for some universal constants* c_1, $c_2 > 0$.

2. *if* λ *and* R *are chosen according to Theorem 2 with the* ϵ_1 *and* ϵ_2 *as in (a) and*

$$n \gtrsim \max\left\{ \frac{\nu_{\max}(\mathbf{\Gamma}^*)^2 |||\mathbf{K}^*|||_1^2}{\nu_{\min}(\mathbf{\Gamma}^*)^2 (1-\rho)^6}, 1 \right\} d^2 \log p, \tag{3.20}$$

   with $\nu_{\max}(\mathbf{\Gamma}^*)$ *and* $\nu_{\min}(\mathbf{\Gamma}^*)$ *defined to be the largest and smallest eigenvalue of* $\mathbf{\Gamma}^*$, *respectively, Theorems 2 and 3 hold with*

$$\left\| \hat{\mathbf{\Theta}} - \mathbf{\Theta}^* \right\|_1 \lesssim d \sqrt{\frac{\log p}{n}}, \tag{3.21}$$

   in particular, with the same probability as in (a).

If we instead choose to use the set of surrogates given by (3.5), the above holds true for

$$\epsilon_1 \lesssim \frac{(1 + 4/(1-\rho)) \max_j \sigma^*_{jj}}{(1-\rho)^3} \sqrt{\frac{\log p}{n}},$$
$$\epsilon_2 = 0.$$

Corollary 4. *Suppose* $\mathbf{x} \in \mathbb{R}^{n\times p}$ *is generated from a non-negative Gaussian distribution with parameter* $\mathbf{K}^*$, *i.e.* $N(0, (\mathbf{K}^*)^{-1})$ *truncated at* $\mathbb{R}^p_+$, $\mathbf{K}^*$ *positive definite, and let* $\mathbf{z} \in \mathbb{R}^{n\times p}$ *be the observed data matrix generated according to (3.1) with parameter* $\rho \in [0, 1)$. *Define* $v_1 = \max_{j,k,l} Var[X_j^2 X_k X_l]$ *and* $v_2 = \max_{j,k} Var[X_j X_k]$. *Then, for the set of surrogates given either by (3.6) or (3.8),*

1. *the deviation conditions (3.16) and (3.17) are satisfied with*

$$\epsilon_1 \lesssim \frac{1}{(1-\rho)^{9/2}}\sqrt{v_1\frac{(\log p)^8}{n}},$$
$$\epsilon_2 \lesssim \frac{1}{(1-\rho)^{5/2}}\sqrt{v_2\frac{(\log p)^4}{n}},$$

respectively, with probability $1-c_1 exp(-c_2 \log p)$ *for some universal constants* c_1, $c_2 > 0$

2. *if* λ *and* R *are chosen according to Theorem 2 with* ϵ_1 *and* ϵ_2 *as in (a) and*

$$n \gtrsim \max\left\{\frac{\max\{v_1, v_2\}|||\mathbf{\Theta}^*|||_1^2}{\nu_{\min}(\mathbf{\Gamma}^*)^2(1-\rho)^9}, 1\right\} d^2(\log p)^8, \tag{3.22}$$

Theorems 2 and 3 hold, with

$$\left|\left|\left|\hat{\mathbf{\Theta}} - \mathbf{\Theta}^*\right|\right|\right|_1 \lesssim d\sqrt{\frac{\log p}{n}}, \tag{3.23}$$

in particular, with the same probability as in (a).

It is unlikely that the lower bound for n in Corollary 4 is optimal in the sense that a lower power for $\log p$ may be sufficient. However, experiments conducted in the Appendix for Chapter 2 suggest that the optimal exponent is not much lower than 8 based on numerical experiments done on a simple chain graph with $d = 2$.

Remark 3. *In both corollaries, we refer to* ρ*, the probability of an observation being missing, as fixed across columns of* $\mathbf{x}$*. It is easy to extend the proofs when the probability varies across columns. Furthermore, it is straightforward to extend to proofs to accommodate the scenario where* ρ *is unknown; compare Lemma 4 in Loh and Wainwright [2012].*

3.4 Numerical experiments

3.4.1 Empirical verification of theoretical results

We conduct a set of experiments to provide empirical support for the claim that the scaling $n = \Omega(d^2 \log p)$ is sharp for the Gaussian setting (Corollary 3). These are similar to those

that can be found in the Appendix of Chapter 2. We attempt two different setups: in the first, we vary p while holding d fixed, and in the second, we vary d while holding p fixed. In both settings, we study how the error curves $\left|\left|\left|\hat{\mathbf{\Theta}} - \mathbf{\Theta}^*\right|\right|\right|_1$, corresponding to different n, change. In these experiments, $\hat{\mathbf{\Theta}}$ refers to the output of the composite gradient descent algorithm. If the lower bound is sharp, the error curves should overlap upon rescaling n with $d^2 \log p$, when all other factors remain constant. The chain graph allows us to naturally vary p while holding d fixed while the star graph is ideal for the other setting. Recall that in the Gaussian setup (Example 1), $\mathbf{\Theta}^* = \mathbf{K}^*$, where $\mathbf{K}^*$ is the true precision matrix.

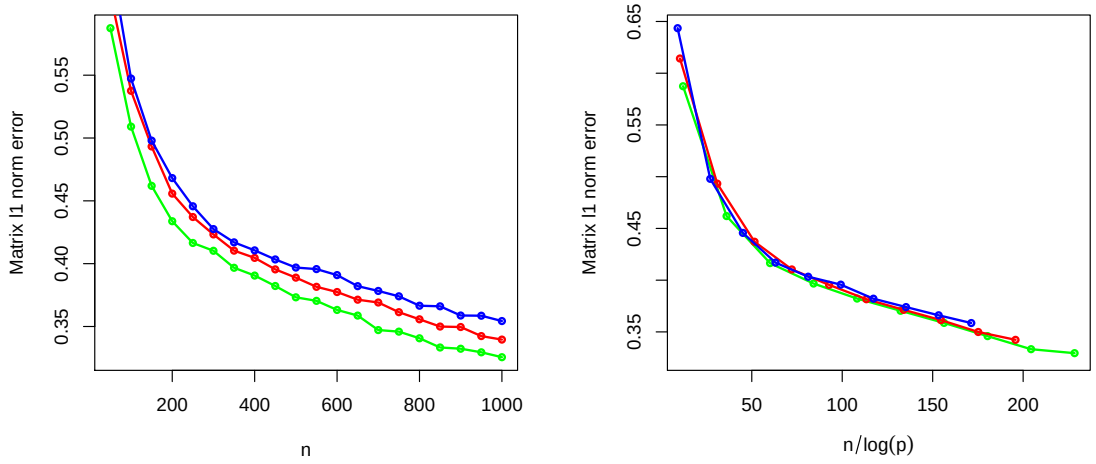
Dependence on number of nodes p

Consider first the case where the underlying conditional independence graph G, as encoded in $\mathbf{K}^*$, is a linear chain of length $p \in \{64, 128, 256\}$. The degree d is always 2, and we choose the tridiagonal precision matrix $\mathbf{K}^*$ to have entries $\kappa^*_{jk} = 0.1$ if $(j, k) \in E$ and $\kappa^*_{jj} = 1$ for $j = 1, 2, \ldots, p$. Here, all other terms in the lower bound forming n in (3.20) are constant (or near constant, and we correct for this) over the range of p. The missing parameter ρ is chosen to be 0.8. Following the theory, we let $\lambda_n = 0.005\sqrt{\log p/n}$.

Figure 3.1 shows $\left|\left|\left|\hat{\mathbf{K}} - \mathbf{K}^*\right|\right|\right|_1$ plotted against sample size n, with different curves corresponding to different p for the 'block' problem with surrogates (3.3). We observe, as expected, that the curves shift rightwards with increasing p. Furthermore, we succeed in confirming that the scaling presented in Corollary 3 is sharp, as rescaling n with $\log p$ causes all three curves to align. Nearly identical figures are obtained for surrogates (3.5), and are consequently excluded.

Dependence on node degree d

We fix the number of nodes to be $p = 100$ and vary d. Suppose that now G is a star graph with varying hub node degree $d \in \{15, 20, 25\}$. The precision matrix $\mathbf{K}^*$ is chosen such that $\kappa^*_{jk} = 0.1$ if $(j, k) \in E$ and $\kappa^*_{jj} = 1$. $\mathbf{K}^*$ is then normalized to have maximum eigenvalue 1.



(a) (b)

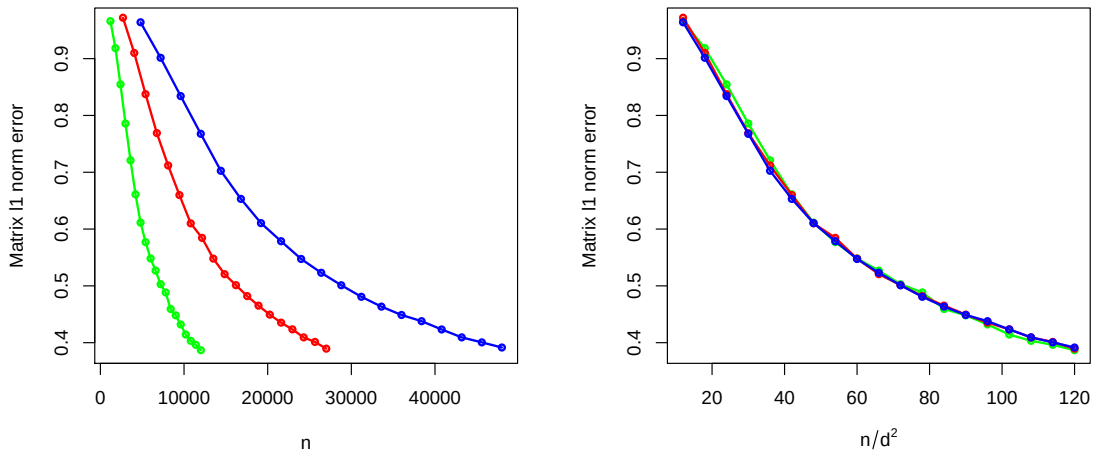
Figure 3.1: Plot of $\left\|\hat{\mathbf{K}} - \mathbf{K}^*\right\|_1$ where $\hat{\mathbf{K}}$ is obtained via composite gradient descent based on the nonconvex 'block' problem where the encoded conditional independence graph is a chain of varying length p. Panels (a) and (b) differ only in the scaling of the x-axis. The colored lines correspond to $p = 64$ (——), $p = 128$ (——), and $p = 256$ (——).

All other terms in the lower bound forming n in (3.20) are constant (or near constant). The missing parameter ρ is chosen to be 0.8. As before, we let $\lambda_n = 0.005\sqrt{\log p/n}$.

Figure 3.2 plots $\left\|\hat{\mathbf{K}} - \mathbf{K}^*\right\|_1$ against sample size n for different d. We observe, as expected, that the curves shift rightwards with increasing d. The scaling presented in Corollary 3 is further validated, as rescaling n with d^2 leads to alignment of the three curves.

3.4.2 Comparisons to other methods

In this section, we compare our method against alternative methods for estimating non-Gaussian pairwise interaction models with missing data when data is drawn from a non-negative Gaussian distribution (Example 2). We demonstrate the performance of both the



(a) (b)

Figure 3.2: Plot of $\left\|\left\|\hat{\mathbf{K}} - \mathbf{K}^*\right\|\right\|_1$ where $\hat{\mathbf{K}}$ is obtained via composite gradient descent based on the nonconvex 'block' problem where the encoded conditional independence graph is a star with varying degree d. Panels (a) and (b) differ only in the scaling of the x-axis. The colored lines correspond to $d = 10$ (—), $d = 15$ (—), and $d = 20$ (—).

'symmetric' and 'block' versions. Particularly, we compare against

(1) *Marginal mean* treatment: all missing values are set to the (empirical) marginal mean of the variable.

(2) *10-nearest neighbors* (10-NN) treatment: missing values are imputed using the function `impute.knn` from the `impute` package by Hastie et al. [2016]. All extra function parameters for the impute function are set to default.

We consider three different structures for the true conditional independence graph G:

(1) *Chain graph* (G1): a linear chain graph so $d = 2$. Set $\kappa^*_{jk} = 0.1$ if $(j, k) \in E$ and $\kappa_{jj} = 1$ for $j = 1, 2, \ldots, p$.

(2) *Lattice graph* (G2): a 4-neighbor lattice graph. Construct the matrix $\mathbf{R}^* = \left(r^*_{jk}\right)$ with entries $r_{jk} = 0.2$ if $(j,k) \in E$, and normalize each row by 1.5 times the absolute row sum. Set $r^*_{jj} = 1$ for $j = 1, 2, \ldots, p$, and symmetrize $\mathbf{R}^*$. Set $\mathbf{R}'^* = (\mathbf{R}^*)^{-1}$, and convert the former into a correlation matrix. Finally, set $\mathbf{K}^* = (\mathbf{R}'^*)^{-1}$

(3) *Erdős-Rényi graph* (G3): set probability of an edge occurring between any two nodes to be $3/p$. Repeat the procedure from *lattice graph* setup to construct $\mathbf{K}^*$.

We also vary the degree of missingness, setting ρ to be 0.6, 0.7, 0.8. At each combination of p and n, we fixed the pattern of missingness across all experiments. Here, ρ, contrary to previous definition provided in (3.1), actually refers to the expected proportion of missing values in the upper-left $(0.8n \times 0.8p)$-submatrix of $\mathbf{z}$. That is, all missing values occur within this pre-specified block of the data matrix $\mathbf{z}$, and all values outside this block are observed.

For all method(s), different $\hat{\mathbf{\Theta}}$'s may be obtained for different tuning parameters λ. Let Λ be the set of tuning parameters considered for each method. To select the tuning parameter, we use a criterion analogous to the Bayesian Information Criterion (BIC) proposed by Schwarz [1978], with the likelihood term substituted with the score matching loss, as suggested in Section 2.3.4. Note that we substitute $\mathbf{\Gamma}(\mathbf{x})$ and $\gamma(\mathbf{x})$ in (2.40) with their surrogates here.

Table 3.1 compares six missing data strategies: (1) imputation via marginal means (MM), (2) imputation via 10 nearest neighbors (10-NN), (3) 'symmetric' problem with surrogates (3.6) (LS), (4) 'symmetric' problem with surrogates (3.8) (KS) , (5) 'block' problem with surrogates (3.6) (LB) and (6) 'block' problem with surrogates (3.8) (KB). It is evident from these results that our strategies vastly outperform the two imputation-based competitors, after accounting for standard deviations, and yet are just as straightforward to implement. Overall, it appears that using surrogates based on the strategy of Kolar and Xing [2012] outperform those based on Loh and Wainwright [2012], but the difference is fairly negligible, when factoring in standard deviations.

Model	(p, n)	ρ	MM	10-NN	LS	KS	LB	KB
	(50, 400)	0.6	3.557 (0.364)	4.202 (0.517)	1.332 (0.319)	1.241 (0.217)	0.885 (0.28)	0.878 (0.274)
		0.7	3.329 (0.442)	3.808 (0.49)	1.139 (0.225)	1.124 (0.204)	0.89 (0.257)	0.882 (0.253)
		0.8	3.215 (0.414)	3.537 (0.434)	1.07 (0.185)	1.069 (0.186)	0.848 (0.198)	0.856 (0.201)
	(100, 1000)	0.6	2.644 (0.248)	3.128 (0.261)	1.159 (0.163)	1.116 (0.157)	0.784 (0.234)	0.775 (0.219)
G1		0.7	2.461 (0.24)	2.798 (0.311)	1.079 (0.133)	1.063 (0.131)	0.767 (0.199)	0.749 (0.189)
		0.8	2.383 (0.23)	2.589 (0.256)	1.025 (0.152)	1.017 (0.151)	0.689 (0.164)	0.688 (0.162)
	(150, 2000)	0.6	3.174 (0.277)	3.821 (0.34)	1.056 (0.135)	1.043 (0.123)	0.697 (0.182)	0.69 (0.181)
		0.7	2.9 (0.365)	3.359 (0.286)	0.979 (0.105)	0.985 (0.107)	0.65 (0.16)	0.645 (0.16)
		0.8	2.509 (0.425)	3.025 (0.339)	0.9 (0.096)	0.908 (0.096)	0.604 (0.14)	0.599 (0.14)
	(49, 750)	0.6	4.125 (0.371)	4.651 (0.478)	2.757 (0.408)	2.552 (0.275)	2.251 (0.509)	2.234 (0.481)
		0.7	3.81 (0.383)	4.28 (0.427)	2.429 (0.259)	2.39 (0.242)	2.139 (0.396)	2.168 (0.393)
		0.8	3.586 (0.48)	3.931 (0.422)	2.339 (0.297)	2.329 (0.285)	2.061 (0.31)	2.068 (0.316)
	(100, 1000)	0.6	8.855 (0.952)	10.798 (0.965)	2.908 (0.369)	2.688 (0.273)	2.373 (0.555)	2.337 (0.514)
G2		0.7	7.28 (1.174)	9.099 (1.143)	2.617 (0.274)	2.576 (0.25)	2.259 (0.454)	2.245 (0.438)
		0.8	6.156 (1.13)	7.406 (1.224)	2.454 (0.207)	2.437 (0.207)	2.123 (0.373)	2.13 (0.37)
	(144, 2000)	0.6	5.322 (0.536)	6.46 (0.463)	2.502 (0.192)	2.409 (0.187)	2.138 (0.456)	2.141 (0.434)
		0.7	4.311 (0.657)	5.562 (0.557)	2.367 (0.186)	2.335 (0.182)	2.049 (0.378)	2.05 (0.383)
		0.8	3.633 (0.601)	4.293 (0.778)	2.201 (0.162)	2.201 (0.164)	1.941 (0.292)	1.952 (0.304)
	(50, 400)	0.6	8.392 (0.996)	9.955 (1.285)	3.961 (1.167)	3.542 (0.661)	4.675 (1.321)	4.372 (1.06)
		0.7	7.655 (0.931)	8.763 (1.04)	3.119 (0.377)	3.078 (0.327)	4.234 (1.043)	4.109 (0.886)
		0.8	7.262 (0.929)	8.055 (1.022)	2.96 (0.292)	2.956 (0.293)	3.982 (0.723)	4.001 (0.697)
	(100, 1000)	0.6	8.863 (1.029)	10.908 (0.851)	3.114 (0.39)	2.952 (0.285)	3.921 (1.356)	3.733 (1.126)
G3		0.7	7.524 (1.282)	9.215 (1.171)	2.84 (0.254)	2.839 (0.239)	3.461 (1.059)	3.432 (0.992)
		0.8	6.47 (1.17)	7.846 (1.316)	2.804 (0.27)	2.809 (0.255)	3.215 (0.878)	3.152 (0.827)
	(150, 2000)	0.6	5.457 (0.583)	6.59 (0.575)	3.058 (0.35)	3.053 (0.342)	3.604 (1.173)	3.25 (0.948)
		0.7	4.655 (0.705)	5.699 (0.539)	2.992 (0.387)	3.01 (0.385)	2.851 (0.852)	2.601 (0.741)
		0.8	3.981 (0.584)	4.487 (0.726)	2.93 (0.408)	2.94 (0.409)	2.129 (0.538)	2.071 (0.534)

Table 3.1: Mean estimation error in the matrix ℓ_1 norm, $\left\|\left\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}^*\right\|\right\|_1$ across the three simulation setups for varying n, p, and ρ. Numbers in the brackets correspond to empirical standard deviations computed based on 100 independent trials.

3.5 Revisiting the RNAseq data example

We apply our proposed method to RNAseq dataset from Section 2.6. We focus explicitly on support recovery of $\mathbf{\Theta}^*$: the estimated graph summarizes gene-gene interactions potentially linked to disease development and, thus, is of scientific interest. As discussed in Section 2.6, truncated Gaussian models may be a suitable alternative to more conventional Gaussian models, so we assume the distribution of the observed data can be characterized by log-densities of the form

$$f_{\mu,\mathbf{K}}(x) \propto \exp\left\{\frac{1}{2}(x-\mu)^T\mathbf{K}(x-\mu)\right\}, \quad x \in \mathbb{R}^p_+,$$

as before.

As a refresher, we have RNAseq data on 487 prostate adenocarcinoma samples made available in The Cancer Genome Atlas dataset. We focus on 350 genes that belong to known cancer pathways in the Kyoto Encyclopedia of Genes and Genomes; of those, we additionally prune out those with more than 10% of its observations missing and whose signal-to-noise ratio exceed 5 (to improve algorithm stability).

We focus on the 'block' problem and use the 'OR' rule to obtain symmetric graph estimates. As in Section 3.4, we compare our results against those generated via marginal mean and 10-nearest neighbor imputation. To draw comparisons and aid visualization, we examine the topologies of the graphs corresponding to the tuning parameter(s) which yield $p = 315$ edges, as was done in the previous chapter.

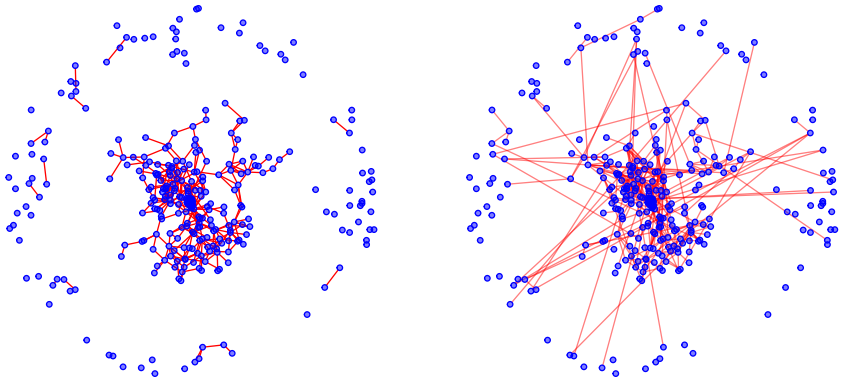
While the direct imputation methods only involve one tuning parameter, ours require two: we need to set the tuning parameter λ and the search radius R. We run the algorithm over a coarse grid of R's, derived by scaling the radius bounds suggested by the imputation methods by a positive factor $\alpha \in \{0.25, 0.5, 1, 1.5, 2\}$. Overall, it appears that the graph topologies are surprisingly stable with respect to R: to illustrate, the graphs generated by scaling the suggested radius by $\alpha = 0.25$ ($R \approx 16$) and $\alpha = 2$ ($R \approx 128$) – an 8-fold increase – share over 60% of their 315 edges.

The graphs generated by imputation methods are nearly identical to one another, as are the graphs generated by de-biasing via use of surrogates. Therefore, we only show results for marginal mean imputation and the 'block' method with surrogates (3.8) and (3.9) only. The topologies of the graphs with $p = 315$ edges for these two methods are presented in Figure 3.3. Because we have preserved the node layout across all graphs, the differences in the topologies we obtain from the two methods is rather noticeable. The 'block' method, based on different R, yields graph estimates which include edges that are not present in marginal mean imputation graph; likewise, it does not include others. On the other hand, it is reassuring to observe that some similarities are preserved between them and that overall the results are similar to those obtained in the previous chapter: after all, in this dataset, only a small number of genes ($\approx 10\%$) have missing observations. The list of hub genes, which we define to be genes with estimated degree $\geqslant 10$, is similar across imputation and de-biasing methods with different R; see Table 3.2.

Unfortunately, as was the case in previous chapter, it is not possible to determine which method produces the graph most representative of a truth, as a true network is not known. However, it is clear from this brief study that by accounting for missingness by de-biasing via surrogates versus traditional methods such as imputation, we will obtain topologically different graphs, which may alter the scope of further exploration.

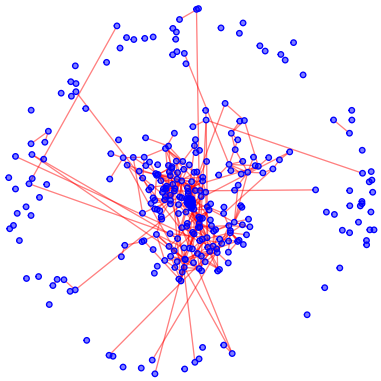
MM	KB, $\alpha = 0.25$	KB, $\alpha = 0.5$	KB, $\alpha = 1$	KB, $\alpha = 1.5$	KB, $\alpha = 2$
CCNE2 (19)	BRCA2 (19)	CCNE2 (24)	CCNE2 (19)	CCNE2 (18)	CCNE2 (18)
PIK3CG (16)	CCNE2 (18)	PIK3CG (23)	PIK3CG (17)	BRCA2 (17)	BRCA2 (17)
BRCA2 (13)	PIK3CG (17)	BRCA2 (20)	BRCA2 (16)	PIK3CG (16)	PIK3CG (16)
BIRC5 (11)	LAMB3 (13)	GTSE1 (12)	BIRC5 (10)	BIRC5 (11)	PIK3CD (12)
E2F2 (11)	EGFR (12)	LAMB3 (12)	E2F2 (10)	CRKL (11)	BIRC5 (11)
LAMB3 (10)	SKP2 (12)	MMP2 (12)	SKP2 (10)	STAT5B (11)	CRKL (11)
PIK3CD (10)	MMP9 (11)	E2R2 (12)	STAT5B (10)	E2F2 (10)	STAT5B (11)
HRAS (9)	MAPK8 (11)	LAMA4(11)	LAMB3 (9)	LAMB3 (10)	E2F2 (10)
SKP2 (9)	PGF (10)	PGF (11)	PGF (9)	PGF (10)	LAMB3 (10)
STAT5B (9)	PIK3CA (10)	SKP2 (11)	CRKL (8)	PIK3CD (10)	PGF (10)
			GTSP1 (8)	SKP2 (10)	SKP2 (10)
			GTSE1 (8)		
			PIK3CD (8)		

Table 3.2: Hub genes (genes with estimated degree $\geqslant 10$) in the estimated networks with $p = 315$ edges. MM stands for marginal mean imputation and KB for 'block' method with surrogates (3.8) and (3.9). Degree of the gene is given in the parentheses.

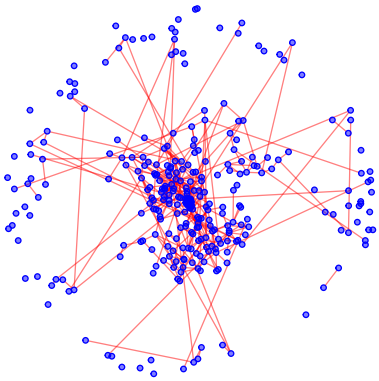


(a) MM

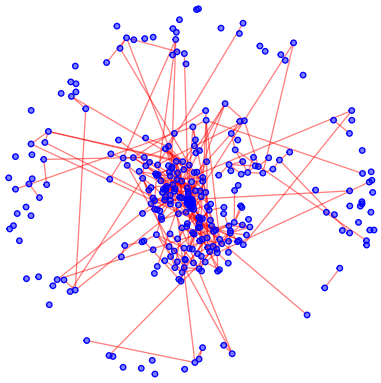
(b) KB, $\alpha = 0.25$



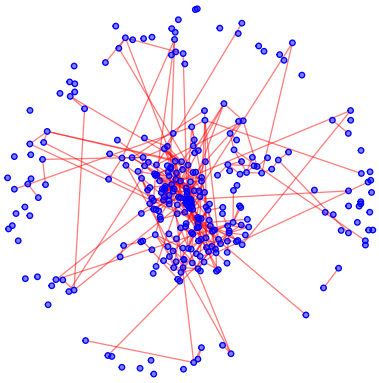
(c) KB, $\alpha = 0.5$



(d) KB, $\alpha = 1$



(e) KB, $\alpha = 1.5$



(f) KB, $\alpha = 2$

Figure 3.3: Topology of inferred networks with $p = 315$ edges. MM stands for marginal mean imputation and KB for 'block' method with surrogates (3.8) and (3.9). Layout of nodes is fixed across all graphs.

3.6 Discussion

In this chapter, we presented an extension of the regularized score matching framework in Chapter 2 for estimating high-dimensional non-Gaussian graphical models while accounting for potential missingness in the data. While the objective is non-convex, we show that if the search region is appropriately constrained, n scales at least with $d^2 \log p$, and λ is on the order of $\sqrt{\log p/n}$, the distance between the global optimum and the truth, as quantified by $\left|\left|\left|\hat{\mathbf{\Theta}} - \mathbf{\Theta}^*\right|\right|\right|_1$, is small in high probability. In our numerical experiments, we verify these theoretical scalings, showing that they are indeed sharp in the Gaussian setting. Furthermore, we show that our method outperforms traditional strategies for missing data, such as marginal imputation. We also returned to the RNAseq example from Chapter 2 and showed how, when missingness is properly accounted for, we can obtain different results from the naive treatment.

Unfortunately, we only managed to address the setting where observations are missing-completely-at-random, which may limit the method's applicability, as in many instances, observations are missing-at-random or missing-not-at-random. Thus, another area we could potentially work on would be extending our regularized score matching framework to accommodate data that is missing-at-random, as an initial first step in this direction. This may entail considering a Ising model-hybrid to explain the missing data mechanism.

Chapter 4

STATISTICAL SIGNIFICANCE IN HIGH-DIMENSIONAL LINEAR MIXED EFFECT MODELS

4.1 Introduction

Modern statistical problems are increasingly high-dimensional, with the number of covariates p potentially vastly exceeding sample size N. This is largely in part due to technological advances that have improved our ability to collect data efficiently. To illustrate, we are now able to measure the expression of many genes in a given specimen at little cost. On the other hand, it remains expensive to procure many replicates/species to experiment on, resulting in $N \ll p$.

Fortunately, significant progress has been made in developing rigorous statistical tools for tackling such problems. While earlier work largely targeted point estimation and/or variable selection, recent years have seen a number of proposals on how to also assign uncertainty, statistical significance and confidence in high-dimensional models. This is of great practical importance, particularly when interpretation of parameters and variables is of key priority.

Early attempts are highly varied in their approach. Stability selection was proposed in Meinshausen and Bühlmann [2010] as a generic method for controlling the expected number of false positive selections; a later modification was proposed in Shah and Samworth [2013]. Another approach that has been explored is sample splitting (i.e. first subsample used to screen, second subsample to perform inference), first implicitly put forward in Wasserman and Roeder [2009] and later made explicit and improved upon in Meinshausen et al. [2009]. Zhao et al. [2017] argue that sample-splitting can be entirely avoided and valid inference can be achieved upon refitting if there is a sufficient gap between 'strong' and 'weak' signal strength; however, this is fairly stringent requirement, as this assumption often does not

hold in practice. In the high-dimensional linear regression setting, a method for constructing confidence intervals without strict assumptions on the design matrix was presented in Meinshausen [2013]. In Juditsky et al. [2012], the authors develop necessary and sufficient conditions for bounding $\|\hat{\beta} - \beta^*\|_\infty$, with β^* being the true parameter value in a regression problem and $\hat{\beta}$ the lasso solution, with probability $1 - \alpha$, allowing one to construct very conservative confidence regions. Ning and Liu [2017] consider penalized M-estimators in greater generality, with the scope of application extending beyond linear models. From a different perspective, Lockhart et al. [2014], Tibshirani et al. [2014] and Lee et al. [2016] build a framework for conditional inference for high-dimensional linear models (i.e. conduct inference given some covariates have been selected).

In this chapter, we propose an (unconditional) inferential framework for high-dimensional linear mixed effect models, with the goal of being able to test null hypotheses of the form

$$H_{0,G} \ : \ \beta_j^* = 0 \text{ for all } j \in G$$

where $\beta^* \in \mathbb{R}^p$ is the vector of fixed effect regression coefficients. The set G may be any subset in $\{1, \ldots, p\}$. Of particular interest is when $G = \{j\}$, i.e. testing if a single regression coefficient β_j^* is 0. A related goal we set out to accomplish is the construction of confidence intervals for β_j^*, $j = 1, \ldots, p$. This problem, to the best of our knowledge, has not been previously addressed in existing literature, despite being of significant practical interest: observations are rarely independent, and linear mixed effect models are a natural extension of linear models for modeling data exhibiting group-structured dependence. Longitudinal data, highly prevalent in clinical studies (and others), are a natural application.

Our framework is inspired by a line of work where a high-dimensional estimator is corrected for bias, and the approximate limiting distribution of the resultant estimator is used to construct p-values and confidence intervals in high-dimensional problems. For example, in the high-dimensional linear regression setting, van de Geer et al. [2014b], and Javanmard and Montanari [2014] propose de-sparsifying the lasso (their formulations differ somewhat): they take the lasso estimator, which is biased, and by 'inverting' the corresponding Karuhn-Kush-

Tucker (KKT) optimality conditions, obtain an estimator that is asymptotically unbiased for β^* and normally distributed. By construction, the de-biased estimator can then be used to derive asymptotically valid confidence intervals of desired coverage and p-values.

Our proposed method bears strongest resemblance to Bühlmann [2013]. Developed for high-dimensional linear models, the framework proposed by Bühlmann [2013] is similar to those put forth by Zhang and Zhang [2014a], van de Geer et al. [2014b], and Javanmard and Montanari [2014], except it uses ridge estimation as a starting point. While the overall framework is similar, there are notable differences in the specifics on how to correct – or rather, approximately correct – for the bias in the ridge estimator, and how to compute an approximation of the limiting distribution of the de-biased estimator, so that we can construct p-values and confidence intervals for elements in β^*. As will be made evident later, these differences are the direct result of having to cope with dependencies induced by the random effects in the linear mixed effect model.

This chapter is organized as follows. The remainder of this section provides a brief overview of the subsequent notation. Section 4.2 makes explicit the form of the high-dimensional linear mixed effect model we are working with. In Section 4.3, we describe the details of our method, and additionally present theory – along with required assumptions – which validates it. Numerical experiments can be found in Section 4.4, followed by a practical application of the method to riboflavin production data in Section 4.5. We conclude with a brief discussion in Section 4.6.

4.2 High-dimensional model setup

In this section, we make explicit the model we would like to perform statistical inference on. Let $m = 1, \ldots, M$ represent group indices, and let $i = 1, \ldots, n_m$ index observations within group m. Write N for the total number of observations: $N = \sum_{m=1}^{M} n_m$. We assume that $n_m = n$ for all groups, implying that $N = nM$ (later theory only needs minor adjustments to accommodate imbalanced groups). For group $m \in \{1, \ldots, M\}$, we observe the response

vector $y_m \in \mathbb{R}^n$, which is generated according to

$$y_m = \mathbf{X}_m \beta^* + \mathbf{Z}_m \upsilon_m + \epsilon_m, \quad m = 1, \ldots, M \tag{4.1}$$

with

1. $\beta^* \in \mathbb{R}^p$ an unknown vector of fixed regression coefficients,

2. $\upsilon_m \in \mathbb{R}^q$, $m = 1, \ldots, M$ unknown vectors of group-specific random effects, with $\upsilon_m \underset{i.i.d.}{\sim} \mathcal{N}(0, \mathbf{\Psi}^*)$, $\mathbf{\Psi}^*$ a $q \times q$ positive definite covariance matrix.

3. errors $\epsilon_m \underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^{*2}\mathbf{I}_{n\times n})$, which are generated independently of $\upsilon_1, \ldots, \upsilon_M$, and

4. $\mathbf{X}_m$ and $\mathbf{Z}_m$ known design matrices of dimensions $n \times p$ and $n \times q$, respectively.

By construction, it is clear that β^* represents the shared effects while υ_m, $m = 1, \ldots, M$, represent group-specific deviations. The above expression can be written more compactly. Let $y = \begin{bmatrix} y_1^T & \ldots & y_M^T \end{bmatrix}^T$, $\upsilon = \begin{bmatrix} \upsilon_1^T & \ldots & \upsilon_M^T \end{bmatrix}^T$, and $\epsilon = \begin{bmatrix} \epsilon_1^T & \ldots & \epsilon_M^T \end{bmatrix}^T$, and defining stacked matrices $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1^T & \ldots & \mathbf{X}_M^T \end{bmatrix}^T$, and $\mathbf{Z} = \text{diag}(\mathbf{Z}_1, \ldots, \mathbf{Z}_M)$, then (4.1) may be written as

$$y = \mathbf{X}\beta^* + \mathbf{Z}\upsilon + \epsilon. \tag{4.2}$$

Marginalizing out the random effects yields

$$y \sim \mathcal{N}(\mathbf{X}\beta^*, \mathbf{V}(\sigma^{*2}, \mathbf{\Psi}^*)) \quad \text{with } \mathbf{V}(\sigma^2, \mathbf{\Psi}) = \sigma^2 \mathbf{I}_{N\times N} + \mathbf{Z}\mathbf{\Psi}^{(B)}\mathbf{Z}^T. \tag{4.3}$$

with $\mathbf{\Psi}^{(B)} = \mathbf{I}_{M\times M} \otimes \mathbf{\Psi}$. This implies that $\mathbf{V}(\sigma^{*2}, \mathbf{\Psi}^*)$ is block-diagonal and observations belonging to different groups are independent. Thus, the inclusion of random effects only induces dependencies between observations belonging to the same group. We will be primarily working with the marginal form (4.3) in subsequent sections.

Before we proceed with introducing our method for constructing confidence intervals and p-values for this model, we first write out a general set of assumptions on the high-dimensional linear mixed effect model (4.3):

1. *What is implied when model is high-dimensional*: We allow p, the number of fixed regression coefficients, to be possibly much larger than N. On the other hand, q, the number of random effect variables, is assumed to be of constant order, or at least smaller than n.

2. *Sparsity of* β^*: We assume β^* is sparse in the sense that the majority of its elements are 0: a clearer specification on the level of sparsity required is detailed in subsequent section.

3. *Structure of* $\mathbf{\Psi}^*$: We primarily consider the scenario of $\mathbf{\Psi}^* = \tau^{*2}\mathbf{I}_{q\times q}$. Readers, however, should note that our method, and corresponding theoretical results, can be readily extended to accommodate the more general scenario of $\mathbf{\Psi}^* = \mathbf{D}^*$ where $\mathbf{D}^*$ is a diagonal $q \times q$ matrix.

4. *Standardization of design matrices*: In the sequel, we assume that the design matrices $\mathbf{X}$ and $\mathbf{Z}$ are *fixed* and standardized such that $\|x_j\|_2^2 = N$ for $j \in \{1, \ldots, p\}$ and $\|z_j\|_2^2 = n$ for $j \in \{1, \ldots, q\}$.

4.3 A ridge-based inferential framework

As mentioned previously, we are interested in testing null hypotheses of the form,

$$H_{0,G} : \beta_j^* = 0 \text{ for all } j \in G.$$

Moreover, we would also to construct confidence intervals for β_j^*. In this section, we formally introduce our inferential framework. We first describe the de-biased ridge estimator which makes up its foundation, and how it can be used to accomplish these tasks. This section of the chapter bears strong similarity to portions in Bühlmann [2013]. We then detail how to assemble the necessary components needed to construct this de-biased ridge estimator and approximate its limiting distribution. Theoretical justification of our approach is provided along the way.

4.3.1 A de-biased ridge estimator

As in Bühlmann [2013], the 'naive' ridge estimator forms the foundation of our approach. The 'naive' ridge estimator is the minimizer of the following objective,

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \|y - \mathbf{X}\beta\|_2^2/N + \lambda\|\beta\|_2^2. \tag{4.4}$$

The word 'naive' references the fact that we have ignored correlation resulting from random effects. The estimator has a simple closed form expression given by

$$\hat{\beta} = \frac{1}{N}\left(\hat{\mathbf{\Sigma}} + \lambda \mathbf{I}_{p\times p}\right)^{-1}\mathbf{X}^T\mathbf{Y}, \tag{4.5}$$

where $\hat{\mathbf{\Sigma}} = \mathbf{X}^T\mathbf{X}/N$. It is straightforward to show that the ridge estimator is normally distributed with covariance matrix, multiplied by a factor of N,

$$\mathbf{\Omega}^* = (\hat{\mathbf{\Sigma}} + \lambda \mathbf{I}_{p\times p})^{-1}\mathbf{X}^T\mathbf{V}(\sigma^{*2}, \tau^{*2})\mathbf{X}(\hat{\mathbf{\Sigma}} + \lambda \mathbf{I}_{p\times p})^{-1}/N. \tag{4.6}$$

As in Bühlmann [2013], we assume that the diagonal entries of $\mathbf{\Omega}^* = (\omega^*_{jk})$ satisfy

$$\omega^*_{\min} \equiv \min_{j\in\{1,\ldots,p\}} \omega^*_{jj} > 0. \tag{4.7}$$

Likewise, we do not require (4.7) to be bounded away from 0 as a function of N or p. This condition, in fact, is fairly mild; it is only violated under special kinds of design matrices. To illustrate, define $R \equiv \text{rank}(\mathbf{X})$ and write the singular value decomposition of $\mathbf{X}$ as

$$\mathbf{X} = \mathbf{Q}\mathbf{D}\mathbf{\Gamma}^T$$

with

$$\mathbf{Q} \in \mathbb{R}^{N\times N},\ \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_{N\times N},$$
$$\mathbf{D} \in \mathbb{R}^{N\times N},\ \mathbf{D} \text{ diagonal with entries } s_1, \ldots, s_R \text{ (i.e. singular values of } \mathbf{X}),$$
$$\mathbf{\Gamma} \in \mathbb{R}^{p\times N},\ \mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{I}_{N\times N}.$$

Lemma 4. *Conditional (4.7) holds if and only if*

$$\min_{j\in\{1,\ldots,p\}} \max_{k\in\{1,\ldots,N\}, s_k\neq 0} \mathbf{\Gamma}^2_{jk} > 0. \tag{4.8}$$

Proof. It is straightforward to show that $\mathbf{\Omega}^*$ can be lower bounded by

$$\mathbf{\Omega}^* \geq \nu_{\min}(\mathbf{V}(\sigma^{*2}, \tau^{*2})) \underbrace{(\hat{\mathbf{\Sigma}} + \lambda \mathbf{I}_{p\times p})^{-1} \hat{\mathbf{\Sigma}} (\hat{\mathbf{\Sigma}} + \lambda \mathbf{I}_{p\times p})^{-1}}_{\tilde{\mathbf{\Omega}}^*}$$

Since σ^{*2} is taken to be greater than 0, $\nu_{\min}(\mathbf{V}(\sigma^{*2}, \tau^{*2})) > 0$. Note that $\tilde{\mathbf{\Omega}}^*$ can alternatively be written as

$$\tilde{\mathbf{\Omega}}^* = \mathbf{\Gamma} \operatorname{diag}\left(\frac{s_1^2}{(s_1^2 + \lambda)^2}, \ldots, \frac{s_N^2}{(s_N^2 + \lambda)^2} \right) \mathbf{\Gamma}^T,$$

which, in turn, implies that

$$\tilde{\omega}^*_{\min} = \min_{j\in\{1,\ldots,p\}} \sum_{k=1}^{N} \frac{s_k^2}{(s_k^2 + \lambda)^2} \mathbf{\Gamma}_{jk}^2,$$

and the claim follows. □

Readers should make note that the parameter β^* is, in fact, not identifiable in model (4.3) when $p > N$. Indeed, having $p > N$ implies that $R \leqslant N < p$, which in turn implies that there exist different vectors $\theta \in \mathbb{R}^p$ such that $\mathbf{X}\beta^* = \mathbf{X}\theta$. Define for generic matrix $\mathbf{A}$, $\mathbf{P}_{\mathbf{A}} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-}\mathbf{A}^T$, i.e. the hat matrix. A natural parameter to consider, as noted in Shao and Deng [2012], is $\theta^* = \mathbf{P}_{\mathbf{X}^T}\beta^* = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-}\mathbf{X}\beta^* = \mathbf{\Gamma}\mathbf{\Gamma}^T\beta^*$, the projection of β^* onto the linear space generated by the rows of $\mathbf{X}$. As it turns out, under condition (4.7), see also (4.8), the ridge estimator $\hat{\beta}$ is a reasonable proxy for θ^* when λ is sufficiently small.

Proposition 4. *Suppose that $\lambda > 0$ and (4.7), or equivalently, (4.8), holds. Then the ridge estimator (4.5) satisfies*

$$\max_{j\in\{1,\ldots,p\}} \left| \mathbb{E}[\hat{\beta}_j] - \theta_j^* \right| \leqslant \lambda \|\theta^*\|_2 \nu_{\min,+}(\hat{\mathbf{\Sigma}})^{-1}$$

$$\min_{j\in\{1,\ldots,p\}} \operatorname{Var}[\hat{\beta}_j] \geqslant N\omega^*_{\min}$$

where $\nu_{\min,+}(\hat{\mathbf{\Sigma}})$ refers to the smallest non-zero eigenvalue of $\hat{\mathbf{\Sigma}}$.

Proof. The proof to Proposition 4 is provided in the Appendix. □

It follows from Proposition 4 that the bias in estimating θ^* with $\hat{\beta}$ is small when $\lambda > 0$ is sufficiently small. We explicitly quantify how small λ needs to be for estimation bias to be smaller than the standard error of $\hat{\beta}$.

Corollary 5. *Suppose that the ridge penalty parameter $\lambda > 0$ satisfies the following inequality:*

$$\lambda / \sqrt{\omega^*_{\min}} \leqslant \nu_{\min,+}(\hat{\mathbf{\Sigma}}) / (\sqrt{N} \|\theta^*\|_2).$$

Furthermore, assume that condition (4.7), or equivalently, (4.8) holds. Then we have

$$\max_{j \in \{1,\ldots,p\}} \left| \mathbb{E}[\hat{\beta}_j] - \theta^*_j \right| \leqslant \min_{j \in \{1,\ldots,p\}} \sqrt{\mathrm{Var}[\hat{\beta}_j]}$$

Our interest, however, lies in β^*, not θ^*. Thus, for $\hat{\beta}$ to be useful, we need to adjust $\hat{\beta}$ for the projection bias $B_j = \theta^*_j - \beta^*_j$. By definition of θ^*, one observes that

$$B_j = (\mathbf{P}_{\mathbf{X}^T} \beta^*)_j - \beta^*_j = (\mathbf{P}_{\mathbf{X}^T})_{jj} \beta^*_j - \beta^*_j + \sum_{k \neq j} (\mathbf{P}_{\mathbf{X}^T})_{jk} \beta^*_k, \tag{4.9}$$

which, under the null hypothesis $H_{0,j} : \beta^*_j = 0$, becomes,

$$B_{H_0,j} = \sum_{k \neq j} (\mathbf{P}_{\mathbf{X}^T})_{jk} \beta^*_k. \tag{4.10}$$

The quantity can be approximated by

$$\hat{B}_{H_0,j} = \sum_{k \neq j} (\mathbf{P}_{\mathbf{X}^T})_{jk} \hat{\beta}^{\text{init}}_k. \tag{4.11}$$

where $\hat{\beta}^{\text{init}}$ a consistent initial estimator of β^* (consistency occurs under additional assumptions) whose distribution may not be of tractable form – i.e. $\hat{\beta}^{\text{init}} - \beta^*$ is small in some norm. Consider the corrected ridge estimator $\hat{\beta}^{\text{corr}}_j$ for testing $H_{0,j}$:

$$\hat{\beta}^{\text{corr}}_j = \hat{\beta}_j - \hat{B}_{H_0,j} = \hat{\beta}_j - \sum_{k \neq j} (\mathbf{P}_{\mathbf{X}^T})_{jk} \hat{\beta}^{\text{init}}_k. \tag{4.12}$$

Proposition 5. *Assume model (4.3). Suppose that* $\min_{j\in\{1,\ldots,p\}} \omega^*_{\min} > 0$. *Without referencing any hypothesis, we can decompose* $\hat{\beta}^{corr}_j$ *as follows:*

$$\begin{aligned}
\hat{\beta}^{corr}_j &= W_j + \gamma_j \\
W_1, \ldots, W_p &\sim \mathcal{N}(0, \mathbf{\Omega}^*/N) \\
\gamma_j &= (\mathbf{P}_{\mathbf{X}^T})_{jj}\beta^*_j - \sum_{k\neq j}(\mathbf{P}_{\mathbf{X}^T})_{jk}(\hat{\beta}^{init}_k - \beta^*_k) + \delta_j(\lambda) \\
\delta_j &= \mathbb{E}[\hat{\beta}_j] - \theta^*_j.
\end{aligned}$$

Alternatively, we can write

$$\frac{\hat{\beta}^{corr}_j}{(\mathbf{P}_{\mathbf{X}^T})_{jj}} - \beta^*_j = \frac{W_j}{(\mathbf{P}_{\mathbf{X}^T})_{jj}} - \sum_{k\neq j}\frac{(\mathbf{P}_{\mathbf{X}^T})_{jk}}{(\mathbf{P}_{\mathbf{X}^T})_{jj}}(\hat{\beta}^{init}_k - \beta^*_k) + \frac{\delta_j}{(\mathbf{P}_{\mathbf{X}^T})_{jj}}. \tag{4.13}$$

The normalizing factors needed to bring the W_j to $N(0,1)$ scale are given by $\kappa_j = \kappa_j(N,p) = \sqrt{N/\omega^*_{jj}}$.

Theorem 6. *Suppose we choose the ridge penalty parameter* $\lambda > 0$ *such that*

$$\lambda/\sqrt{\omega^*_{\min}} = o(\nu_{\min\neq 0}(\hat{\mathbf{\Sigma}})/(\sqrt{N}\|\theta^*\|_2)), \quad (N,p\to\infty), \tag{4.14}$$

and assume that for our choice of $\hat{\beta}^{init}$, *there exist constants* $C_j = C_j(N,p)$ *such that*

$$\mathbb{P}\left[\bigcap_{j=1}^{p}\left\{\left|\kappa_j(N,p)\sum_{k\neq j}(\mathbf{P}_{\mathbf{X}^T})_{jk}(\hat{\beta}^{init}_k - \beta^*_k)\right| \leqslant C_j(N,p)\right\}\right] \to 1 \quad (N,p\to\infty). \tag{4.15}$$

Supposing that $H_{0,j}$ *is true, then for all* $w > 0$,

$$\limsup_{N,p\to\infty} \mathbb{P}\left[\left|\kappa_j\hat{\beta}^{corr}_j\right| > w\right] - \mathbb{P}[|\tilde{W}| + C_j > w] \leqslant 0. \tag{4.16}$$

where $\tilde{W} \sim N(0,1)$. *In addition, for any sequence of subsets* G_p, $G_p \subseteq \{1,\ldots,p\}$, *if* H_{0,G_p} *is true, then for any* $w > 0$,

$$\limsup_{N,p\to\infty} \mathbb{P}\left[\max_{j\in G_p}\ \left|\kappa_j\hat{\beta}^{corr}_j\right| > w\right] - \mathbb{P}[\max_{j\in G_p}\ (|\tilde{W}| + C_j) > w] \leqslant 0. \tag{4.17}$$

Proof. The proof is quite straightforward. It follows, from Proposition 4, that

$$\begin{aligned}\max_j \ \kappa_j|\delta_j| &= \max_j \ \kappa_j \left|\mathbb{E}[\hat{\beta}_j] - \theta_j^*\right| \\ &\leqslant \frac{\lambda\|\theta^*\|_2 \nu_{\min\neq 0}(\hat{\boldsymbol{\Sigma}})^{-1}}{N^{-1/2}\omega_{jj}^{*1/2}} \\ &\leqslant \frac{\lambda\|\theta^*\|_2 \nu_{\min\neq 0}(\hat{\boldsymbol{\Sigma}})^{-1}}{N^{-1/2}\omega_{\min}^{*1/2}},\end{aligned}$$

which, due to our choice of ridge penalty parameter $\lambda > 0$ (4.14), is $o(1)$, $N, p \to \infty$. Then the claim follows from Proposition 5 and the assumption given by (4.15). □

The specific scaling we require of N and p in relation to one another will be made more apparent in subsequent sections, as it is derived based on our theoretical needs. Based on the asymptotic distributions in Theorem 6, we can construct p-values for testing $H_{0,G}$, $G \subseteq \{1, \ldots, p\}$. For testing the individual null hypothesis $H_{0,j}$, we define the p-value for the two-sided alternative as

$$\varrho_j = 2(1 - \Phi((\kappa_j|\hat{\beta}_j^{\text{corr}}| - C_j)_+)) \tag{4.18}$$

where Φ is the distribution function of the standard normal. For testing the group null hypothesis $H_{0,G}$, $|G| > 1$, we define the p-value for its complement to be

$$\varrho_G = 1 - \mathbb{P}\left[\max_{j\in G}(\kappa_j|W_j| + C_j) \leqslant \kappa_j|\hat{\beta}_j^{\text{corr}}|\right], \tag{4.19}$$

where $W_1, \ldots, W_p$ are as in Proposition 5. From Theorem 6, we can derive the following corollary.

Corollary 6. *Under the conditions in Theorem 6, for any $\alpha \in (0, 1)$, the following statements hold:*

$$\begin{aligned}\limsup_{N,p\to\infty} \mathbb{P}[\varrho_j \leqslant \alpha] - \alpha \leqslant 0 \quad &\textit{if } H_{0,j} \textit{ is true} \\ \limsup_{N,p\to\infty} \mathbb{P}[\varrho_G \leqslant \alpha] - \alpha \leqslant 0 \quad &\textit{if } H_{0,G} \textit{ is true.}\end{aligned}$$

4.3.2 Consistent estimation of variance parameters

To be able to apply this de-biased ridge framework, we need to know, or at least approximate, $\hat{\sigma}^{*2}$ and $\hat{\tau}^{*2}$. It is highly unlikely that we know $\hat{\sigma}^{*2}$ and $\hat{\tau}^{*2}$, so we consider the less trivial scenario where they are unknown and we need to construct consistent estimators.

Let S denote the support of β^*, i.e. $S = \{j \ : \ \beta_j^* \neq 0\}$, and $d = |S|$, the cardinality of S. We employ a two-step approach.

1. Use Lasso [Tibshirani, 1996] with an appropriate choice of tuning parameter λ_L to identify an initial guess of the elements (i.e. indices) in S. The Lasso estimator $\hat{\beta}^L$ is the minimizer of the objective,

$$\hat{\beta}^L = \arg\min_{\beta \in \mathbb{R}^p} \|y - \mathbf{X}\beta\|_2^2/N + 2\lambda_L\|\beta\|_1.$$

   and we define $\hat{S} = \{j \ : \ \hat{\beta}_j^L \neq 0\}$ to be our guess for the support S. By properties of the Lasso, $|\hat{S}| \leqslant N$.

2. Based on the (potentially misspecified) random effects model,

$$y = \mathbf{X}_{\hat{S}}\beta^* + \mathbf{Z}b + \epsilon \tag{4.20}$$

   apply Henderson's Method III [Henderson, 1953] to form $\hat{\sigma}^2$ and $\hat{\tau}^2$. As elaborated in subsequent paragraphs, Henderson's Method III is particularly tractable theoretically and enables us to more easily study consistency in the scenario where (4.20) is actually misspecified, i.e., $|S \backslash \hat{S}| > 0$.

Readers may be unfamiliar with Henderson's methods as they have been largely supplanted by alternatives such as restricted maximum likelihood (REML) [Harville, 1977] for variance component estimation (the variance of the random effects are also known as variance components). Thus, we provide a brief overview of what Henderson's Method III entails. (Four approaches were proposed in the 1953 paper; of them, Method III has been shown to be most appropriate for mixed effect models [Searle, 1968]). Consider the low-dimensional

model (4.3) with $p < N$. To simplify notation in the following explanation, we momentarily define $\tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X} & \mathbf{Z} \end{bmatrix}$. By not distinguishing between fixed and random effects, the idea behind Henderson's methods is to match the differences in the reductions in the sum-of-squares between sub-models of (4.3) to its expected value, not unlike a method-of-moments approach. To elaborate, in fitting (4.3) to data y, the reduction in the sum of squares is

$$\mathcal{R}(\beta, \upsilon) = y^T \mathbf{P}_{\tilde{\mathbf{X}}} y. \tag{4.21}$$

Likewise, the decrease in the sum of squares due to fitting the reduced model $y = \mathbf{X}\beta + \epsilon$ is

$$\mathcal{R}(\beta) = y^T \mathbf{P}_{\mathbf{X}} y. \tag{4.22}$$

The expected difference in the reductions $\mathcal{R}(\upsilon|\beta) \equiv \mathcal{R}(\beta, \upsilon) - \mathcal{R}(\beta)$ is

$$\mathbb{E}[\mathcal{R}(\upsilon|\beta)] = \tau^{*2} \mathrm{tr}(\mathbf{Z}^T [\mathbf{I}_{N\times N} - \mathbf{P}_{\mathbf{X}}]\mathbf{Z}) + \sigma^{*2}[\mathrm{rank}(\tilde{\mathbf{X}}) - \mathrm{rank}(\mathbf{X})]. \tag{4.23}$$

Likewise,

$$\mathbb{E}[y^T y - R(\beta, \upsilon)] = \sigma^{*2}[N - \mathrm{rank}(\tilde{\mathbf{X}})] \tag{4.24}$$

Together, (4.23) and (4.24), when matching theoretical expectations to empirical averages, form a triangular system of linear equations, from which we derive $\hat{\sigma}^2$ and $\hat{\tau}^2$. To be explicit, we find

$$\hat{\sigma}^2 = \frac{y^T(\mathbf{I}_{N\times N} - \mathbf{P}_{\tilde{\mathbf{X}}})y}{N - \mathrm{rank}(\mathbf{X})}, \tag{4.25}$$

$$\hat{\tau}^2 = \frac{y^T(\mathbf{P}_{\tilde{\mathbf{X}}} - \mathbf{P}_{\mathbf{X}})y - \hat{\sigma}^2[\mathrm{rank}(\tilde{\mathbf{X}}) - \mathrm{rank}(\mathbf{X})]}{\mathrm{tr}(\mathbf{Z}^T(\mathbf{I}_{N\times N} - \mathbf{P}_{\mathbf{X}})\mathbf{Z})}. \tag{4.26}$$

It is straightforward to see that the $\hat{\sigma}^2$ and $\hat{\tau}^2$ generated from (4.25) and (4.26) are unbiased, presuming that the true model is $y = \mathbf{X}\beta + \mathbf{Z}\upsilon + \epsilon$. (For consistency, some additional assumptions are needed, which we will discuss later in this section).

Referring to our two-step procedure and high-dimensional setup, Step 1 identifies a candidate low-dimensional sub-model, which is then used in Step 2 to obtain variance component estimates. We do not require the candidate model to encompass the truth; however, ideally,

λ_L is chosen such that $\hat{S}$, from Step 1, should at least reliably capture the indices corresponding to the 'strong' signals in β^*: the idea being that failing to include 'weak' signals should only negligibly affect consistency of $\hat{\sigma}^2$ and $\hat{\tau}^2$ in Step 2. We now proceed to show that this two-step procedure yields consistent estimators $\hat{\sigma}^2$ and $\hat{\tau}^2$ (i.e. $|\hat{\sigma}^2 - \sigma^{*2}| = o_P(1)$ and $|\hat{\tau}^2 - \tau^{*2}| = o_P(1)$) in the setting where $N \to \infty$ (specifically, n is fixed, but $M \to \infty$) and $d^2 \log p/M = o(1)$, provided some additional technical assumptions hold. From here on, this will also be the same scaling assumed for Theorem 6, as well as Corollary 6. Define, for some $\xi > 1$, the cone

$$\mathcal{C}(\xi, S) = \{u \in \mathbb{R}^p \ : \ \|u_{S^c}\|_1 \leqslant \xi \|u_S\|_1\}. \tag{4.27}$$

Assumption 2. *For some constant $\xi > 1$,*

$$\zeta \equiv \inf \left\{ \frac{\|\hat{\boldsymbol{\Sigma}} u\|_\infty}{\|u_A\|_\infty} : u \in \mathcal{C}_-(\xi, S), |A \backslash S| \leqslant p \right\} \gtrsim 1 \tag{4.28}$$

with

$$\mathcal{C}_-(\xi, S) \equiv \{u : u \in \mathcal{C}(\xi, S), \ u_j \boldsymbol{\Sigma}_{j,\cdot} u \leqslant 0 \ \forall j \notin S\},$$

the sign-restricted version of (4.27).

The quantity ζ (4.28) in Assumption 2 is defined more generally in Ye and Zhang [2010], where it is referred to as a sign-restricted cone invertibility factor (SCIF). As we demonstrate in the proof of Lemma A.9 (provided in the Appendix), this quantity naturally appears when deriving an upper bound for $\|\hat{\beta}^L - \beta^*\|_\infty$. Lemma A.9 claims that if Assumption 2 is satisfied, and we choose λ_L in Step 1 according to

$$\lambda_L = \frac{(\xi + 1)}{(\xi - 1)} \sqrt{\frac{2(\sigma^{*2} + \tau^{*2} q n)(\log p - \log(\varepsilon/2))}{N}} \asymp \sqrt{\frac{\log p}{M}} = o(1), \tag{4.29}$$

(i.e. an 'appropriate' choice), with ξ as in Assumption 2, then

$$\|\hat{\beta}^L - \beta^*\|_\infty \leqslant 2\xi \lambda_L / \zeta(\xi + 1) = o(1)$$

with probability exceeding $1 - \varepsilon$, where $\varepsilon > 0$ can be made arbitrarily small. A direct implication is that if the lemma conditions are satisfied, $S \backslash \hat{S}$ only includes indices corresponding

to 'weak' signals in β^* of magnitude less than $4\xi\lambda_L/\zeta(\xi+1) = o(1)$ with close to certainty, which is part of what Step 1 sets out to achieve.

Assumption 3. *There exists an integer $N' \lesssim d$ such that for the same constant $\xi > 1$ as in Assumption 2,*

$$\frac{d\xi^2}{\kappa^2(\xi, S)} < \frac{N'}{\kappa_+(N', S)}, \tag{4.30}$$

where

$$\kappa(\xi, S) = \min\left\{ \frac{d^{1/2}\|\mathbf{X}u\|_2}{n^{1/2}\|u_S\|_1} \ : \ u \in \mathcal{C}(\xi, S), u \neq 0 \right\}, \tag{4.31}$$

and $\kappa_+(N', S)$ is given by

$$\kappa_+(N', S) = \max_{\mathcal{A}\cap S=\emptyset, |A|\leqslant N'} \nu_{\min}\left(\frac{\mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}}}{N}\right),$$

the sparse upper eigenvalue of models disjoint with S.

Assumption 3 is needed to control the number of false positive selections in $\hat{S}$ from Step 1. In particular, it is possible to show that under this assumption and our choice of λ_L from before (4.29), the total number of false selections in Step 1 is bounded by N', with probability exceeding $1-\varepsilon$ (Lemma A.10 in the Appendix).

Assumption 4. *For the same N' in Assumption 3,*

$$rank([\mathbf{I}_{N\times N} - \mathbf{P}_{\tilde{\mathbf{X}}}]\mathbf{Z}) = rank(\mathbf{Z}) = qM, \tag{4.32}$$

$$\mathbf{Z}^T[\mathbf{I}_{N\times N} - \mathbf{P}_{\tilde{\mathbf{X}}}]\mathbf{Z} \gtrsim qM, \tag{4.33}$$

and the qM singular values of $[\mathbf{I}_{N\times N} - \mathbf{P}_{\tilde{\mathbf{X}}}]\mathbf{Z}$, $s_1, \ldots, s_{qM}$, satisfy

$$\frac{\sum_{i=1}^{qM} s_i^4}{\left(\sum_{i=1}^{qM} s_i^2\right)^2} = o(1), \tag{4.34}$$

where $\tilde{\mathbf{X}}$, here, is formed from joining any N' columns in $\mathbf{X}$ with $\beta_j^* = 0$ to the d support columns in $\mathbf{X}$.

By (4.32) in Assumption 4, the fixed data matrix $\mathbf{Z}$ has full column rank, and no column vector of $\mathbf{Z}$ can be represented as a linear combination of the column vectors of any 'feasible' $\mathbf{X}_{\hat{S}}$, supposing we choose λ_L according to (4.29). After all, $N' + d$ is the upper bound on the number of selected fixed effects with probability exceeding $1-\varepsilon$ (Lemma A.10). Additionally, by (4.33), the sum of the squared perpendicular distances between each column vector in $\mathbf{Z}$ and its projection onto the linear subspace spanned by the column vectors of feasible $\mathbf{X}_{\hat{S}}$'s is at least on the order of qM (substantial, given there are qM columns in $\mathbf{Z}$). On the other hand, the latter half of Assumption 4 seems to require all columns of $(\mathbf{I}_{N\times N} - \mathbf{P}_{\tilde{\mathbf{X}}})\mathbf{Z}$ are 'close' to being linearly independent from one another and 'contribute equally' to its rank. In particular, note that (4.34) is satisfied if

$$c_1 < \frac{s_j}{s_k} < c_2 \quad \text{for } j \neq k \text{ and some constants } c_1, c_2 > 0 \tag{4.35}$$

It is thus clear that (4.32) and (4.33) imply that random effects must not be confounded from any 'feasible' set of fixed effects (from Step 1) while (4.34) implies that the random effects are are not confounded from one another. Analogous conditions were shown to be necessary to prove consistency of REML estimators in Jiang [1996].

Assumption 5. *For any j in S such that $|\beta_j^*| < 4\xi\lambda_L/\zeta(\xi+1)$, with λ_L defined as in (4.29),*

$$\|\mathbf{\Gamma}_{\tilde{\mathbf{X}}} x_j\|_\infty \asymp 1.$$

Here, $\mathbf{\Gamma}_{\tilde{\mathbf{X}}}\mathbf{D}_{\tilde{\mathbf{X}}}\mathbf{\Gamma}_{\tilde{\mathbf{X}}}^T$ is the eigendecomposition of $\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^-\tilde{\mathbf{X}}^T$ with $\tilde{\mathbf{X}} = \begin{bmatrix}\bar{\mathbf{X}} & \mathbf{Z}\end{bmatrix}$, where $\bar{\mathbf{X}}$ is formed by joining any N' columns in $\mathbf{X}$ with $\beta_j^* = 0$ to the $d-1$ support (excluding j) columns in $\mathbf{X}$. The N' referenced here is the same as in Assumptions 3 and 4.

Assumption 5 relates to the irrepresentability conditions needed for model selection consistency in Lasso – see, for example, Zhao and Yu [2006a] – and non-confounding between fixed effects. Essentially, it states that covariates corresponding to weak (but non-zero) signals in β^* (of which we cannot quantify a bound on the probability they are to be included in $\mathbf{X}_{\hat{S}}$) cannot be too strongly correlated to covariates in $\mathbf{X}_{\hat{S}}$ nor covariates associated with the random effects.

Theorem 7. *Consider $N, p \to \infty$ (n fixed, $M \to \infty$), and $d^2 q \log p/M = o(1)$. Suppose Assumptions 2 through 5 are satisfied and λ_L is chosen according to (4.29) with $\varepsilon \propto 1/p$. Then, $\hat{\sigma}^2$ and $\hat{\tau}^2$ are consistent for σ^{*2} and τ^{*2}, respectively, i.e.,*

$$|\hat{\sigma}^2 - \sigma^{*2}| = |\hat{\tau}^2 - \tau^{*2}| = o_P(1) \quad (N, p \to \infty) \tag{4.36}$$

Proof. The proof to Theorem 7 is provided in the Appendix. □

Thus, we have validated our two-step procedure under a set of fairly standard conditions. For practical applications, REML can work as a substitute for Henderson's Method III for Step 2. We opted for Henderson's Method III largely due to convenience: because the estimators are derived from a set of linear equations, i.e. linear in variance parameters, consistency is straightforward to prove, despite the potential omission of fixed effects in Step 1. A possible avenue for further exploration is developing and proving analogous theory for REML.

Because $|\hat{\sigma}^2 - \sigma^{*2}|$ and $|\hat{\tau}^2 - \tau^{*2}|$ are both $o_P(1)$, we can use $\hat{\sigma}^2$ and $\hat{\tau}^2$ as plug-in values for σ^{*2} and τ^{*2}, respectively. From there, we can form a consistent estimator of $\mathbf{\Omega}^*$ and normalizing constants κ_j.

4.3.3 An initial estimator for β^* and our choice of C_j

To form $\hat{\beta}^{\text{init}}$, we consider the ordinary least-squares (OLS) fit restricted to $\hat{S}$, i.e.,

$$\hat{\beta}^{\text{init}} = \arg \min_{\beta \in \mathbb{R}^p : \beta_{\hat{S}^c} = 0} \|y - \mathbf{X}\beta\|_2^2. \tag{4.37}$$

We proceed to demonstrate that the error $\hat{\beta}^{\text{init}} - \beta^*$ is $o(1)$ in ℓ_1 norm.

Assumption 6. *For the same N' as in Assumptions 3, 4, 5, the sparse lower eigenvalue for models containing S of cardinality smaller than $d + N'$ is constant and greater than 0,*

$$\kappa_-(N', S) = \max_{\mathcal{A} \supset S, |\mathcal{A} \setminus S| \leqslant N'} \nu_{\max} \left(\frac{\mathbf{X}_{\mathcal{A}}^T \mathbf{X}_{\mathcal{A}}}{N} \right) \gtrsim 1,$$

Assumption 6, in conjunction with previous assumptions and choice of λ_L (4.29), can be used to control the ℓ_1 norm of the estimation error $\hat{\beta}^{\text{init}} - \beta^*$.

Theorem 8. *Suppose Assumptions 2 through 6 hold. Under the same conditions as in Theorem 7,*

$$\|\hat{\beta}^{init} - \beta^*\|_1 \leqslant Cd\sqrt{\frac{q \log p}{M}} \tag{4.38}$$

for some universal constant $C > 0$ with probability converging to 1 as $N, p \to \infty$.

Proof. The proof to Theorem 8 is provided in the Appendix. □

Theorem 8 implies that we have the following crude bound, based on Hölder's inequality,

$$\begin{aligned}\left|\kappa_j \sum_{k \neq j} (\mathbf{P}_{\mathbf{X}^T})_{jk} (\hat{\beta}_k^{\text{init}} - \beta_k^*)\right| &\leqslant \kappa_j \max_{k \neq j} |(\mathbf{P}_{\mathbf{X}^T})_{jk}| \, \|\hat{\beta}^{\text{init}} - \beta^*\|_1 \\ &\leqslant \kappa_j \max_{k \neq j} |(\mathbf{P}_{\mathbf{X}^T})_{jk}| \, Cd\lambda_L. \end{aligned} \tag{4.39}$$

Corollary 7. *Suppose the conditions in Theorem 8 are satisfied, and that d, the sparsity of β^*, satisfies*

$$d \leqslant C^{-1} \left(\frac{M}{q \log p}\right)^{\eta},$$

with C as in Theorem 8 and $\eta \in (0, 1/2)$, then we can choose C_j to be

$$C_j = \max_{k \neq j} |\kappa_j (\mathbf{P}_{\mathbf{X}^T})_{jk}| \left(\frac{q \log p}{M}\right)^{1/2 - \eta}. \tag{4.40}$$

for condition (4.15) to be satisfied in Theorem 6.

Proof. This follows directly from the crude bound (4.39). □

4.4 Numerical experiments

4.4.1 A practical choice for λ_L

In practical applications, we run into the issue of not being able to set λ_L according to (4.29), as it involves knowing τ^* and σ^*. However, we can derive a (slightly ad-hoc) approximation of what λ_L should be. Upon closer examination of the proof of Lemma A.8, we can substitute the term $\sigma^{*2} + \tau^{*2} qn$ with $\nu_{\max}(\mathbf{V}(\sigma^*, \tau^*)) = \sigma^{*2} + \tau^{*2} \nu_{\max}(\mathbf{Z}^T \mathbf{Z})$. The latter can be approximated according to the following procedure, assuming that the ratio τ^*/σ^* is sufficiently large:

1. Apply scaled lasso [Sun and Zhang, 2012] to obtain an initial 'average' noise estimate. The solution to the scaled lasso problem is characterized by

$$(\hat{\beta}^{\text{scaled}}, \hat{\sigma}^{\text{scaled}}) \in \arg\min_{\beta,\sigma} \frac{\|y - \mathbf{X}\beta\|_2^2}{2N\sigma} + \frac{\sigma}{2} + \lambda_{\text{univ}}\|\beta\|_1 \tag{4.41}$$

   with $\lambda_{\text{univ}} = \sqrt{2\log p/N}$.

2. Take $\hat{\sigma}^{\text{scaled}}\lambda_{\text{univ}}$ and adjust it by a factor of

$$\sqrt{\frac{\nu_{\max}(\mathbf{Z}^T\mathbf{Z})}{\text{tr}(\mathbf{Z}^T\mathbf{Z})/N}} \tag{4.42}$$

   to form λ_L.

We provide a heuristic justification. Ignoring the finer details involved in the theory, for the scaled lasso, $\hat{\sigma}^{\text{scaled},2}$ serves as a good approximation for $\|\epsilon^*\|_2^2/N$, where we have defined $\epsilon^* = y - \mathbf{X}\beta^*$. In linear models, ϵ^* holds i.i.d. observations drawn from a $N(0, \sigma^{*2})$ distribution. By law of large numbers, $\|\epsilon^*\|_2^2/N$ converges to σ^{*2} for large N.

Under a heteroskedastic error model, with ϵ^* independent and $\epsilon_i^* \sim N(0, \sigma_i^{*2})$, we can match $\|\epsilon^*\|_2^2/N$ to its expectation, which is given by $\sum_{i=1}^N \sigma_i^{*2}/N$, so $\hat{\sigma}^{\text{scaled},2}$ can be used to approximate the 'average' noise level. If $\epsilon^* \sim N(0, \mathbf{V}(\sigma^*, \tau^*))$, then using a similar expectation matching argument, we can expect $\hat{\sigma}^{\text{scaled},2}$ to act as a surrogate for

$$\sigma^{*2} + \frac{\tau^{*2}\text{tr}(\mathbf{Z}\mathbf{Z}^T)}{N}, \tag{4.43}$$

which follows from the fact that $\|\mathbf{\Gamma}\epsilon^*\|_2 = \|\epsilon^*\|_2$ for any $N \times N$ orthogonal matrix $\mathbf{\Gamma}$. What we actually need is $\sigma^{*2} + \tau^{*2}\nu_{\max}(\mathbf{Z}^T\mathbf{Z})$. Then in the scenario where ratio τ^{*2}/σ^{*2} is sufficiently large, (4.42) should give us a choice of λ_L that is close to the desired one from (4.29). Our choice of λ_L is constructed according to the above procedure for all subsequent numerical experiments.

4.4.2 A look into p-values

Denote the 'unblocked' version of $\mathbf{Z}$ as $\mathbf{Z}_u$; i.e., $\mathbf{Z}_u$ is a $N \times q$ matrix formed by row-wise concatenating the M diagonal blocks in $\mathbf{Z}$. We generate data from model (4.1) according to following schemes:

(M1) For $p \in \{300, 600\}$, $q \in \{1, 2\}$, we construct $\begin{bmatrix}\mathbf{X} & \mathbf{Z}_u\end{bmatrix}$ from N i.i.d. realizations from a $\mathcal{N}(0, \mathbf{\Phi}^*)$ distribution with $\mathbf{\Phi}^* = \{\phi_{jk}\}$ a $(p+q) \times (p+q)$ matrix with $\phi^*_{jk} = 0.2^{|j-k|}$. $\mathbf{X}$ and $\mathbf{Z}$ (the 'blocked' version) are then normalized such that $\|x_j\|_2^2 = N$ and $\|z_j\|_2^2 = n$ for all j. For the fixed regression coefficient, we have

$$\beta = [\underbrace{b, \ldots, b}_{d \text{ times}}, 0, \ldots, 0]$$

with $b \in \{0.5, 1\}$ and $d \in \{5, 10\}$. The variance parameters σ^* and τ^* are set to 0.5 and 1 respectively.

(M2) Same as (M1) except with $\mathbf{\Phi}^* = \mathbf{I}_{(p+q)\times(p+q)}$.

The numerical experiment are setup similarly to those in Bühlmann [2013] and Schelldorfer et al. [2011]. We set the ridge penalty parameter λ to $1/N$ for all experiments. Additionally, we set C_j according to Corollary 7 with $\eta = 0.005$.

We first consider null hypotheses of the form

$$H_{0,j} : \beta_j = 0. \tag{4.44}$$

We consider decision rules based on a significance level $\alpha = 0.05$, i.e.,

$$\text{reject } H_{0,j} \text{ if } \underbrace{\varrho_j \leqslant 0.05}_{\text{Event } E_j}.$$

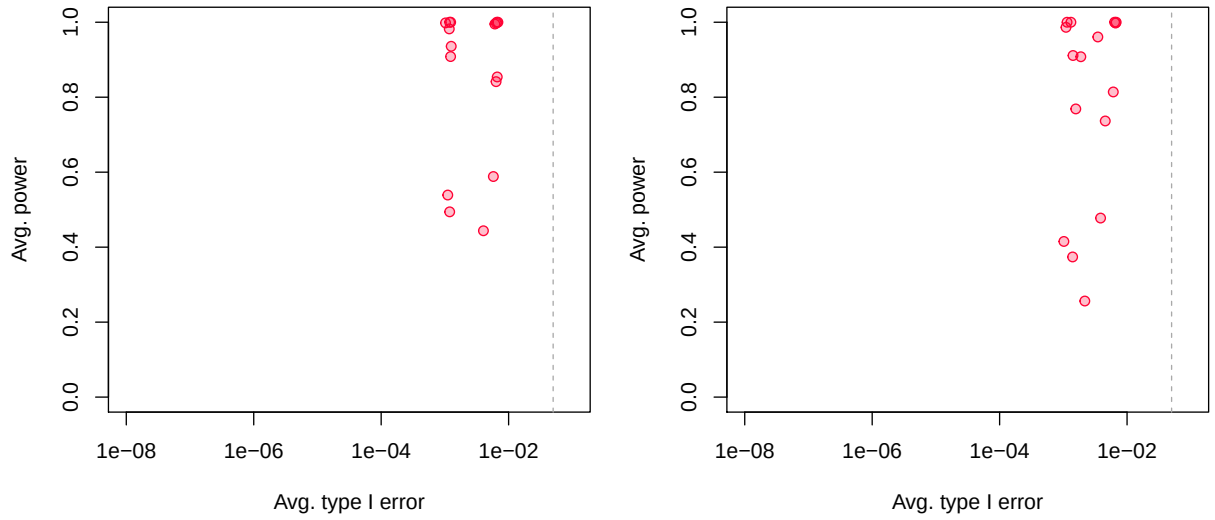
where ϱ_j is as defined in (4.18). Following Bühlmann [2013], we evaluate the performance of the tests based on the type I error, averaged over the non-support indices,

$$\text{Avg. type I error} = (p-d)^{-1} \sum_{j \in S^c} \hat{\mathbb{P}}(E_j) \tag{4.45}$$

and the power, averaged over the support indices,

$$\text{Avg. power} = d^{-1} \sum_{j \in S} \hat{\mathbb{P}}(E_j) \tag{4.46}$$

where $\hat{\mathbb{P}}$ denotes the empirical probability over the 1000 simulations we ran. Results are presented in Figure 4.1. Overall, it appears that Type I error is well-controlled for all attempted combinations of p, q, b and d for the two different models. Power is high in most scenarios, but also appears to vary with the aforementioned quantities, noticeably decreasing with b. However, this is to be expected.



(a) M1 (b) M2

Figure 4.1: Average power vs. average type I error for testing groups of coefficients under the two models for different combinations of p, q, b and d.

We also consider null hypotheses of the form

$$H_{0,G} : \beta_j = 0 \text{ for all } j \in G. \tag{4.47}$$

with G taken either to be $\{1, \ldots, 100\}$ ($G1$), or $\{101, \ldots, 200\}$ ($G2$). By construction, the hypothesis $H_{0,G1}$ should be accepted while $H_{0,G2}$ rejected. We consider decision rules based

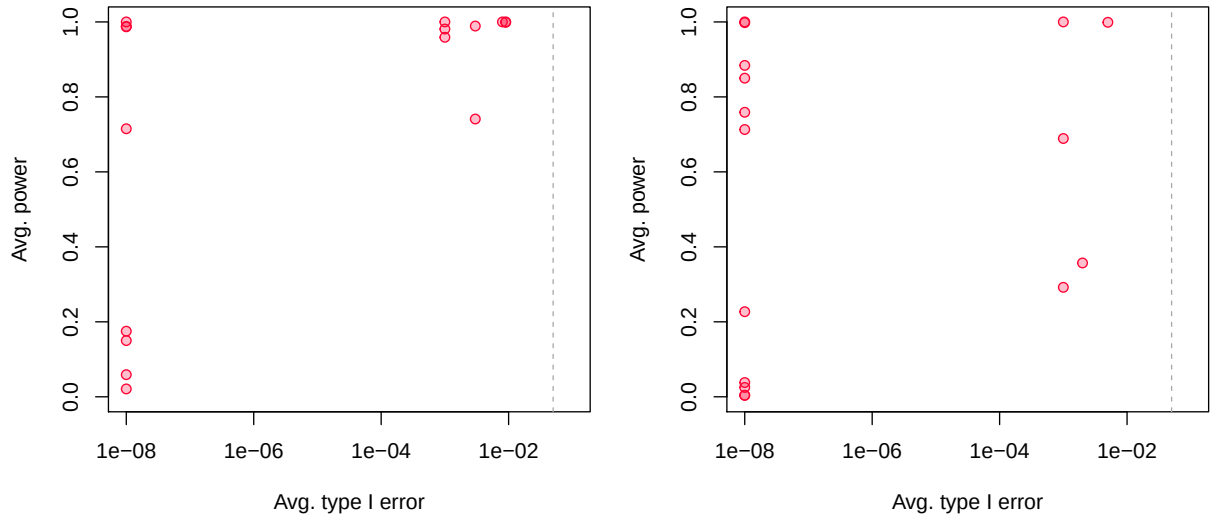
on a significance level $\alpha = 0.05$:

$$\text{reject } H_{0,G} \text{ if } \underbrace{\varrho_G \leqslant 0.05}_{\text{Event } E_G}$$

with ϱ_G is as defined in (4.19). To evaluate the performance of these tests, we consider type I and power, which can be represented by

$$\hat{\mathbb{P}}(E_{G2}) \quad \text{and} \quad \hat{\mathbb{P}}(E_{G1}),$$

respectively, where again, $\hat{\mathbb{P}}$ denotes the empirical probability over the 1000 simulations we ran. Figure 4.2 gives a visual representation of the results.



(a) M1 (b) M2

Figure 4.2: Average power vs. average type I error for testing groups of coefficients under the two models for different combinations of p, q, b and d.

4.4.3 Comparisons with existing methods

One question the reader may ask is if we can instead 'naively' apply inferential procedures for high-dimensional linear models and achieve similar performance. We briefly address this

via a short numerical experiment.

Consider Model (M1) from Section 4.4.2, with $p = 300$ and $q = 1$. Let

$$\beta^* = \begin{bmatrix} 0.05 & 2 & 4 & 3 & 0.1 & 0 & \dots & 0 \end{bmatrix}. \tag{4.48}$$

We compare our method against

1. *Bühlmann [2013]*: an analogue of our method developed for high-dimensional linear models.

2. *van de Geer et al. [2014b]*: one of the inferential approaches entailing de-sparsifying a Lasso estimator, also developed for high-dimensional linear models.

The differences are fairly evident when comparing confidence interval coverage. For any $\alpha \in (0, 1)$, define $\mathbb{Q}_\alpha[W_j]$ as the α-th quantile of the distribution of W_j. Under the conditions of Theorem 6, if the assumed model is correct, Proposition 5 suggests that confidence intervals of the form

$$\left[\frac{\hat{\beta}_j^{\text{corr}}}{(\mathbf{P}_{\mathbf{X}^T})_{jj}} - \frac{\mathbb{Q}_{1-\alpha/2}[W_j] + C_j}{(\mathbf{P}_{\mathbf{X}^T})_{jj}}, \frac{\hat{\beta}_j^{\text{corr}}}{(\mathbf{P}_{\mathbf{X}^T})_{jj}} + \frac{\mathbb{Q}_{1-\alpha/2}[W_j] + C_j}{(\mathbf{P}_{\mathbf{X}^T})_{jj}} \right]$$

should guarantee coverage of at least $(1-\alpha)\%$. Rather than setting C_j according to Corollary 7, we set them to be the same as the 'C_j-analogues's from Bühlmann [2013], to make the two methods comparable. Our choice of C_j are larger than theirs, so if anything, this ad-hoc decision provides Bühlmann [2013]'s method an unfair advantage. In Figure 4.3, we examine 95% confidence interval coverage for the three methods, based on the above modifications.

Overall, our method, which accounts for random effects, performs best at attaining the target guaranteed coverage across all β_j^*'s, compared to the methods proposed in Bühlmann [2013] and van de Geer et al. [2014b]. While Bühlmann [2013]'s method does come close, coverage falls short at 16 indices: minimum coverage achieved was 92.9% (with 1000 simulations, this is statistically significant from 0.95). At initial glance it appears that the lasso-based method from van de Geer et al. [2014b] performs quite well – even outperforming ours, a

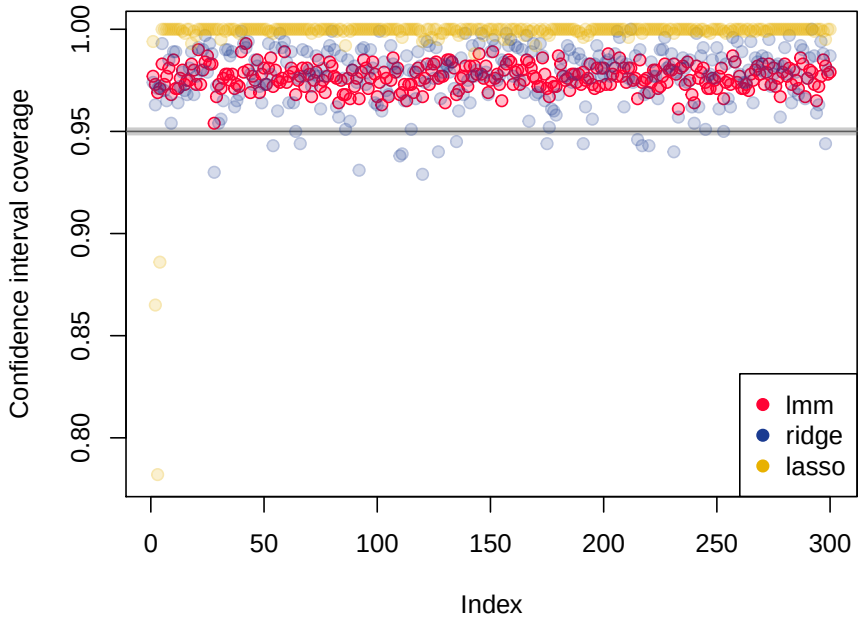


Figure 4.3: Confidence interval coverage for β_j^*, $j = 1, \ldots, p$; target coverage is 95% (with 1000 simulations, the standard deivation is $\approx 0.69\%$). Color to method legend: our method (——), Bühlmann [2013](——), and van de Geer et al. [2014b](——).

closer examination of the results reveals otherwise. Specifically, the lasso-based method does extremely poorly over the some of the support indices, as made evident in Table 4.1.

4.5 An application to riboflavin production data

In this section, we apply our proposed methodology to data on riboflavin (vitamin B_2) production by *Bacillus subtilis* made publicly available by Bühlmann et al. [2014] (original data was provided by DSM (Switzerland). In this dataset (referenced as *riboflavinGrouped* in Bühlmann et al. [2014]), we have $M = 28$ specimens measured at two to six time points, resulting in $N = 111$ observations in total. For each specimen at each time point, we record a single real valued response variable, the log-transformed riboflavin production rate, as well as the expression levels of $p = 4088$ genes. We are interested in identifying which gene is significantly correlated with riboflavin production.

	Our method	Bühlmann [2013]	van de Geer et al. [2014b]
β_1^*	0.977	0.974	0.994
β_2^*	0.973	0.963	0.865
β_3^*	0.969	0.971	0.782
β_4^*	0.971	0.972	0.886
β_5^*	0.983	0.993	1.000

Table 4.1: Confidence interval coverage for β_j^*, $j = 1, \ldots, 5$; target coverage is 95%.

To account for correlations induced by repeated measurements, a natural model to consider is the random intercept model, in which we assume that

$$y_m = \mathbf{X}_m \beta^* + v_m + \epsilon_m, \tag{4.49}$$

with v_m, $m = 1, \ldots, M$ i.i.d. with $v_m \sim N(0, \tau^{*2})$, and ϵ_m, $m = 1, \ldots, M$, independent with $\epsilon_m \sim N(0, \sigma^{*2}\mathbf{I}_{n_m \times n_m})$, and generated independently of $v_1, \ldots, v_m$. Readers may note that (4.49) can be represented by (4.1) with the $\mathbf{Z}_m$'s taken to be column vectors of 1s of lengths n_m.

We apply the our proposed framework and compute the marginal p-values for testing $\beta_j^* = 0$. Controlling the family-wise error rate (FWER) at 5%, via a simple Bonferroni correction, we find a single significant gene in riboflavin production: *YXLD-at*. This result matches previous results obtained by Javanmard and Montanari [2014] and Meinshausen et al. [2009] using an homogeneous dataset with $N = 71$ samples provided by the same source (*riboflavin* in Bühlmann et al. [2014]). Like us, Meinshausen et al. [2009] makes a single discovery, *YXLD-at*, while Javanmard and Montanari [2014] also labels *YXLE-at* as significant. We do note that the method of Bühlmann [2013], on the other hand, makes no discoveries.

4.6 Discussion

In this chapter, we developed a novel framework for constructing asymptotically valid p-values and confidence intervals for the fixed effects in high-dimensional linear mixed effect models, a problem that has not been addressed in existing literature. It entails de-biasing a 'naive' ridge estimator, whose asymptotic distribution we can approximately sufficiently well supposing that M scales at least with $d^2 q \log p$ and which in turn can be used to infer β^*. The results are promising: we show in simulations that level α tests developed using this method sufficiently control Type I error *at* level α for the array of high-dimensional setups considered. In addition, the conclusions we draw from the grouped riboflavin data using our method match those obtained by earlier papers on high-dimensional inference using the homogeneous dataset from the same source [Javanmard and Montanari, 2014, Meinshausen et al., 2009].

It is evident that several improvements can be made. For one, our proposed method for selecting the tuning parameter λ_L is admittedly ad-hoc and relies on the assumption that τ^{*2}/σ^{*2} is large, although it appears to work well in practice. Perhaps, an iterative scheme can be implemented where we repeatedly update λ_L based on the resultant estimates of σ^{*2} and τ^{*2}: this can be readily implemented in practice but may be difficult to validate theoretically. Additionally, our method requires q to be quite small (treated as a constant in theory). We can possibly work around this by taking $\mathbf{\Psi}^*$ to be a general diagonal matrix, i.e. $\mathbf{\Psi}^* = \text{diag}(\tau_1^{*2}, \ldots, \tau_q^{*2})$, and assuming that only a select number of τ_j^{*2}'s are nonzero, i.e. cardinality of $T \equiv \{j : \tau_j^{*2} \neq 0\}$ is small, less than n. Then, rather than just screen for fixed effects in Step 1 of the variance component estimation procedure, we screen both fixed and random effects by incorporating a double penalization scheme as in Wang et al. [2010] This way, in Step 2, both $|\hat{S}|$ and $|\hat{T}|$ are small, and we can apply Henderson's method III as before.

There are some details we did not discuss which should be mentioned, for completeness. One is multiple testing, which we omitted to avoid having to repeat content from Bühlmann

[2013]. Essentially, multiple testing can be handled the same way it was in Bühlmann [2013]. He proposes a procedure similar to the Westfall-Young procedure which strongly controls the familywise error rate. Define

$$F(a) = \mathbb{P}\left[\min_{1 \leqslant j \leqslant p} 2(1 - \kappa_j |W_j|) \leqslant a\right],$$

and corrected p-values,

$$\varrho_j^{\text{corr}} = F(\varrho_j + \zeta)$$

with $\zeta > 0$ an arbitrarily small number, set to 0.01 in Bühlmann [2013] in most practical applications. This multiple testing adjustment can directly be used in conjunction with our method for generating p-values for the individual hypothesis tests, and analogous theoretical guarantees on familywise error rate control should carry over.

Also clearly omitted is any discussion of power; however, this is due to the fact that we know that the ridge-based framework from Bühlmann [2013], which ours is significantly based on, is *not* optimal in this sense. Bühlmann [2013] shows that the detection rate may be larger than $1/\sqrt{N}$; on the other hand, Zhang and Zhang [2014a], who propose a de-biased lasso approach, establish, under certain conditions, that the detection limit is indeed in the $1/\sqrt{N}$ range. With that in mind, one line of future work is on how to build a lasso-based inferential framework for high-dimensional linear mixed effect models. (That said, in practice, we find for simulated homogeneous datasets, the ridge-based approach tends to outperform the de-biased lasso approach in controlling Type I error).

BIBLIOGRAPHY

Alekh Agarwal, Sahand Negahban, and Martin J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *Ann. Statist.*, 40(5):2452–2482, 2012. ISSN 0090-5364. URL http://dx.doi.org/10.1214/12-AOS1032.

Tommi J Ahonen, Jianwu Xie, Matthew J LeBaron, Jianqiong Zhu, Martti Nurmi, Kalle Alanen, Hallgeir Rui, and Marja T Nevalainen. Inhibition of transcription factor stat5 induces cell death of human prostate cancer cells. *Journal of Biological Chemistry*, 278 (29):27287–27292, 2003.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

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

Genevera I. Allen and Zhandong Liu. A log-linear graphical model for inferring genetic networks from high-throughput sequencing data. In *IEEE International Conference on Bioinformatics and Biomedicine*, pages 1–6. 2012.

Genevera I. Allen and Zhandong Liu. A local Poisson graphical model for inferring networks from sequencing data. *IEEE Trans. NanoBioscience*, 12(3):189–198, 2013.

Barry C. Arnold, Enrique Castillo, and José María Sarabia. *Conditional specification of statistical models*. Springer-Verlag, New York, 1999. ISBN 0-387-98761-4.

Barry C. Arnold, Enrique Castillo, José Maria Sarabia, and Laureano González-Vega. Mul-

tiple modes in densities with normal conditionals. *Statistics & Probability Letters*, 49(4): 355–363, 2000.

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

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

Rina Foygel Barber and Mathias Drton. High-dimensional Ising model selection with Bayesian information criteria. *Electron. J. Stat.*, 9:567–607, 2015. ISSN 1935-7524. URL http://dx.doi.org/10.1214/15-EJS1012.

S. Basu, A. Shojaie, and G. Michailidis. Network granger causality with inherent grouping structure. *Journal of Machine Learning Research*, 2014.

Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236, 1974.

A. Bhattacharyya. On some sets of sufficient conditions leading to the normal bivariate distribution. *Sankhya: The Indian Journal of Statistics*, 6(4):pp. 399–406, 1943.

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009. ISSN 0090-5364. URL http://dx.doi.org/10.1214/08-AOS620.

P.J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227, 2008.

P.J. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008b.

W. Bischoff and W. Fieger. Characterization of the multivariate normal distribution by conditional normal distributions. *Metrika*, 38(3-4):239–248, 1991.

Wolfgang Bischoff. On distribution whose conditional distributions are normal. A vector space approach. *Math. Methods Statist.*, 5(4):443–463, 1996.

Søren L. Buhl. On the existence of maximum likelihood estimators for graphical Gaussian models. *Scand. J. Statist.*, 20(3):263–270, 1993.

Peter Bühlmann. Boosting for high-dimensional linear models. *Ann. Statist.*, 34(2):559–583, 2006. ISSN 0090-5364. URL http://dx.doi.org/10.1214/009053606000000092.

Peter Bühlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4):1212–1242, 2013. ISSN 1350-7265. URL http://dx.doi.org.offcampus.lib.washington.edu/10.3150/12-BEJSP11.

Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data*. Springer, Heidelberg, 2011. ISBN 978-3-642-20191-2. URL http://dx.doi.org.offcampus.lib.washington.edu/10.1007/978-3-642-20192-9.

Peter Bühlmann, Markus Kalisch, and Lukas Meier. High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1:255–278, 2014.

T. T. Cai, W. Liu, and H. Zhou. Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *The Annals of Statistics*, page to appear, 2014.

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

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

Anthony Carbery and James Wright. Distributional and L^q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Math. Res. Lett.*, 8(3):233–248, 2001. ISSN 1073-2780. URL http://dx.doi.org.offcampus.lib.washington.edu/10.4310/MRL.2001.v8.n3.a1.

S. Carter, C. Brechbühler, M. Griffin, and A. Bond. Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, 20(14):2242–2250, 2004a.

Scott L Carter, Christian M Brechbühler, Michael Griffin, and Andrew T Bond. Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, 20(14):2242–2250, 2004b.

E. Castillo and J. Galambos. Conditional distributions and the bivariate normal distribution. *Metrika*, 36(3-4):209–214, 1989.

Jiahua Chen and Zehua Chen. Extended Bayesian information criterion for model selection with large model space. *Biometrika*, 95:759–771, 2008.

Shizhe Chen, Daniela Witten, and Ali Shojaie. Selection and estimation for mixed graphical models. *Biometrika (accepted)*, 2014. arXiv:1311.0085.

Michael Chichignoud, Johannes Lederer, and Martin Wainwright. Tuning lasso for sup-norm optimality. arXiv:1410.0247, 2014.

PT Chun, RJ McPherson, LC Marney, SZ Zangeneh, BA Parsons, Ali Shojaie, RE Synovec, and SE Juul. Metabolomic biomarkers of hypoxic ischemic encephalopathy in a nonhuman primate model. *Developmental Neuroscience (tentatively accepted)*, 2014.

Robert G. Cowell, A. Philip Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter. *Probabilistic networks and expert systems*. Statistics for Engineering and Information Science. Springer-Verlag, New York, 1999. ISBN 0-387-98767-3.

Patrick Danaher, Pei Wang, and Daniela M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 76 (2):373–397, 2014.

E.H. Davidson, D.R. McClay, and L. Hood. Regulatory gene networks and the properties of the developmental process. *Proceedings of the National Academy of Sciences*, 100:1475–1480, 2003.

A. Philip Dawid and Monica Musio. Estimation of spatial processes using local scoring rules. *AStA Adv. Stat. Anal*, 97(2):173–179, 2013.

Alexander Philip Dawid and Monica Musio. Theory and applications of proper scoring rules. *Metron*, 72(2):169–183, 2014.

Aaron Defazio and Tiberio S. Caetano. A convex formulation for learning scale-free networks via submodular relaxation. *Adv. Neural Inf. Process. Syst.*, pages 1250–1258, 2012.

Arthur P Dempster. Covariance selection. *Biometrics*, pages 157–175, 1972.

A. Dobra and M. West. Bayesian covariance selection. *Duke Statistics Discussion Papers*, 23, 2004.

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

Adrian Dobra and Alex Lenkoski. Copula Gaussian graphical models and their application to modeling functional disability data. *Ann. Appl. Stat.*, 5(2A):969–993, 2011.

M. Drton and M.D. Perlman. Multiple testing and error control in Gaussian graphical model selection. *Statistical Science*, 22(3):430–449, 2007a.

M. Drton and M.D. Perlman. A sinful approach to Gaussian graphical model selection. *Journal of Statistical Planning and Inference*, 138(4):1179–1200, 2008.

Mathias Drton. Discrete chain graph models. *Bernoulli*, 15(3):736–753, 2009a.

Mathias Drton. Likelihood ratio tests and singularities. *The Annals of Statistics*, 37(2): 979–1012, 2009b.

Mathias Drton and Caroline J. Klivans. A geometric interpretation of the characteristic polynomial of reflection arrangements. *Proc. Amer. Math. Soc.*, 138(8):2873–2887, 2010.

Mathias Drton and Marloes Maathuis. Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4(1):null, 2017. URL http://www.annualreviews.org/optdoi/abs/10.1146/annurev-statistics-060116-053803.

Mathias Drton and Michael D. Perlman. Multiple testing and error control in Gaussian graphical model selection. *Statist. Sci.*, 22(3):430–449, 2007b. ISSN 0883-4237. URL http://dx.doi.org.offcampus.lib.washington.edu/10.1214/088342307000000113.

Mathias Drton and Benjamin Williams. Quantifying the failure of bootstrap likelihood ratio tests. *Biometrika*, 98(4):919–934, 2011.

Mathias Drton and Han Xiao. Smoothness of Gaussian conditional independence models. In *Algebraic methods in statistics and probability II*, volume 516 of *Contemp. Math.*, pages 155–177. Amer. Math. Soc., Providence, RI, 2010a.

Mathias Drton and Han Xiao. Finiteness of small factor analysis models. *Ann. Inst. Statist. Math.*, 62(4):775–783, 2010b.

Mathias Drton and Han Xiao. Wald tests of singular hypotheses. *Bernoulli*, page to appear, 2014. arXiv:1304.6746.

Mathias Drton and Josephine Yu. On a parametrization of positive semidefinite matrices with zeros. *SIAM J. Matrix Anal. Appl.*, 31(5):2665–2680, 2010.

Mathias Drton, Bernd Sturmfels, and Seth Sullivant. *Lectures on algebraic statistics*, volume 39 of *Oberwolfach Seminars*. Birkhäuser Verlag, Basel, 2009. ISBN 978-3-7643-8904-8.

Mathias Drton, Rina Foygel, and Seth Sullivant. Global identifiability of linear structural equation models. *The Annals of Statistics*, 39(2):865–886, 2011a.

Mathias Drton, Rina Foygel, and Seth Sullivant. Global identifiability of linear structural equation models. *The Annals of Statistics*, 39(2):865–886, 2011b.

John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 272–279. ACM, 2008.

David Edwards. *Introduction to graphical modelling*. Springer-Verlag, New York, second edition, 2000. ISBN 0-387-95054-0. URL http://dx.doi.org.offcampus.lib.washington.edu/10.1007/978-1-4612-0493-0.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 2004. With discussion, and a rejoinder by the authors.

Werner Ehm and Tilmann Gneiting. Local proper scoring rules of order two. *The Annals of Statistics*, 40(1):609–637, 2012.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360, 2001. ISSN 0162-1459. URL http://dx.doi.org/10.1198/016214501753382273.

Jianqing Fan, Yang Feng, and Yichao Wu. Network exploration via the adaptive lasso and SCAD penalties. *Annals of Applied Statistics*, 3(2):521–541, 2009.

S Fan, Q Meng, K Auborn, T Carter, and EM Rosen. Brca1 and brca2 as molecular targets for phytochemicals indole-3-carbinol and genistein in breast and prostate cancer cells. *Brit. J. Cancer*, 94(3):407–426, 2006.

Bernd Fellinghauer, Peter Bühlmann, Martin Ryffel, Michael von Rhein, and Jan D. Reinhardt. Stable graphical model estimation with random forests for discrete, con-

tinuous, and mixed variables. *Comput. Statist. Data Anal.*, 64:132–152, 2013. URL http://dx.doi.org.offcampus.lib.washington.edu/10.1016/j.csda.2013.02.022.

Fernando A Ferrer, Lauri J Miller, Ramez I Andrawis, Scott H Kurtzman, Peter C Albertsen, Vincent P Laudone, and Donald L Kreutzer. Vascular endothelial growth factor (vegf) expression in human prostate cancer: in situ and in vitro expression of vegf by human prostate cancer cells. *J. Urol.*, 157(6):2329–2333, 1997.

Michael Finegold and Mathias Drton. Robust graphical modeling of gene networks using classical and alternative t-distributions. *Ann. Appl. Stat.*, 5(2A):1057–1080, 2011.

Michael Finegold and Mathias Drton. Robust Bayesian graphical modeling using Dirichlet t-distributions. *Bayesian Anal.*, pages 521–550, 2014.

Peter G. M. Forbes and Steffen Lauritzen. Linear estimating equations for exponential families with application to Gaussian linear concentration models. *Linear Algebra Appl.*, 473:261–283, 2015.

Christopher J. Fox, Andreas Käufl, and Mathias Drton. On the causal interpretation of acyclic mixed graphs under multivariate normality. *Linear Algebra Appl.*, page in press, 2014.

Rina Foygel and Mathias Drton. Extended Bayesian information criteria for Gaussian graphical models. *Adv. Neural Inf. Process. Syst.*, 23:2020–2028, 2010a.

Rina Foygel and Mathias Drton. Exact block-wise optimization in group lasso for linear regression. arXiv:1010.3320, 2010b.

Rina Foygel, Jan Draisma, and Mathias Drton. Half-trek criterion for generic identifiability of linear structural equation models. *The Annals of Statistics*, 40(3):1682–1713, 2012.

Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332, 2007.

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

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Applications of the lasso and grouped lasso to the estimation of sparse graphical models. Technical report, Stanford University, 2010.

N. Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303 (5659):799–805, 2004.

Ling-jie Fu and Bing Wang. Investigation of the hub genes and related mechanism in ovarian cancer via bioinformatics analysis. *Journal of Ovarian Research*, 6(1):92, 2013.

Xin Gao, Daniel Q. Pu, Yuehua Wu, and Hong Xu. Tuning parameter selection for penalized likelihood estimation of Gaussian graphical model. *Statist. Sinica*, 22(3):1123–1146, 2012. ISSN 1017-0405.

Simon A Gayther, Karen AF de Foy, Patricia Harrington, Paul Pharoah, William D Dunsmuir, Stephen M Edwards, Cheryl Gillett, Audrey Ardern-Jones, David P Dearnaley, Douglas F Easton, et al. The frequency of germ-line mutations in the breast cancer predisposition genes brca1 and brca2 in familial prostate cancer. *Cancer Res.*, 60(16): 4513–4518, 2000.

Andrew Gelman and Xiao-Li Meng. A note on bivariate distributions that are conditionally normal. *Amer. Statist.*, 45(2):125–126, 1991.

Elizabeth Gross, Mathias Drton, and Sonja Petrovic. Maximum likelihood degree of variance component models. *Electron. J. Statist.*, 6:993–1016, 2012.

Lei Gu, Paraskevi Vogiatzi, Martin Puhr, Ayush Dagvadorj, Jacqueline Lutz, Amy Ryder, Sankar Addya, Paolo Fortina, Carlton Cooper, Benjamin Leiby, Abhijit Dasgupta, Terry Hyslop, Lukas Bubendorf, Kalle Alanen, Tuomas Mirtti, and Marja T

Nevalainen. Stat5 promotes metastatic behavior of human prostate cancer cells in vitro and in vivo. *Endocrine-Related Cancer*, 17(2):481–493, 2010a. URL http://erc.endocrinology-journals.org/content/17/2/481.abstract.

Lei Gu, Paraskevi Vogiatzi, Martin Puhr, Ayush Dagvadorj, Jacqueline Lutz, Amy Ryder, Sankar Addya, Paolo Fortina, Carlton Cooper, Benjamin Leiby, et al. Stat5 promotes metastatic behavior of human prostate cancer cells in vitro and in vivo. *Endocr. Relat. Cancer*, 17(2):481–493, 2010b.

Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, 2011.

Magnús M Halldórsson and Jaikumar Radhakrishnan. Greed is good: Approximating independent sets in sparse and bounded-degree graphs. *Algorithmica*, 18(1):145–163, 1997.

Jing-Dong J Han, Nicolas Bertin, Tong Hao, Debra S Goldberg, Gabriel F Berriz, Lan V Zhang, Denis Dupuy, Albertha JM Walhout, Michael E Cusick, Frederick P Roth, et al. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature*, 430(6995):88–93, 2004.

Naftali Harris and Mathias Drton. PC algorithm for nonparanormal graphical models. *J. Mach. Learn. Res.*, 14:3365–3383, 2013.

David A. Harville. Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc.*, 72(358):320–340, 1977. ISSN 0162-1459. URL http://links.jstor.org.offcampus.lib.washington.edu/sici?sici=0162-1459(197706)72:358<320:MLATVC>2.0.CO;2-9&origin=MSN. With a comment by J. N. K. Rao and a reply by the author.

Trevor Hastie, Robert Tibshirani, Balasubramanian Narasimhan, and Gilbert Chu. *impute: impute: Imputation for microarray data*, 2016. R package version 1.46.0.

C. R. Henderson. Estimation of variance and covariance components. *Biometrics*, 9:226–252, 1953. ISSN 0006-341X. URL http://dx.doi.org.offcampus.lib.washington.edu/10.2307/3001853.

Holger Höfling and Robert John Tibshirani. Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *J. Mach. Learn. Res.*, 10:883–906, 2009.

Søren Højsgaard and Steffen L. Lauritzen. Graphical Gaussian models with edge and vertex symmetries. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 70(5):1005–1027, 2008.

Cho-Jui Hsieh, Arindam Banerjee, Inderjit S Dhillon, and Pradeep K Ravikumar. A divide-and-conquer method for sparse inverse covariance estimation. In *Adv. Neural Inf. Process. Syst.*, pages 2330–2338, 2012.

Jian Huang, Shuangge Ma, and Cun-Hui Zhang. Adaptive Lasso for sparse high-dimensional regression models. *Statist. Sinica*, 18(4):1603–1618, 2008. ISSN 1017-0405.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6:695–709, 2005.

Aapo Hyvärinen. Some extensions of score matching. *Comput. Statist. Data Anal.*, 51(5): 2499–2512, 2007.

Ali Jalali, Pradeep D Ravikumar, Vishvas Vasuki, and Sujay Sanghavi. On learning discrete graphical models using group-sparse regularization. In *AISTATS 2011*, pages 378–387, 2011.

E. Janofsky. *Exponential series approaches for nonparametric graphical models*. PhD thesis, The University of Chicago, 2015.

A. Jauhiainen, A. Shojaie, M. Kallitsis, and G. Michailidis. *RIPE: an R package for inferring regulatory networks from interventional and steady-state measurements*, 2013.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909, 2014. URL http://jmlr.org/papers/v15/javanmard14a.html.

Hawoong Jeong, Sean P Mason, A-L Barabási, and Zoltan N Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.

Jiming Jiang. REML estimation: asymptotic behavior and related topics. *Ann. Statist.*, 24(1):255–286, 1996. ISSN 0090-5364. URL http://dx.doi.org/10.1214/aos/1033066209.

Christopher C Johnson, Ali Jalali, and Pradeep D Ravikumar. High-dimensional sparse inverse covariance estimation using greedy methods. In *International Conference on Artificial Intelligence and Statistics*, pages 574–582, 2012.

B. Jones and M. West. Covariance decomposition in undirected Gaussian graphical models. *Biometrika*, 92(4):779–786, 2005.

Anatoli Juditsky, Fatma Kilinç Karzan, Arkadi Nemirovski, and Boris Polyak. Accuracy guaranties for ℓ_1 recovery of block-sparse signals. *Ann. Statist.*, 40(6):3077–3107, 2012. ISSN 0090-5364. URL http://dx.doi.org/10.1214/12-AOS1057.

Akash K. Kaushik, Shaiju K. Vareed, Sumanta Basu, Vasanta Putluri, Nagireddy Putluri, Katrin Panzitt, Christine A. Brennan, Arul M. Chinnaiyan, Ismael A. Vergara, Nicholas Erho, Nancy L. Weigel, Nicholas Mitsiades, Ali Shojaie, Ganesh Palapattu, George Michailidis, and Arun Sreekumar. Metabolomic profiling identifies biochemical pathways associated with castration-resistant prostate cancer. *Journal of Proteome Research*, 13(2): 1088–1100, 2014.

Kshitij Khare, Sang-Yun Oh, and Bala Rajaratnam. A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *J. Roy.*

Statist. Soc. Ser. B, 77(4):803–825, 2015. ISSN 1369-7412. URL http://dx.doi.org.offcampus.lib.washington.edu/10.1111/rssb.12088.

Diederik P Kingma and Yann LeCun. Regularized estimation of image statistics by score matching. In *Adv. Neural Inf. Process. Syst.*, pages 1126–1134, 2010.

Hirofumi Kishi, Mikio Igawa, Nobuyuki Kikuno, Tateki Yoshino, Shinji Urakami, and Hiroaki Shiina. Expression of the survivin gene in prostate cancer: correlation with clinicopathological characteristics, proliferative activity and apoptosis. *J. Urology*, 171(5):1855–1860, 2004.

Mladen Kolar and Eric P Xing. Consistent covariance selection from data with missing values. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 551–558, 2012.

Daphne Koller and Nir Friedman. *Probabilistic graphical models*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2009. ISBN 978-0-262-01319-2. Principles and techniques.

Urs Köster and Aapo Hyvärinen. A two-layer ICA-like model estimated by score matching. In *ICANN 2007*, pages 798–807. Springer, 2007.

Clifford Lam and Jianqing Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics*, 37(6B):4254–4278, 2009. URL http://dx.doi.org.offcampus.lib.washington.edu/10.1214/09-AOS720.

Johanna W Lampe, Sandi L Navarro, Meredith AJ Hullar, and Ali Shojaie. Inter-individual differences in response to dietary intervention: integrating omics platforms towards personalized dietary recommendations. *Proceedings of the Nutrition Society*, 72(2):207–218, 2013.

S.L. Lauritzen. *Graphical models*, volume 17. Oxford University Press, 1996.

Steffen L. Lauritzen. Some modern applications of graphical models. In *Highly structured stochastic systems*, volume 27 of *Oxford Statist. Sci. Ser.*, pages 13–44. Oxford Univ. Press, Oxford, 2003. With part A by Nanny Wermuth and part B by Julia Mortera.

Quoc V Le, Alexandre Karpenko, Jiquan Ngiam, and Andrew Y Ng. Ica with reconstruction cost for efficient overcomplete feature learning. In *Adv. Neural Inf. Process. Syst.*, pages 1017–1025, 2011.

Robert D Leclerc. Survival of the sparsest: robust gene networks are parsimonious. *Mol. Syst. Biol.*, 4(1):213, 2008.

Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the lasso. *Ann. Statist.*, 44(3):907–927, 2016. ISSN 0090-5364. URL http://dx.doi.org/10.1214/15-AOS1371.

Su-In Lee, Varun Ganapathi, and Daphne Koller. Efficient structure learning of Markov networks using ℓ_1-regularization. In B. Schölkopf, J.C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 817–824. MIT Press, 2007.

Shao Li, Lijiang Wu, and Zhongqi Zhang. Constructing biological networks through combined literature mining and microarray analysis: a lmma approach. *Bioinformatics*, 22(17):2143–2150, 2006.

Lina Lin, Mathias Drton, and Ali Shojaie. Estimation of high-dimensional graphical models using regularized score matching. *Electron. J. Stat.*, 10(1):806–854, 2016. ISSN 1935-7524. URL http://dx.doi.org.offcampus.lib.washington.edu/10.1214/16-EJS1126.

Roderick J. A. Little and Donald B. Rubin. *Statistical analysis with missing data*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York, 1987. ISBN 0-471-80254-9.

Roderick J. A. Little and Donald B. Rubin. *Statistical analysis with missing data*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2002. ISBN 0-471-18386-5. URL http://dx.doi.org/10.1002/9781119013563.

H. Liu, J. Lafferty, and L. Wasserman. The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.*, 10:2295–2328, 2009.

H. Liu, M. Xu, H. Gu, A. Gupta, J. Lafferty, and L. Wasserman. Forest density estimation. *J. Mach. Learn. Res.*, 12:907–951, 2011.

H. Liu, F. Han, M. Yuan, J. Lafferty, and L. Wasserman. The nonparanormal skeptic. *Proceedings of the 29th International Conference on Machine Learning*, 2012a.

Han Liu, Kathryn Roeder, and Larry Wasserman. Stability approach to regularization selection (stars) for high dimensional graphical models. In *Adv. Neural Inf. Process. Syst.*, pages 1432–1440, 2010.

Han Liu, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman. High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.*, 40(4):2293–2326, 2012b.

Han Liu, Fang Han, and Cun-hui Zhang. Transelliptical graphical models. In *Adv. Neural Inf. Process. Syst.*, pages 809–817, 2012c.

Qiang Liu and Alexander T Ihler. Learning scale free networks by reweighted l1 regularization. In *International Conference on Artificial Intelligence and Statistics*, pages 40–48, 2011.

Weidong Liu and Xi Luo. Fast and adaptive sparse precision matrix estimation in high dimensions. *J. Multivariate Anal.*, 135:153–162, 2015. ISSN 0047-259X. URL http://dx.doi.org/10.1016/j.jmva.2014.11.005.

Richard Lockhart, Jonathan Taylor, Ryan J. Tibshirani, and Robert Tibshirani. A significance test for the lasso. *Ann. Statist.*, 42(2):413–468, 2014. ISSN 0090-5364. URL http://dx.doi.org/10.1214/13-AOS1175.

P.-L. Loh and M. J. Wainwright. Support recovery without incoherence: A case for nonconvex regularization. *ArXiv e-prints*, December 2014.

Po-Ling Loh and Martin J. Wainwright. High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity. *Ann. Statist.*, 40(3):1637–1664, 2012. ISSN 0090-5364. URL http://dx.doi.org/10.1214/12-AOS1018.

Shan Lu, Juwon Lee, Monica Revelo, Xiaohong Wang, Shan Lu, and Zhongyun Dong. Smad3 is overexpressed in advanced human prostate cancer and necessary for progressive growth of prostate cancer cells in nude mice. *Clinical Cancer Research*, 13(19):5692–5702, 2007. ISSN 1078-0432. URL http://clincancerres.aacrjournals.org/content/13/19/5692.

Zhiyun Lu, Zhirong Yang, and Erkki Oja. Selecting β-divergence for nonnegative matrix factorization by score matching. In *Artificial Neural Networks and Machine Learning–ICANN 2012*, pages 419–426. Springer, 2012.

Jing Ma, Ali Shojaie, and George Michailidis. Network enrichment analysis with incomplete network information. *under review*, 2014a.

Jing Ma, Ali Shojaie, and George Michailidis. Network enrichment analysis with incomplete network information. *under review*, 2014b.

S. Ma, Q. Gong, and H.J. Bohnert. An arabidopsis gene network based on the graphical gaussian model. *Genome Research*, 17:1614–1625, 2007.

V. Mameli, M. Musio, and A. P. Dawid. Comparisons of Hyv\"arinen and pairwise estimators in two simple linear time series models. *ArXiv e-prints*, September 2014.

Daniel Marbach, James C Costello, Robert Kuffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Manolis Kellis, James J Collins, and Gustavo Stolovitzky. Wisdom of crowds for robust gene network inference. *Nature Methods*, 9(8):796–804, 2012.

K.V. Mardia, J. Kent, and J.M. Bibby. *Multivariate analysis*. Academic Press, 1979.

R. Mazumder and T. Hastie. Exact covariance thresholding into connected components for large-scale graphical lasso. *J. Mach. Learn. Res.*, 13:781–794, 2012.

Tyler McCormick, Hedwig Lee, Nina Cesare, and Ali Shojaie. Using twitter for demographic and social science research: Tools for data collection. *Sociological Methods and Research, (submitted for second review)*, 2013.

Lukas Meier, Sara van de Geer, and Peter Bühlmann. The group Lasso for logistic regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 70(1):53–71, 2008. ISSN 1369-7412. URL http://dx.doi.org/10.1111/j.1467-9868.2007.00627.x.

N. Meinshausen. Group-bound: confidence intervals for groups of variables in sparse high-dimensional regression without assumptions on the design. *ArXiv e-prints*, September 2013.

Nicolai Meinshausen. Relaxed Lasso. *Comput. Statist. Data Anal.*, 52(1):374–393, 2007. ISSN 0167-9473. URL http://dx.doi.org/10.1016/j.csda.2006.12.019.

Nicolai Meinshausen. A note on the Lasso for Gaussian graphical model selection. *Statist. Probab. Lett.*, 78(7):880–884, 2008.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 2006.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *J. Roy. Statist. Soc. Ser. B*, 72(4):417–473, 2010.

Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. p-values for high-dimensional regression. *J. Amer. Statist. Assoc.*, 104(488):1671–1681, 2009. ISSN 0162-1459. URL http://dx.doi.org/10.1198/jasa.2009.tm08647.

A Mitra, C Fisher, CS Foster, C Jameson, Y Barbachanno, J Bartlett, E Bancroft, R Doherty, Z Kote-Jarai, S Peock, et al. Prostate cancer in male brca1 and brca2 mutation carriers has a more aggressive phenotype. *Brit. J. Cancer*, 98(2):502–507, 2008.

Masashi Miyamura and Yutaka Kano. Robust Gaussian graphical modeling. *J. Multivariate Anal.*, 97(7):1525–1550, 2006. URL http://dx.doi.org.offcampus.lib.washington.edu/10.1016/j.jmva.2006.02.006.

Christian Moser, Petra Ruemmele, Sebastian Gehmert, Hedwig Schenk, Marina P Kreutz, Maria E Mycielska, Christina Hackl, Alexander Kroemer, Andreas A Schnitzbauer, Oliver Stoeltzing, et al. Stat5b as molecular target in pancreatic cancer?inhibition of tumor growth, angiogenesis, and metastases. *Neoplasia*, 14(10):915–IN12, 2012.

K.P. Murphy. *Dynamic Bayesian networks: representation, inference and learning*. PhD thesis, University Of California, 2002.

Yu. Nesterov. Gradient methods for minimizing composite objective function. CORE Discussion Papers 2007076, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2007. URL http://EconPapers.repec.org/RePEc:cor:louvco:2007076.

Xiao Ni, Daowen Zhang, and Hao Helen Zhang. Variable selection for semiparametric mixed models in longitudinal studies. *Biometrics*, 66(1):79–88, 2010. ISSN 0006-341X. URL http://dx.doi.org/10.1111/j.1541-0420.2009.01240.x.

Yang Ning and Han Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist.*, 45(1):158–195, 2017. ISSN 0090-5364. URL http://dx.doi.org/10.1214/16-AOS1448.

Masashi Okamoto. Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *Ann. Statist.*, 1:763–765, 1973. ISSN 0090-5364.

Matthew Parry, A. Philip Dawid, and Steffen Lauritzen. Proper local scoring rules. *The Annals of Statistics*, 40(1):561–592, 2012.

J. Peng, P. Wang, N. Zhou, and J. Zhu. Partial correlation estimation by joint sparse regression models. *J. Amer. Statist. Assoc.*, 104(486):735–746, 2009.

Deanna Petrochilos, Ali Shojaie, John Gennari, and Neil Abernethy. Using random walks to identify cancer-associated modules in expression data. *BioData Mining*, 6(1):17, 2013.

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

Pradeep Ravikumar, Martin J. Wainwright, and John D. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.*, 38(3):1287–1319, 2010. URL http://dx.doi.org.offcampus.lib.washington.edu/10.1214/09-AOS691.

Pradeep Ravikumar, Martin J. Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.*, 5:935–980, 2011.

Guilherme V. Rocha, Peng Zhao, and Bin Yu. A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (SPLICE). Technical report, University of California, Berkeley, 2008.

Saharon Rosset and Ji Zhu. Piecewise linear regularized solution paths. *Ann. Statist.*, 35(3):1012–1030, 2007. ISSN 0090-5364. URL http://dx.doi.org/10.1214/009053606000001370.

Volker Roth and Bernd Fischer. The group-lasso for generalized linear models: uniqueness of

solutions and efficient algorithms. In *ICML*, pages 848–855, 2008. ISBN 978-1-60558-205-4. URL http://doi.acm.org/10.1145/1390156.1390263.

Adam J. Rothman, Peter J. Bickel, Elizaveta Levina, and Ji Zhu. Sparse permutation invariant covariance estimation. *Electron. J. Stat.*, 2:494–515, 2008. URL http://dx.doi.org.offcampus.lib.washington.edu/10.1214/08-EJS176.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308:523–529, 2005.

José María Sarabia. The centered normal conditionals distribution. *Comm. Statist. Theory Methods*, 24(11):2889–2900, 1995.

K. Scheinberg and I. Rish. Sinco - a greedy coordinate ascent method for sparse inverse covariance selection problem. *Advances in Neural Information Processing Systems*, 2009. Preprint available at `http://www.optimizationonline. org/DB HTML/2009/07/2359.html`.

K. Scheinberg, S. Ma, and D. Goldfarb. Sparse inverse covariance selection via alternating linearization methods. *Advances in Neural Information Processing Systems*, 2010.

Katya Scheinberg and Irina Rish. Learning sparse Gaussian Markov networks using a greedy coordinate ascent approach. In *Machine Learning and Knowledge Discovery in Databases*, pages 196–212. Springer, 2010.

Jrg Schelldorfer, Peter Bhlmann, and Sara van de Geer. Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. *Scand. J. Stat.*, 38(2):197–214, 2011. ISSN 0303-6898. URL http://dx.doi.org/10.1111/j.1467-9469.2011.00740.x.

Gideon E. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978.

S. R. Searle. Another look at henderson's methods of estimating variance components. *Biometrics*, 24(4):749–787, 1968. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2528870.

Nafiseh Sedaghat, Takumi Saegusa, Timothy Randolph, and Ali Shojaie. Comparative study of computational methods for reconstructing genetic networks of cancer-related pathways. *Cancer Informatics*, 13(Suppl 2):55, 2014.

Rajen D. Shah and Richard J. Samworth. Variable selection with error control: another look at stability selection. *J. Roy. Statist. Soc. Ser. B*, 75(1):55–80, 2013. ISSN 1369-7412. URL http://dx.doi.org/10.1111/j.1467-9868.2011.01034.x.

Jun Shao and Xinwei Deng. Estimation in high-dimensional linear models with deterministic design matrices. *Ann. Statist.*, 40(2):812–831, 2012. ISSN 0090-5364. URL http://dx.doi.org/10.1214/12-AOS982.

Shahrokh F Shariat, Yair Lotan, Hossein Saboorian, Seyed M Khoddami, Claus G Roehrborn, Kevin M Slawin, and Raheela Ashfaq. Survivin expression is associated with features of biologically aggressive prostate carcinoma. *Cancer*, 100(4):751–757, 2004.

A. Shojaie. Link prediction using penalized multi-mode exponential random graph models. In *Proceedings of the 13th KDD Workshop on Learning and Mining with Graphs*. ACM, 2013.

A. Shojaie and G. Michailidis. Discovering graphical Granger causality using the truncating lasso penalty. *Bioinformatics*, 26(18):i517–i523, 2010.

A. Shojaie and N. Sedaghat. How similar are estimated networks of different cancer subtypes? In S. E. Ahmed, editor, *Big and Complex Data Analysis: Statistical Methodologies and Applications*. Springer, New York, 2016.

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *J. Comput. Graph. Statist.*, 22(2):231–245, 2013. ISSN 1061-8600. URL http://dx.doi.org.offcampus.lib.washington.edu/10.1080/10618600.2012.681250.

Nicolas Städler and Peter Bühlmann. Missing values: sparse inverse covariance estimation and an extension to sparse regression. *Stat. Comput.*, 22(1):219–235, 2012. ISSN 0960-3174. URL http://dx.doi.org/10.1007/s11222-010-9219-7.

Hokeun Sun and Hongzhe Li. Robust Gaussian graphical modeling via ℓ_1 penalization. *Biometrics*, 68(4):1197–1206, 2012.

Siqi Sun, Mladen Kolar, and Jinbo Xu. Learning structured densities via infinite dimensional exponential families. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2287–2295. Curran Associates, Inc., 2015a. URL http://papers.nips.cc/paper/6006-learning-structured-densities-via-infinite-dimensional-exponential-families.pdf.

Siqi Sun, Mladen Kolar, and Jinbo Xu. Learning structured densities via infinite dimensional exponential families. In *Advances in Neural Information Processing Systems*, pages 2287–2295, 2015b.

Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012. ISSN 0006-3444. URL http://dx.doi.org/10.1093/biomet/ass043.

Kean Ming Tan, Daniela Witten, and Ali Shojaie. The cluster graphical lasso for improved estimation of Gaussian graphical models. *Comput. Statist. Data Anal.*, 2014. provisionally accepted.

R. J. Tibshirani, J. Taylor, R. Lockhart, and R. Tibshirani. Exact Post-Selection Inference for Sequential Regression Procedures. *ArXiv e-prints*, January 2014.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

Robert Tibshirani, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan Taylor, and Ryan J. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 74(2):245–266, 2012. ISSN 1369-7412. URL http://dx.doi.org.offcampus.lib.washington.edu/10.1111/j.1467-9868.2011.01004.x.

Ryan J. Tibshirani. The lasso problem and uniqueness. *Electron. J. Stat.*, 7:1456–1490, 2013. ISSN 1935-7524. URL http://dx.doi.org/10.1214/13-EJS815.

Olga A Timofeeva, Xueping Zhang, Habtom W Ressom, Rency S Varghese, Bhaskar VS Kallakury, Kan Wang, Youngmi Ji, Amrita Cheema, Mira Jung, Milton L Brown, et al. Enhanced expression of sos1 is detected in prostate cancer epithelial cells from african-american men. *International journal of oncology*, 35(4):751, 2009.

Joel A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, 50(10):2231–2242, 2004. ISSN 0018-9448. URL http://dx.doi.org/10.1109/TIT.2004.834793.

Laufey Tryggvadóttir, Linda Vidarsdóttir, Tryggvi Thorgeirsson, Jon Gunnlaugur Jonasson, Elinborg Jona Ólafsdóttir, Gudridur Helga Ólafsdóttir, Thorunn Rafnar, Steinunn Thorlacius, Eirikur Jonsson, Jorunn Erla Eyfjord, et al. Prostate cancer progression and survival in brca2 mutation carriers. *Journal of the National Cancer Institute*, 99(12): 929–935, 2007.

Paul Tseng. Convergence of a block coordinate descent method for non-differentiable minimization. *J. Optim. Theory Appl.*, 109(3):475–494, 2001.

Sara van de Geer, Peter Bühlmann, and Shuheng Zhou. The adaptive and the thresholded

Lasso for potentially misspecified models (and a lower bound for the Lasso). *Electron. J. Stat.*, 5:688–749, 2011. ISSN 1935-7524. URL http://dx.doi.org/10.1214/11-EJS624.

Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014a.

Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3): 1166–1202, 2014b. ISSN 0090-5364. URL http://dx.doi.org/10.1214/14-AOS1221.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Comput.*, 23(7):1661–1674, 2011.

Daniel Vogel and Roland Fried. Elliptical graphical modelling. *Biometrika*, 98(4):935–951, 2011.

A. Voorman. *spacejam: R-package for graph estimation with joint additive models*, 2013.

A. Voorman. *lassoscore: R-package for inference in high dimensions with penalized score test*, 2014.

Arend Voorman, Ali Shojaie, and Daniela Witten. Graph estimation with joint additive models. *Biometrika*, 101(1):85–101, 2014.

Arie Voorman, Ali Shojaie, and Daniela Witten. Inference in high-dimensions with the penalized score test. *under review*, 2013.

Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory*, 55(5):2183–2202, 2009a. ISSN 0018-9448. URL http://dx.doi.org.offcampus.lib.washington.edu/10.1109/TIT.2009.2016018.

Martin J. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inform. Theory*, 55(12):5728–5741, 2009b. URL http://dx.doi.org.offcampus.lib.washington.edu/10.1109/TIT.2009.2032816.

Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Information Theory*, 55 (5):2183–2202, 2009c. URL http://dx.doi.org.offcampus.lib.washington.edu/10.1109/TIT.2009.2016018.

Martin J. Wainwright. Structured regularizers for high-dimensional problems: Statistical and computational issues. *Annu. Rev. Statist. Appl.*, 1:233–253, 2014.

Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

Hao Wang. Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis*, 7(2):771–790, 2012.

Hongbo Wang, Daqian Sun, Peng Ji, James Mohler, and Liang Zhu. An ar-skp2 pathway for proliferation of androgen-dependent prostate-cancer cells. *Journal of Cell Science*, 121 (15):2578–2587, 2008.

Sijian Wang, Peter Xuewin Song, and Ji Zhu. Doubly regularized reml for estimation and selection of fixed and random effects in linear mixed-effects models. 2010.

Zhiwei Wang, Daming Gao, Hidefumi Fukushima, Hiroyuki Inuzuka, Pengda Liu, Lixin Wan, Fazlul H Sarkar, and Wenyi Wei. Skp2: a novel potential therapeutic target for prostate cancer. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1825(1):11–17, 2012.

Larry Wasserman and Kathryn Roeder. High-dimensional variable selection. *Ann. Statist.*, 37 (5A):2178–2201, 2009. ISSN 0090-5364. URL http://dx.doi.org/10.1214/08-AOS646.

Peter H Westfall and S Stanley Young. *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons, 1993.

George Wilding, Eva Valverius, Cornelius Knabbe, and Edward P Gelmann. Role of transforming growth factor-α in human prostate cancer cell growth. *The Prostate*, 15(1):1–12, 1989.

A. Wille, P. Zimmermann, E. Vranová, A. Fürholz, O. Laule, S. Bleuler, L. Hennig, A. Prelíc, P.V. Rohr, L. Thiele, E. Zitzler, Wilhelm. Gruissem, and P. Bühlmann. Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. *Genome Biology*, 5:1–13, 2004.

D. M. Witten, J. Friedman, and N. Simon. New insights and faster computations for the graphical lasso. *J. Comp. Graph. Stat.*, 20(4):892–900, 2011.

Daniela M Witten, Ali Shojaie, and Fan Zhang. The cluster elastic net for high-dimensional regression with unknown variable grouping. *Technometrics*, 56(1):112–122, 2014.

Zhong Wu, HyungJun Cho, Garret M Hampton, and Dan Theodorescu. Cdc6 and cyclin e2 are pten-regulated genes associated with human prostate cancer metastasis. *Neoplasia*, 11 (1):66–76, 2009.

L. Xue and H. Zou. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics*, 40(5):2541–2571, 2012.

Eunho Yang, Genevera Allen, Zhandong Liu, and Pradeep K Ravikumar. Graphical models via generalized linear models. In *Adv. Neural Inf. Process. Syst.*, pages 1358–1366, 2012.

Eunho Yang, Pradeep Ravikumar, Genevera I Allen, and Zhandong Liu. On graphical models via univariate exponential family distributions. arXiv:1301.4183, 2013.

Guang Yang, Gustavo Ayala, Angelo De Marzo, Weihua Tian, Anna Frolov, Thomas M Wheeler, Timothy C Thompson, and J Wade Harper. Elevated skp2 protein expression

in human prostate cancer association with loss of the cyclin-dependent kinase inhibitor p27 and pten and with reduced recurrence-free survival. *Clinical Cancer Research*, 8(11): 3419–3426, 2002.

Fei Ye and Cun-Hui Zhang. Rate minimaxity of the lasso and dantzig selector for the lq loss in lr balls. *J. Mach. Learn. Res.*, 11:3519–3540, December 2010. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1756006.1953043.

Suhong Yu, Xingtian Yang, Yewei Zhu, Fangwei Xie, Yusheng Lu, Ting Yu, Cuicui Yan, Jingwei Shao, Yu Gao, Fan Mo, et al. Systems pharmacology of mifepristone (ru486) reveals its 47 hub targets and network: Comprehensive analysis and pharmacological focus on fak-src-paxillin complex. *Scientific reports*, 5, 2015.

M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(10):19–35, 2007.

Ming Yuan. Efficient computation of ℓ_1 regularized estimates in Gaussian graphical models. *Journal of Computational and Graphical Statistics*, 17(4):809–826, 2008.

Ming Yuan. High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11:2261–2286, 2010.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B*, 68(1):49–67, 2006.

B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4:Article 17, 2005.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942, 2010. ISSN 0090-5364. URL http://dx.doi.org/10.1214/09-AOS729.

Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 76(1):217–242, 2014a. ISSN 1369-7412. URL http://dx.doi.org/10.1111/rssb.12026.

Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 76(1):217–242, 2014b. URL http://dx.doi.org.offcampus.lib.washington.edu/10.1111/rssb.12026.

Teng Zhang and Hui Zou. Sparse precision matrix estimation via lasso penalized D-trace loss. *Biometrika*, 101(1):103–120, 2014. ISSN 0006-3444. URL http://dx.doi.org/10.1093/biomet/ast059.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563, 2006a.

Peng Zhao and Bin Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7: 2541–2563, 2006b.

S. Zhao, A. Shojaie, and D. Witten. In Defense of the Indefensible: A Very Naive Approach to High-Dimensional Inference. *ArXiv e-prints*, May 2017.

Sihai Dave Zhao, T. Tony Cai, and Hongzhe Li. Direct estimation of differential networks. *Biometrika*, 101(2):253–268, 2014.

Tuo Zhao, Han Liu, Kathryn Roeder, John Lafferty, and Larry Wasserman. The `huge` package for high-dimensional undirected graph estimation in `R`. *J. Mach. Learn. Res.*, 13: 1059–1062, 2012.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

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

Hui Zou, Trevor Hastie, and Robert Tibshirani. On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35(5):2173–2192, 2007.

Appendix A

SUPPLEMENT TO CHAPTER 2

A.1 Technical Lemmas

Lemma A.1 (Ravikumar et al., 2011). *If* $\left(X_1, \ldots, X_p\right)$ *is a zero-mean random vector with covariance matrix* $\mathbf{\Sigma}^*$ *such that* $X_i/\sqrt{\Sigma^*_{ii}}$ *is sub-Gaussian with scale parameter* σ^**, then the sample covariance matrix* $\mathbf{W}$*, for* n *i.i.d. samples, satisfies the bound*

$$\mathbb{P}[|\mathbf{W}_{jk} - \mathbf{\Sigma}^*_{jk}| > \delta] \leqslant 4 \exp\left\{ -\frac{n\delta^2}{128(1+4\sigma^{*2})^2 \max\limits_{j=1,\ldots,p} (\mathbf{\Sigma}^*_{jj})^2} \right\} \tag{A.1}$$

for any fixed choice of two indices $1 \leqslant j, k \leqslant p$ *and for all* $\delta \in (0, 40 \max\limits_{j=1,\ldots,p} \mathbf{\Sigma}^*_{jj})$.

Lemma A.2 (Carbery and Wright, 2001). *Let* $\mathcal{X}$ *be a Banach space, and let* $f : \mathbb{R}^p \to \mathcal{X}$ *be a polynomial of degree at most* z*. Suppose* $0 < \zeta_1 \leqslant \zeta_2 < \infty$ *and* μ *is a log-concave probability measure on* $\mathbb{R}^p$*. Then*

$$\left(\int \|f(x)\|^{\zeta_2/z} d\mu(x) \right)^{1/\zeta_2} \leqslant L \frac{\max(\zeta_2, 1)}{\max(\zeta_1, 1)} \left(\int \|f(x)\|^{\zeta_1/z} d\mu(x) \right)^{1/\zeta_1}, \tag{A.2}$$

where $L > 0$ *is an absolute constant.*

Lemma A.3. *Consider a degree* z *polynomial* $f(X) = f(X_1, \ldots, X_p)$*, where* $X_1, \ldots, X_p$ *are possibly dependent random variables with log-concave joint distribution on* $\mathbb{R}^p$*. Let* $L > 0$ *be the constant from Lemma A.2. Then, for all* δ *such that*

$$K := \frac{2}{L} \left(\frac{\delta}{e\sqrt{\mathrm{Var}[f(X)]}} \right)^{1/z} \geqslant 2, \tag{A.3}$$

we have,

$$\mathbb{P}|f(X) - \mathbb{E}[f(X)]| > \delta] \leqslant \exp\left\{ -\frac{2}{L} \left(\frac{\delta}{\sqrt{\mathrm{Var}[f(X)]}} \right)^{1/z} \right\}. \tag{A.4}$$

Proof. Choosing $\zeta_1 = 2z$ and $\zeta_2 = Kz$ in Lemma A.2, we have

$$\mathbb{E}[|f(X) - E[f(X)]|^K]^{\frac{1}{K}} \leqslant \left(\frac{LK}{2}\right)^z \sqrt{\mathrm{Var}[f(X)]}.$$

Hence, by Markov's inequality, for any δ satisfying (A.3),

$$\begin{aligned}
\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| > \delta] &\leqslant \frac{\mathbb{E}[|f(X) - \mathbb{E}[f(X)]|^K]}{\delta^K} && \text{(A.5)}\\
&\leqslant \left[\left(\frac{LK}{2}\right)^z \frac{\sqrt{\mathrm{Var}[f(X)]}}{\delta}\right]^K && \text{(A.6)}\\
&= \exp\{-K\} && \text{(A.7)}\\
&= \exp\left\{-\frac{2}{L}\left(\frac{\delta}{\sqrt{\mathrm{Var}[f(X)]}}\right)^{\frac{1}{z}}\right\}, && \text{(A.8)}
\end{aligned}$$

and the proof is complete. □

A.2 Proofs for Section 2.4

A.2.1 Proof of Theorem 1

First, we note that claim (b) is an immediate consequence of claim (a). To show (a), we apply the primal-dual witness method (PDW) from Wainwright [2009a]. As explained in detail below, PDW entails construction of a pair $(\tilde{\theta}, \tilde{z})$, with $\tilde{\theta} \in \mathbb{R}^{p^2}$ and $\tilde{z} \in \partial\|\tilde{\theta}\|_1$, that satisfies the KKT optimality conditions from (2.45) and has the support of $\tilde{\theta}$ included in S. If the construction is successful then it ensures that the rSME problem admits a unique solution such that the rSME $\hat{\theta}$ is equal to $\tilde{\theta}$ and inherits all the properties the latter has by definition. These properties include the ℓ_∞ bound on estimation error in addition to the claim about the support.

Replacing $\mathbf{\Gamma}$ by $\mathbf{\Gamma}^*$ and γ by γ^* in the empirical (basic or non-negative) score matching loss recovers the population loss which, in the present exponential family context, is quadratic and minimized when $\theta = \theta^*$. (Recall that the score matching loss is consistent.) It follows that r_3 from (2.44) is zero as it is the gradient of the population loss. In block form, (2.45)

becomes

$$\begin{bmatrix} \mathbf{\Gamma}^*_{SS} & \mathbf{\Gamma}^*_{SS^c} \\ \mathbf{\Gamma}^*_{S^cS} & \mathbf{\Gamma}^*_{S^cS^c} \end{bmatrix} \begin{bmatrix} \hat{\theta}_S - \theta^*_S \\ \hat{\theta}_{S^c} - \theta^*_{S^c} \end{bmatrix} + \begin{bmatrix} \mathbf{R}_{1,SS} & \mathbf{R}_{1,SS^c} \\ \mathbf{R}_{1,S^cS} & \mathbf{R}_{1,S^cS^c} \end{bmatrix} \begin{bmatrix} \hat{\theta}_S \\ \hat{\theta}_{S^c} \end{bmatrix} + \begin{bmatrix} r_{2,S} \\ r_{2,S^c} \end{bmatrix} + \lambda \begin{bmatrix} \hat{z}_S \\ \hat{z}_{S^c} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \quad \text{(A.9)}$$

We construct the PDW pair $(\tilde{\theta}, \tilde{z})$ according to the following steps:

1. Take $\tilde{\theta}$ to be the unique solution to the support-restricted problem, that is,

$$\tilde{\theta} = \arg\min_{\theta_{S^c}=0} \frac{1}{2}\theta^T \mathbf{\Gamma}\theta - \gamma^T\theta + \lambda\|\theta\|_1. \quad \text{(A.10)}$$

2. Choose

$$\tilde{\omega}_S \in \partial\|\tilde{\theta}_S\|_1.$$

3. Solving (A.9), set

$$\tilde{\omega}_{S^c} = \frac{1}{\lambda}\Big[-\mathbf{\Gamma}^*_{S^cS}(\mathbf{\Gamma}^*_{SS})^{-1}\left(\mathbf{R}_{1,SS}\tilde{\theta}_S + r_{2,S}\right) + \mathbf{R}_{1,S^cS}\tilde{\theta}_S + r_{2,S^c} + \lambda\mathbf{\Gamma}^*_{S^cS}(\mathbf{\Gamma}^*_{SS})^{-1}\tilde{z}_S\Big]. \quad \text{(A.11)}$$

4. Check the *strict dual feasibility* condition that

$$\|\tilde{\omega}_{S^c}\|_\infty < 1. \quad \text{(A.12)}$$

By step (i), $\tilde{\theta}$ has support contained in S. By step (iii), $(\tilde{\theta}, \tilde{z})$ is guaranteed to fulfill the equations from (A.9). By step (ii), the S-coordinates of $\tilde{z}$ satisfy 'their part' of the subgradient condition. Thus, if the strict dual feasibility from step (iv) holds, then $(\tilde{\theta}, \tilde{z})$ satisfies the KKT conditions from (2.45). Having a strict inequality in (A.12) ensures that every solution to the original rSME problem has support contained in the true support S and since $\mathbf{\Gamma}^*_{SS}$ is assumed invertible, there is then only one solution [Wainwright, 2009a, Lemma 1]. The invertibility of $\mathbf{\Gamma}^*_{SS}$ is also what guarantees the uniqueness in step (i).

If the PDW construction is successful, that is, if the strict dual feasibility condition can be established, then we may conclude the rSME $\hat{\theta}$ possesses all the desired properties. Indeed, $\hat{\theta}$ equals $\tilde{\theta}$ which has these properties by construction.

Let $\tilde{\Delta} = \tilde{\theta} - \theta^*$, where $\tilde{\theta}$ is the solution to the support-restricted regularized score matching problem from (A.10). By definition, $\|\tilde{\Delta}\|_\infty = \|\tilde{\Delta}_S\|_\infty$. Furthermore, by step (iii) in the PDW construction,

$$\tilde{\omega}_{S^c} = \frac{1}{\lambda}\Big[\mathbf{\Gamma}^*_{S^cS}(\mathbf{\Gamma}^*_{SS})^{-1}(\mathbf{R}_{1,SS}(\theta^*_S + \Delta_S) + r_{2,S}) - \mathbf{R}_{1,S^cS}(\theta^*_S + \Delta_S) - r_{2,S^c}\Big] + \mathbf{\Gamma}^*_{S^cS}(\mathbf{\Gamma}^*_{SS})^{-1}\tilde{\omega}_S. \quad \text{(A.13)}$$

By Assumption 1, and the triangle inequality for the ℓ_∞ norm,

$$\begin{aligned}
\|\tilde{\omega}_{S^c}\|_\infty &\leqslant \frac{1}{\lambda}\Big[(1-\alpha)\left(\|\mathbf{R}_{1,SS}(\theta^*_S + \Delta_S)\|_\infty + \|r_{2,S}\|_\infty\right) \\
&\qquad + \|\mathbf{R}_{1,S^cS}(\theta^*_S + \Delta_S)\|_\infty + \|r_{2,S^c}\|_\infty\Big] + (1-\alpha) \\
&\leqslant \frac{(2-\alpha)}{\lambda}\Big[\|\mathbf{R}_{1,\cdot S}(\theta^*_S + \Delta_S)\|_\infty + \|r_2\|_\infty\Big] + (1-\alpha) \\
&= \frac{(2-\alpha)}{\lambda}\Big[\|\mathbf{R}_1\theta^* + \mathbf{R}_{1,\cdot S}\Delta_S\|_\infty + \|r_2\|_\infty\Big] + (1-\alpha) \\
&\leqslant \underbrace{\frac{(2-\alpha)}{\lambda}\|\mathbf{R}_1\theta^*\|_\infty}_{=a_1} + \underbrace{\frac{(2-\alpha)}{\lambda}|||\mathbf{R}_{1,\cdot S}|||_\infty\|\Delta_S\|_\infty}_{=a_2} + \underbrace{\frac{(2-\alpha)}{\lambda}\|r_2\|_\infty}_{=a_3} + (1-\alpha),
\end{aligned}$$

where the equality in the second to last line follows from the fact that $\theta^*_{S^c} = 0$.

We observe that

$$a_1 = \frac{(2-\alpha)}{\lambda} \times \|\mathbf{\Theta}^*_{\text{wide}}\text{vec}(\mathbf{R}_{1,\text{blocks}})\|_\infty \quad \text{(A.14)}$$

where

$$\mathbf{\Theta}^*_{\text{wide}} = \begin{bmatrix}
\theta_1^{*T} & 0 & \dots & \dots & 0 \\
0 & \theta_1^{*T} & 0 & \dots & \vdots \\
\vdots & 0 & \ddots & \ddots & \dots \\
\vdots & \vdots & \ddots & \theta_p^{*T} & 0 \\
\vdots & \vdots & \vdots & 0 & \theta_p^{*T}
\end{bmatrix}$$

is an $p^2 \times p^3$ matrix whose diagonal blocks are given by the rows of the the interaction matrix $\mathbf{\Theta}^*$, each row being replicated p times. Moreover, $\text{vec}(\mathbf{R}_{1,\text{blocks}})$ refers to the vectorization of

the p diagonal blocks of $\mathbf{R}_1$ that are each of size $p \times p$; recall Lemma 2. More precisely, if $\mathbf{R}_{1,1}, \ldots, \mathbf{R}_{1,p}$ are the diagonal blocks of $\mathbf{R}_1$, then $\text{vec}(\mathbf{R}_{1,\text{blocks}})$ is obtained by concatenating $\text{vec}(\mathbf{R}_{1,1}), \ldots, \text{vec}(\mathbf{R}_{1,p})$ in that order. Equation (A.14) is the only argument relying on the block-diagonality of $\mathbf{\Gamma}$ and $\mathbf{R}_1$.

From (A.14), we obtain that

$$a_1 \leqslant \frac{(2-\alpha)}{\lambda} |||\mathbf{\Theta}^*_{\text{wide}}|||_\infty \|\text{vec}(\mathbf{R}_1)\|_\infty < \frac{(2-\alpha)}{\lambda} |||\mathbf{\Theta}^*_{\text{wide}}|||_\infty \epsilon_1.$$

since we have assumed that $\|\text{vec}(\mathbf{R}_1)\|_\infty = \|\mathbf{R}_1\|_\infty < \epsilon_1$. By construction, $|||\mathbf{\Theta}^*_{\text{wide}}|||_\infty = |||\mathbf{\Theta}^*|||_\infty = c_{\mathbf{\Theta}^*}$. It follows, from our choice of λ that $a_1 < \alpha/3$.

By the assumption that $\|r_2\|_\infty < \epsilon_2$, we have

$$a_3 < \frac{(2-\alpha)}{\lambda} \epsilon_2 < \frac{\alpha}{3},$$

and it remains to similarly bound a_2. We treat $|||\mathbf{R}_{1,\cdot S}|||_\infty$ and $\|\tilde{\Delta}_S\|_\infty$ separately.

We note that the rows of $\mathbf{R}_{1,\cdot S}$ have at most d non-zero elements. It follows that $|||\mathbf{R}_{1,\cdot S}|||_\infty \leqslant d\|\mathbf{R}_1\|_\infty < d\epsilon_1 < \alpha/6c_{\mathbf{\Gamma}}$, where the last inequality holds by assumption. Since $\mathbf{\Gamma}_{SS}$ is assumed invertible, we have from the top block of equations in (A.9) that

$$\tilde{\Delta}_S = (\mathbf{\Gamma}_{SS})^{-1}(-\mathbf{R}_{1,SS}\theta^*_S - r_{2,S} - \lambda\tilde{\omega}_S).$$

Application of the triangle inequality yields

$$\begin{aligned}
\|\tilde{\Delta}_S\|_\infty &\leqslant |||(\mathbf{\Gamma}_{SS})^{-1}|||_\infty \Big[\|\mathbf{R}_{1,SS}\theta^*_S\|_\infty + \|r_{2,S}\|_\infty + \lambda \Big] \\
&< |||(\mathbf{\Gamma}_{SS})^{-1}|||_\infty \Big[|||\mathbf{\Theta}^*_{\text{wide}}|||_\infty \|\text{vec}(\mathbf{R}_1)\|_\infty + \|r_2\|_\infty + \lambda \Big] \\
&\leqslant |||(\mathbf{\Gamma}_{SS})^{-1}|||_\infty \times \frac{(6-\alpha)}{3(2-\alpha)} \lambda. \qquad \text{(A.15)}
\end{aligned}$$

Since $\|\mathbf{R}_1\|_\infty < \epsilon_1$, we have $|||\mathbf{R}_{1,SS}|||_\infty \leqslant d\epsilon_1 < 1/c_{\mathbf{\Gamma}^*}$. This implies that

$$|||(\mathbf{\Gamma}^*_{SS})^{-1}\mathbf{R}_{1,SS}|||_\infty \leqslant |||(\mathbf{\Gamma}^*_{SS})^{-1}|||_\infty |||\mathbf{R}_{1,SS}|||_\infty < 1,$$

which gives us the following bound in the error in the inverse in the matrix ℓ_∞ norm,

$$\begin{aligned}\left|\left|\left|(\mathbf{\Gamma}_{SS})^{-1} - (\mathbf{\Gamma}^*_{SS})^{-1}\right|\right|\right|_\infty &\leqslant \frac{\left|\left|\left|(\mathbf{\Gamma}^*_{SS})^{-1}\mathbf{R}_{1,SS}\right|\right|\right|_\infty}{1 - \left|\left|\left|(\mathbf{\Gamma}^*_{SS})^{-1}\mathbf{R}_{1,SS}\right|\right|\right|_\infty} \times \left|\left|\left|(\mathbf{\Gamma}^*_{SS})^{-1}\right|\right|\right|_\infty \\ &\leqslant \frac{\left|\left|\left|(\mathbf{\Gamma}^*_{SS})^{-1}\right|\right|\right|_\infty \left|\left|\left|\mathbf{R}_{1,SS}\right|\right|\right|_\infty}{1 - \left|\left|\left|(\mathbf{\Gamma}^*_{SS})^{-1}\right|\right|\right|_\infty \left|\left|\left|\mathbf{R}_{1,SS}\right|\right|\right|_\infty} \times \left|\left|\left|(\mathbf{\Gamma}^*_{SS})^{-1}\right|\right|\right|_\infty. \qquad \text{(A.16)}\end{aligned}$$

Application of the triangle inequality, along with our definition of $c_{\mathbf{\Gamma}^*} = \left|\left|\left|(\mathbf{\Gamma}^*_{SS})^{-1}\right|\right|\right|_\infty$, yields

$$\begin{aligned}\left|\left|\left|(\mathbf{\Gamma}_{SS})^{-1}\right|\right|\right|_\infty &\leqslant \left|\left|\left|(\mathbf{\Gamma}^*_{SS})^{-1}\right|\right|\right|_\infty + \left|\left|\left|(\mathbf{\Gamma}_{SS})^{-1} - (\mathbf{\Gamma}^*_{SS})^{-1}\right|\right|\right|_\infty \\ &= \left|\left|\left|(\mathbf{\Gamma}^*_{SS})^{-1}\right|\right|\right|_\infty \times \frac{1}{1 - \left|\left|\left|(\mathbf{\Gamma}^*_{SS})^{-1}\right|\right|\right|_\infty \left|\left|\left|\mathbf{R}_{1,SS}\right|\right|\right|_\infty} \\ &\leqslant \frac{c_{\mathbf{\Gamma}^*}}{1 - d c_{\mathbf{\Gamma}^*}\epsilon_1} \\ &\leqslant \frac{c_{\mathbf{\Gamma}^*}}{1 - \alpha/6}, \qquad \text{(A.17)}\end{aligned}$$

where the last inequality uses the assumption that $d\epsilon_1 \leqslant \alpha/6c_{\mathbf{\Gamma}^*}$. Substituting (A.17) into (A.15), it is straightforward to show that $a_2 < \alpha/3$. Therefore, $a_1 + a_2 + a_3 < \alpha$, which yields that $\|\tilde{\omega}_{S^c}\| < 1$.

Along the way we have also proven the second part of the claim. Indeed, from (A.15) and (A.17), we have

$$\|\tilde{\Delta}_S\|_\infty \leqslant \frac{c_{\mathbf{\Gamma}^*}}{1 - \alpha/6} \times \frac{(6-\alpha)}{3(2-\alpha)}\lambda = \frac{2c_{\mathbf{\Gamma}^*}\lambda}{2-\alpha}.$$

A.2.2 Proof of Corollary 1

We need to show that the conditions in Theorem 1, specifically those in (2.46), hold with the claimed probability. Since $r_2 = \gamma - \gamma^* = \text{vec}(\mathbf{I}_{p\times p}) - \text{vec}(\mathbf{I}_{p\times p}) = 0$, the second inequality in (2.46) can be trivially satisfied with any $\epsilon_2 > 0$. Thus, we only need to show that we can bound $\|\mathbf{R}_1\|_\infty$ by some suitable ϵ_1 with sufficiently large probability. To do so, we apply a Bernstein-type concentration inequality for the entries of W that is also used by Ravikumar et al. [2011]. Lemma A.1 below states the inequality, as given in their paper.

The matrix $\mathbf{R}_1$ features only entries in $\mathbf{W} - \mathbf{\Sigma}^*$. By taking a union bound over the p^2 entries of $\mathbf{W}$, plugging in our lower bound for n and observing that $\sigma^* = 1$ in the Gaussian

case, Lemma A.1 yields that

$$\mathbb{P}\left[\|\mathbf{R}_1\|_\infty \geqslant \sqrt{\frac{c^*(\log p^{\tau_1} + \log 4)}{n}}\right] \leqslant \exp\{-\log p^\tau + 2\log p\} = \frac{1}{p^{\tau-2}}.$$

In addition, each row in $|||\mathbf{R}_{\cdot S}|||_\infty$ features at most d entries from the matrix $\mathbf{W} - \mathbf{\Sigma}^*$. Hence, it follows from another union bound, and choosing n at least

$$c^* c_1^2 d^2 (\log p^\tau + \log 4)$$

where c^* and c_1 are defined in the corollary statement, that

$$\mathbb{P}\left[|||\mathbf{R}_{\cdot S}|||_\infty > \frac{1}{c_1}\right] \leqslant \frac{1}{p^{\tau-2}}.$$

Thus, applying Theorem 1 with

$$\epsilon_1 = \sqrt{\frac{c^*(\log p^\tau + \log 4)}{n}}$$

shows that our choices for λ and n give the high probability statement in Corollary 1.

When looking back at the proof of Theorem 1, we see that as a consequence of having $r_2 = 0$, we need only be concerned with bounding terms a_1 and a_2. We may thus bound a_1 and a_2 each by $\alpha/2$ instead of $\alpha/3$ and ignore the a_3 term entirely, as it is 0. This leads to having $c_1 = (\alpha c_{\mathbf{\Gamma}^*}/4)^{-1}$, as opposed to the expected $(\alpha c_{\mathbf{\Gamma}^*}/6)^{-1}$.

A.2.3 Proof of Corollary 2

We proceed as for the proof of Corollary 1 and use concentration results to satisfy the bounds from (2.46) in Theorem 1. However, we now bound $\|\mathbf{R}_1\|_\infty$ and $\|r_2\|_\infty$ using concentration inequalities for general log-concave measures (any truncated multivariate normal density is log-concave).

Let $X^{(i)} = (X_{i1}, \ldots, X_{ip})$ be i.i.d. according to $N(0, (\mathbf{K}^*)^{-1})$ with truncation to $\mathbb{R}_+^p$. Take

$$\epsilon_1 = \frac{\left[\left(\frac{L}{2}\right)(\log p^\tau + \log 2)\right]^4}{\sqrt{n}} \sqrt{\max_{j,k,l} \operatorname{Var}[X_j^2 X_k X_l]}, \tag{A.18}$$

$$\epsilon_2 = \frac{\left[\left(\frac{L}{2}\right)(\log p^\tau + \log 2)\right]^2}{\sqrt{n}} \sqrt{\max_{j,k} \operatorname{Var}[X_j X_k]}. \tag{A.19}$$

We now want to see if we can apply Lemma A.3 with $\delta = \epsilon_1$ from (A.18) and $\delta = \epsilon_2$ from (A.19). It thus needs to be checked that condition (A.3) holds in these two cases. Indeed, the condition holds as long as

$$p \geqslant \exp\left\{\frac{2\sqrt{e} - \log 2}{\tau_2}\right\}. \tag{A.20}$$

To see this, we substitute ϵ_1 and ϵ_2 for δ in (A.3), take $z = 4$ and 2 respectively, to find a term that is lower bounded by $(\tau \log p + \log 2)/e^2$. Here, the $1/\sqrt{n}$ factor in ϵ_1 and ϵ_2 cancels out with the $1/\sqrt{n}$ term generated by the $\sqrt{\text{Var}[f(X)]}$ term in the denominator. (Recall that in our scenario $f(X)$ is an empirical average). The more stringent condition on p comes from ϵ_2 and is stated in (A.20). Thus, if (A.20) holds, (A.3) is satisfied. Since $\tau > 3$, the right-hand side of (A.20) never exceeds

$$\exp\left\{\frac{1}{3}(2\sqrt{e} - \log 2)\right\} < 3.$$

Hence, in our application of Lemma A.3, the condition from (A.3) holds for $p \geqslant 3$.

Now applying Lemma A.3, we know that for the absolute constant L specified in Lemma A.2, we have,

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_{ij}X_{ik}X_{i\ell}^2 - \mathbb{E}[X_jX_kX_\ell^2]\right| > \epsilon_1\right] < \exp\left\{-\frac{2}{L}\left(\frac{\sqrt{n}\epsilon_1}{\sqrt{\max_{j,k,l} \text{Var}[X_j^2X_kX_l]}}\right)^{\frac{1}{4}}\right\},$$

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_{ij}X_{ik} - \mathbb{E}[X_jX_k]\right| > \epsilon_2\right] < \exp\left\{-\frac{2}{L}\left(\frac{\sqrt{n}\epsilon_2}{\sqrt{\max_{j,k} \text{Var}[X_jX_k]}}\right)^{\frac{1}{2}}\right\}$$

for all $j, k, \ell = 1, \ldots, p$. By a union bound over no more than $2p^3$ events, we have both $\|\mathbf{R}_1\|_\infty < \epsilon_1$ and $\|r_2\|_\infty < \epsilon_2$ with probability at least $1 - 1/p^{\tau-3}$ as $p \to \infty$. Applying Theorem 1 with the chosen ϵ_1 and ϵ_2 thus shows that our choices for λ and n lead to the claim in Corollary 2.

Appendix B

SUPPLEMENT TO CHAPTER 3

B.1 Technical Lemmas

Due to Chapters 2 and 3 being closely interlinked, some of the technical lemmas used in Chapter 3 can be found in the Appendix for Chapter 2.

Lemma A.4. *Suppose that* $X^{(i)} = (X_1^{(i)}, X_2^{(i)}, \ldots, X_p^{(i)}) \in \mathbb{R}^p$, $i = 1, \ldots, n$ *are i.i.d. and follow a non-negative Gaussian distribution with parameter* $\mathbf{K}^*$, *i.e.* $X^{(i)} \sim N_+(0, (\mathbf{K}^*)^{-1})$. *Furthermore, suppose* $Z_j^{(i)}$ *for* $i = 1, \ldots, n$ *is defined as*

$$Z_j^{(i)} = \begin{cases} X_j^{(i)}, & \text{with probability } 1-\rho, \\ 0, & \text{otherwise.} \end{cases} \tag{B.1}$$

Define $v_1 = \max_{j,k,l} Var[X_j^2 X_k X_l]$ *and* $v_2 = \max_{j,k} Var[X_j X_k]$. *Then,*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\left(Z_j^{(i)}\right)^2 Z_k^{(i)} Z_l^{(i)} - \mathbb{E}[Z_j^2 Z_k Z_l]\right| > \epsilon\right] \leqslant c_2 \exp\left\{-c_1\left(\epsilon\sqrt{\frac{n(1-\rho)^3}{v_1}}\right)^{1/4}\right\}, \tag{B.2}$$

for $j, k, l \in \{1, 2, \ldots, p\}$ *not necessarily distinct. Similarly,*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} Z_j^{(i)} Z_k^{(i)} - E[Z_j Z_k]\right| > \epsilon\right] \leqslant c_2 \exp\left\{-c_1\left(\epsilon\sqrt{\frac{n(1-\rho)^2}{v_2}}\right)\right\}. \tag{B.3}$$

Proof. We prove (B.2). The proof to (B.3) is similar requiring only minor adjustments. Write $N_0 = \#\{i : \left(Z_j^{(i)}\right)^2 Z_k^{(i)} Z_l^{(i)} \neq 0\}$, and $\phi_0 := P[Z_j \neq 0, Z_k \neq 0, Z_l \neq 0] = (1-\rho)^3$. Assume, without loss of generality, that the product $\left(Z_j^{(i)}\right)^2 Z_k^{(i)} Z_l^{(i)}$ is nonzero for the first N_0 observations. Observe that the left-hand side of (B.2) is equal to

$$\mathbb{P}\left[\left|\left(\frac{N_0}{n}\right)\left(\frac{1}{N_0}\sum_{i=1}^{n}\left(X_j^{(i)}\right)^2 X_k^{(i)} X_l^{(i)}\right) - \phi_0 E[X_j^2 X_k X_l]\right| > \epsilon\right],$$

which in turn, by triangle inequality, is upper bounded by

$$\mathbb{P}\left[\left|\left(\frac{N_0}{n}\right)\left[\left(\frac{1}{N_0}\sum_{i=1}^{N_0}\left(X_j^{(i)}\right)^2 X_k^{(i)} X_l^{(i)}\right) - \mathbb{E}[X_j^2 X_k X_l]\right]\right| > \frac{\epsilon}{2}\right] + \mathbb{P}\left[\left|\left(\frac{N_0}{n}\right) E[X_j^2 X_k X_l] - \phi_0 E[X_j^2 X_k X_l]\right| > \frac{\epsilon}{2}\right].$$

We deal with the second term first. Note that since $\mathbb{E}[X_j^2 X_k X_l]$ is positive, we have,

$$\begin{aligned}\mathbb{P}\left[\left|\left(\frac{N_0}{n}\right)\mathbb{E}[X_j^2 X_k X_l] - \phi_0 \mathbb{E}[X_j^2 X_k X_l]\right| > \epsilon\right] &= \mathbb{P}\left[\left|\frac{N_0}{n} - \phi_0\right| > \frac{\epsilon}{2\mathbb{E}[X_j^2 X_k X_l]}\right] \\ &\leqslant c'' \exp\left\{-c' n\right\}, \end{aligned} \tag{B.4}$$

where the last line follows from Hoeffding's inequality for binomial counts (c' and c'' are both positive constants).

For the first term, we have

$$\begin{aligned}&\mathbb{P}\left[\left|\left(\frac{N_0}{n}\right)\left[\left(\frac{1}{N_0}\sum_{i=1}^{N_0}\left(X_j^{(i)}\right)^2 X_k^{(i)} X_l^{(i)}\right) - \mathbb{E}[X_j^2 X_k X_l]\right]\right| > \frac{\epsilon}{2}\right] \\ &\leqslant \mathbb{P}\left[\left|\left[\left(\frac{1}{N_0}\sum_{i=1}^{N_0}\left(X_j^{(i)}\right)^2 X_k^{(i)} X_l^{(i)}\right) - E[X_j^2 X_k X_l]\right]\right| > \frac{\epsilon}{2}\right] \\ &\leqslant \mathbb{P}[N_0 > (\phi_0/2)n] \\ &+ \mathbb{P}\left[\left|\left[\left(\frac{1}{N_0}\sum_{i=1}^{N_0}\left(X_j^{(i)}\right)^2 X_k^{(i)} X_l^{(i)}\right) - \mathbb{E}[X_j^2 X_k X_l]\right]\right| > \frac{\epsilon}{2}\middle| N_0 < \frac{\phi_0}{2}n\right]\mathbb{P}[N_0 < \frac{\phi_0}{2}n] \\ &\leqslant c'' \exp\{-c'\phi_0 n\} + 2\exp\left\{c^{(3)}\left(\epsilon\sqrt{\frac{n(1-\rho)^3}{v_1}}\right)^{\frac{1}{4}}\right\}, \end{aligned} \tag{B.5}$$

for some constants c', c'', $c^{(3)} > 0$, as a result of Lemma A.3 and another application of Hoeffding's inequality. Combining (B.4) and (B.5) gives us (B.2). □

Lemma A.5. *Suppose that $X^{(i)} = (X_1^{(i)}, X_2^{(i)}, \ldots, X_p^{(i)}) \in \mathbb{R}^p$, $i = 1, \ldots, n$ are i.i.d. and follow a Gaussian distribution with 0 mean and covariance matrix $\mathbf{\Sigma}^* = \left(\sigma_{jk}^*\right)$. Let $Z_j^{(i)}$ for $i = 1, \ldots, n$ be obtained via the process defined by (B.1). Define $N_{jk} = \#\{i : Z_j^{(i)} \neq$*

0 *and* $Z_k^{(i)} \neq 0\}$, *and suppose without loss of generality that the first* N_{jk} *observations satisfy this condition. Suppose that* $n = \Omega(\log p/(1-\rho)^2)$. *Then*

$$\mathbb{P}\left[\max_{j,k}\left|\frac{1}{N_{jk}}\sum_{i=1}^{N_{jk}} Z_j^{(i)} Z_k^{(i)} - \sigma^*_{jk}\right| > \epsilon\right] \leqslant c_1 \exp(-c_2 \log p) + \exp\left\{-c_3 \frac{n(1-\rho)^2\epsilon^2}{(\max_j \sigma^*_{jj})^2} + 2\log p + \log 4\right\}, \quad \text{(B.6)}$$

with c_1, c_2, *and* $c_3 > 0$.

Proof. The structure of the proof is similar to that of Lemma A.4. Observe that

$$\begin{aligned}\mathbb{P}\left[\max_{j,k}\left|\frac{1}{N_{jk}}\sum_{i=1}^{N_{jk}} Z_j^{(i)} Z_k^{(i)} - \sigma^*_{jk}\right| > \epsilon\right] &\leqslant \mathbb{P}\left[\max_{j,k}\left|N_{jk} - n(1-\rho)^2\right| > \frac{n(1-\rho)^2}{2}\right] \\ + \mathbb{P}\left[\max_{j,k}\left|\frac{1}{N_{jk}}\sum_{i=1}^{N_{jk}} Z_j^{(i)} Z_k^{(i)} - \sigma^*_{jk}\right| > \epsilon \;\middle|\; \max_{j,k}\left|N_{jk} - n(1-\rho)^2\right| &\leqslant \frac{n(1-\rho)^2}{2}\right]. \quad \text{(B.7)}\end{aligned}$$

By union bound of p^2 events, Hoeffding's inequality, and our assumption on n, we observe that

$$\mathbb{P}\left[\max_{j,k}\left|N_{jk} - n(1-\rho)^2\right| > \frac{n(1-\rho)^2}{2}\right] \leqslant c_1 \exp(-c_2 \log p). \quad \text{(B.8)}$$

For the second term on the right-hand side of (B.7), we apply Lemma 1 in Ravikumar et al. [2011] with a union bound of p^2 events to obtain

$$\begin{aligned}&\mathbb{P}\left[\max_{j,k}\left|\frac{1}{N_{jk}}\sum_{i=1}^{N_{jk}} Z_j^{(i)} Z_k^{(i)} - \sigma^*_{jk}\right| > \epsilon \;\middle|\; \max_{j,k}\left|N_{jk} - n(1-\rho)^2\right| \leqslant \frac{n(1-\rho)^2}{2}\right] \\ &\leqslant \exp\left\{-\frac{n(1-\rho)^2\epsilon^2}{6400(\max_j \sigma^*_{jj})^2} + 2\log p + \log 4\right\}.\end{aligned}$$

□

B.2 Proofs for Section 3.3

B.2.1 Proof of Theorem 2

Since the 'block' problem (3.11) decouples across the p columns of $\mathbf{\Theta}$, we need only consider the jth optimization sub-problem given by (3.13). By assumption, $\hat{\theta}_j$ is the global optimum

to (3.13), so we must have

$$\frac{1}{2}(\hat{\theta}_j - \theta_j^*)^T \check{\mathbf{\Gamma}}_j (\hat{\theta}_j - \theta_j^*) \leqslant (\hat{\theta}_j - \theta_j^*)^T (\check{\gamma}_j - \check{\mathbf{\Gamma}}_j \theta_j^*) + \lambda\{\|\theta_j^*\|_1 - \|\hat{\theta}_j\|_1\}. \tag{B.9}$$

Define $\Delta_j = \hat{\theta}_j - \theta_j^*$. Then, by expansion,

$$\begin{aligned}\frac{1}{2}\Delta_j^T \mathbf{\Gamma}_j^* \Delta_j + \frac{1}{2}\Delta_j^T (\check{\mathbf{\Gamma}}_j - \mathbf{\Gamma}_j^*)\Delta_j \\ \leqslant \Delta_j^T (\check{\gamma}_j - \mathbf{\Gamma}_j^* \theta_j^* + \mathbf{\Gamma}_j^* \theta_j^* - \check{\mathbf{\Gamma}}_j \theta_j^*) + \lambda\{\|\theta_j^*\|_1 - \|\theta_j^* + \Delta_j\|_1\}.\end{aligned}$$

We can lower bound the left-hand side by $\nu_{\min}(\mathbf{\Gamma}^*)\|\Delta_j\|_2^2/2$, which, combined with repeated use of the triangle inequality and Hölder's inequality, gives us

$$\begin{aligned}\frac{\nu_{\min}(\mathbf{\Gamma}^*)}{2}\|\Delta_j\|_2^2 &\leqslant \|\check{\gamma}_j - \mathbf{\Gamma}_j^* \theta_j^*\|_\infty \|\Delta_j\|_1 + \left\| \left(\mathbf{\Gamma}_j^* - \check{\mathbf{\Gamma}}_j\right) \left(\frac{\hat{\theta}_j}{2} + \frac{\theta_j^*}{2}\right) \right\|_\infty \|\Delta_j\|_1 \\ &\quad + \lambda\{\|\theta_j^*\|_1 - \|\theta_j^* + \Delta_j\|_1\} \\ &\leqslant \|\check{\gamma}_j - \mathbf{\Gamma}_j^* \theta_j^*\|_\infty \|\Delta_j\|_1 + \frac{1}{2}\|\mathbf{\Gamma}_j^* - \check{\mathbf{\Gamma}}_j\|_\infty \|\Delta_j\|_1^2 + \|\mathbf{\Gamma}_j^* - \check{\mathbf{\Gamma}}_j\|_\infty \|\theta_j^*\|_1 \|\Delta_j\|_1 \\ &\quad + \lambda\{\|\theta_j^*\|_1 - \|\theta_j^* + \Delta_j\|_1\}. \end{aligned} \tag{B.10}$$

Since $\mathbf{\Gamma}^*$ is positive definite, $\nu_{\min}(\mathbf{\Gamma}^*)$ is positive, so we can lower bound the left-hand side of the inequality by 0. Furthermore, observe that the ℓ_1 constraint implies that

$$\|\Delta_j\|_1 \leqslant \|\theta_j^*\|_1 + \|\hat{\theta}_j\|_1 \leqslant 2|||\mathbf{\Theta}^*|||_1.$$

Referring to assumptions (3.16) and (3.17), we arrive at

$$0 \leqslant \left(2\epsilon_1 |||\mathbf{\Theta}^*|||_1 \|\Delta_j\|_1 + \epsilon_2\right) \|\Delta_j\|_1 + \lambda_n\{\|\theta_j^*\|_1 - \|\theta_j^* + \Delta_j\|_1\}.$$

Let $S = \{k : \theta_{jk}^* \neq 0\}$. Then, we additionally have

$$\|\theta_j^* + \Delta_j\|_1 - \|\theta_j^*\|_1 \geqslant \|\theta_{j,S}^*\|_1 + \|\Delta_{j,S^c}\|_1 - \|\Delta_{j,S}\|_1 + \|\theta_j^*\|_1 = \|\Delta_{j,S^c}\|_1 - \|\Delta_{j,S}\|_1,$$

which implies that

$$\begin{aligned}0 &\leqslant \left(2\epsilon_1 |||\mathbf{\Theta}^*|||_1 + \epsilon_2\right) \left(\|\Delta_{j,S}\|_1 + \|\Delta_{j,S^c}\|_1\right) + \lambda\{\|\Delta_{j,S}\|_1 - \|\Delta_{j,S^c}\|_1\} \\ &\leqslant \frac{3\lambda}{2}\|\Delta_{j,S}\|_1 - \frac{\lambda_n}{2}\|\Delta_{j,S^c}\|_1,\end{aligned}$$

based on our choice of λ. Therefore,

$$\|\Delta_j\|_1 \leqslant 4\|\Delta_{j,S}\|_1 \leqslant 4\sqrt{d}\|\Delta_{j,S}\|_2. \tag{B.11}$$

Furthermore, one can observe from (B.10) that

$$\begin{aligned}\frac{\nu_{\min}(\mathbf{\Gamma}^*)}{2}\|\Delta_j\|_2^2 \;\leqslant\; (2\epsilon_1|||\mathbf{\Theta}^*|||_1 + \epsilon_2)\left(\|\Delta_{j,S}\|_1 + \|\Delta_{j,S^c}\|_1\right) + \lambda\{\|\Delta_{j,S}\|_1 - \|\Delta_{j,S^c}\|_1\} \\ \leqslant\; \frac{3}{2}\lambda\|\Delta_j\|_1,\end{aligned}$$

which implies that

$$\begin{aligned}\|\Delta_j\|_2^2 &\leqslant \frac{3\lambda_n}{\nu_{\min}(\mathbf{\Gamma}^*)}\|\Delta_j\|_1 \leqslant \frac{12\lambda\sqrt{d}}{\nu_{\min}(\mathbf{\Gamma}^*)}\|\Delta_j\|_2 \\ \|\Delta_j\|_2 &\leqslant \frac{12\sqrt{d}}{\nu_{\min}(\mathbf{\Gamma}^*)}\lambda \\ \|\Delta_j\|_1 &\leqslant \frac{48d}{\nu_{\min}(\mathbf{\Gamma}^*)}\lambda,\end{aligned}$$

where the last line follows from (B.11). This then implies that

$$\max_j \|\Delta_j\|_1 \leqslant \frac{48d}{\nu_{\min}(\mathbf{\Gamma}^*)}\lambda,$$

as claimed. To see that $\hat{\mathbf{\Theta}}^{(2)}$ satisfies (3.18) scaled up by 2, observe that, from definition,

$$\left|\left|\left|\hat{\mathbf{\Theta}}^{(2)} - \mathbf{\Theta}^*\right|\right|\right|_1 \leqslant \left|\left|\left|\hat{\mathbf{\Theta}}^{(2)} - \hat{\mathbf{\Theta}}^{(1)}\right|\right|\right|_1 + \left|\left|\left|\hat{\mathbf{\Theta}}^{(1)} - \mathbf{\Theta}^*\right|\right|\right|_1 \leqslant 2 \times \left|\left|\left|\hat{\mathbf{\Theta}}^{(1)} - \mathbf{\Theta}^*\right|\right|\right|_1,$$

via triangle inequality.

B.2.2 Proof of Theorem 3

We say that a generic $p \times p$ matrix $\mathbf{A}$ satisfies the *lower restricted eigenvalue* (lower-RE) condition with curvature $s_1 > 0$ and tolerance $s_2 > 0$ if

$$u^T\mathbf{A}u \geqslant s_1\|u\|_2^2 - s_2\|u\|_1^2 \quad \text{for all } u \in \mathbb{R}^p. \tag{B.12}$$

Furthermore, $\mathbf{A}$ satisfies the *upper restricted eigenvalue* (upper-RE) condition with curvature $s_1 > 0$ and tolerance $s_2 > 0$ if

$$u^T\mathbf{A}u \leqslant s_1\|u\|_2^2 + s_2\|u\|_1^2 \quad \text{for all } u \in \mathbb{R}^p. \tag{B.13}$$

One can show that matrices $\tilde{\mathbf{\Gamma}}_j$, $j = 1, 2, \ldots, p$, satisfy the lower- and upper-RE conditions, as defined in (B.12) and (B.13), for some set of parameters.

We observe that

$$\begin{aligned} u^T \check{\mathbf{\Gamma}}_j u &= u^T \mathbf{\Gamma^*}_j u + u^T (\check{\mathbf{\Gamma}}_j - \mathbf{\Gamma}^*_j) u \\ &\geqslant \nu_{\min}(\mathbf{\Gamma}^*) \|u\|_2^2 + u^T (\check{\mathbf{\Gamma}}_j - \mathbf{\Gamma}^*_j) u \\ &\geqslant \nu_{\min}(\mathbf{\Gamma}^*) \|u\|_2^2 - \|(\check{\mathbf{\Gamma}}_j - \mathbf{\Gamma}^*_j) u\|_\infty \|u\|_1 \\ &\geqslant \nu_{\min}(\mathbf{\Gamma}^*) \|u\|_2^2 - \|(\check{\mathbf{\Gamma}}_j - \mathbf{\Gamma}^*_j)\|_\infty \|u\|_1^2 \\ &\geqslant \nu_{\min}(\mathbf{\Gamma}^*) \|u\|_2^2 - \epsilon_1 \|u\|_1^2, \end{aligned}$$

and likewise,

$$u^T \check{\mathbf{\Gamma}}_j u \leqslant \nu_{\max}(\mathbf{\Gamma}^*) \|u\|_2^2 + \epsilon_1 \|u\|_1^2,$$

with $\nu_{\max}(\mathbf{\Gamma}^*)$ referring to the largest eigenvalue of the matrix $\mathbf{\Gamma}^*$. Thus, the matrices $\tilde{\mathbf{\Gamma}}_j$ satisfy the lower-RE condition with curvature $s_{1,L} \equiv \nu_{\min}(\mathbf{\Gamma}^*)$, and tolerance $s_2(n,p) = \epsilon_1$. These matrices also satisfy the upper-RE condition with a different curvature $s_{1,U} \equiv \nu_{\max}(\mathbf{\Gamma}^*)$, but same tolerance $s_2(n,p)$.

From here, we can apply Theorem 2 in Agarwal et al. [2012]. While the theorem considers convex functions, careful examination of their proof shows that we can apply their theorem in this case as well, since lower- and upper-RE conditions tie into their restricted convexity and smoothness assumptions. In the notation of Agarwal et al. [2012], define $\mathcal{L}_n$ to be the empirical loss function, and $\mathcal{M}$ to be the subspace of all vectors with support contained within the support set of θ^*_j. Then, it is apparent that $\psi(\mathcal{M}) = \sqrt{d}$; here, $\psi(\mathcal{M})$ corresponds to the Lipschitz constant of the penalty with respect to the error norm, when restricted to $\mathcal{M}$. Hence, the compound contraction coefficient $\kappa(\mathcal{L}_n; \mathcal{M})$, which determines how quickly the optimization error shrinks, can be computed to be

$$\kappa(\mathcal{L}_n; \mathcal{M}) = \left(1 - \frac{s_{1,L}}{4 s_{1,U}} + \frac{64 d \epsilon_1}{s_{1,U}}\right) \xi(\mathcal{M}) \in (0, 1). \tag{B.14}$$

with

$$\xi(\mathcal{M}) = \left(1 - \frac{64d\epsilon_1}{s_{1,U}}\right)^{-1} \lesssim 1 \tag{B.15}$$

based on the assumption that $d\epsilon_1 \lesssim 1$. Furthermore, using the notation of Agarwal et al. [2012], we have

$$\beta(\mathcal{M}) = 2\left(\frac{(s_{1,L} - 64d\epsilon_1)}{4s_{1,U}} + \frac{128d\epsilon_1}{(s_{1,L} - 64d\epsilon_1)}\right)\epsilon_1 + 10c'\epsilon_1^2 \tag{B.16}$$

and constraint radius $\bar{\rho} = |||\mathbf{\Theta}^*|||_1$. Put together, one finds that the compound tolerance parameter, which determines the radius up to which geometric convergence can be achieved, is given by $\varepsilon^2 \lesssim d\epsilon_1 \|\hat{\theta}_j^2 - \theta_j^*\|_2^2 \lesssim \|\hat{\theta}_j^2 - \theta_j^*\|_2^2$, and the proof is complete.

B.2.3 Proof of Corollary 3

Our first goal is to show that the deviation bounds (3.16) and (3.17) hold with high probability with

$$\epsilon_1 = c'\frac{(1 + 4/(1-\rho))\max_j \sigma_{jj}^*}{(1-\rho)^2}\sqrt{\frac{\log p}{n}}, \tag{B.17}$$

$$\epsilon_2 = 0. \tag{B.18}$$

for some appropriate choice of fixed $c' > 0$. Plugging these values into Theorems 2 and 3 gives the claim. For the latter, (3.17) is trivially satisfied with $\epsilon_2 = 0$, since $\mathbf{\Gamma}_j^*\theta_j^* = \gamma_j^* = \check{\gamma}_j$ for all j. Thus, we need only show that (3.16) holds for the ϵ_1 given above with high probability.

In the Gaussian case, $\mathbf{\Gamma}_j^*$ is the true covariance matrix $\mathbf{\Sigma}^* = \left(\sigma_{jk}^*\right)$ for all j, and the p blocks of $\check{\mathbf{\Gamma}}$ are identical. It is straightforward to see that

$$\begin{aligned}
\|\mathbf{\Gamma}^* - \check{\mathbf{\Gamma}}\|_\infty &= \|\mathbf{\Gamma}_j^* - \check{\mathbf{\Gamma}}_j\|_\infty \quad \text{for all } j \\
&= \left\|\frac{\mathbf{z}^T\mathbf{z}}{n} \oplusdot \mathbf{M} - \mathbf{\Sigma}^*\right\|_\infty \\
&\leqslant \frac{1}{(1-\rho)^2}\left\|\frac{\mathbf{z}^T\mathbf{z}}{n} - \mathbf{\Sigma}_Z^*\right\|_\infty,
\end{aligned}$$

where $\mathbf{\Sigma}_Z^* = \mathbb{E}[\mathbf{z}^T\mathbf{z}/n] = \left(\sigma_{Z,jk}^*\right)$.

One observes, by definition, that X_j is sub-Gaussian with parameter $(\sigma^*_{jj})^{1/2}$. It is easy to show that Z_j is also sub-Gaussian with parameter $(\sigma^*_{jj})^{1/2}$ from definition:

$$\begin{aligned}\mathbb{E}[\exp(tZ_j)] &= \rho[\exp(0)] + (1-\rho)E[\exp(tX_j)] \\ &\leqslant \rho + (1-\rho)\exp\left(\frac{\sigma^*_{jj}t^2}{2}\right) \\ &\leqslant \rho\exp\left(\frac{\sigma^*_{jj}t^2}{2}\right) + (1-\rho)\exp\left(\frac{\sigma_{jj}t^2}{2}\right) \\ &\leqslant \exp\left(\frac{\sigma^*_{jj}t^2}{2}\right),\end{aligned}$$

which implies that $Z_j/(\sigma^*_{Z,jj})^{1/2}$ is sub-Gaussian with scale parameter $(\sigma^*_{jj}/\sigma^*_{Z,jj})^{1/2} = (1-\rho)^{-1/2}$. Applying Lemma A.1, which states a Bernstein-type concentration inequality for sub-Gaussian random variables, and performing a union bound over at most p^2 events gives us,

$$\mathbb{P}\left[\|\mathbf{\Gamma}(\mathbf{z}) - \mathbf{\Sigma}^*_Z\|_\infty > \epsilon_1\right] \leqslant \exp\left\{-\frac{n\epsilon_1^2}{128[1+4/(1-\rho)]^2(\max_j \sigma^*_{jj})^2} + 2\log p + \log 4\right\}. \quad \text{(B.19)}$$

Given our choice of ϵ_1 from (B.17), it follows that

$$\mathbb{P}[\|\check{\mathbf{\Gamma}} - \mathbf{\Gamma}^*\|_\infty \leqslant \epsilon_1] > 1 - c_1\exp(-c_2\log p). \quad \text{(B.20)}$$

By Theorem 2, our choice of λ should be

$$\lambda \propto \frac{\max_j \sigma^*_{jj}|||\mathbf{\Theta}^*|||_1[1+4/(1-\rho)]}{(1-\rho)^2}\sqrt{\frac{\log p}{n}}. \quad \text{(B.21)}$$

Then if our choice of n satisfies (3.20), we arrive at the claim statement.

To prove the second half of the theorem, we take instead

$$\begin{aligned}\epsilon_1 &= c'\frac{\max_j \sigma^*_{jj}}{(1-\rho)^3}\sqrt{\frac{\log p}{n}}, \\ \epsilon_2 &= 0.\end{aligned}$$

The rest of the proof is the same. We then use Lemma A.5 from Section B.1 to show that (3.16) holds with our choice of ϵ_1 with probability greater than $1 - c_1\exp(-c_2\log p)$.

B.2.4 Proof of Corollary 4

We consider the surrogates given in (3.6) first. Our goal then is to show that the deviation bounds (3.16) and (3.17) hold with high probability with

$$\epsilon_1 = c' \frac{1}{(1-\rho)^{9/2}} \sqrt{v_1 \frac{(\log p)^8}{n}}, \tag{B.22}$$

$$\epsilon_2 = c'' \frac{1}{(1-\rho)^{5/2}} \sqrt{v_2 \frac{(\log p)^4}{n}}, \tag{B.23}$$

for some appropriate constants c' and c''. Substituting these results into Theorems 2 and 3 gives the desired result.

First, observe that

$$\begin{aligned} \|\check{\mathbf{\Gamma}} - \mathbf{\Gamma}^*\|_\infty &= \|\mathbf{\Gamma}(\mathbf{z}) \oslash (\mathbf{I}_{p\times p} \otimes \mathbf{M}_+) - \mathbf{\Gamma}^*\|_\infty \\ &\leqslant \frac{1}{(1-\rho)^3} \|\mathbf{\Gamma}(\mathbf{z}) - \mathbf{\Gamma}_Z^*\|_\infty , \end{aligned}$$

and similarly,

$$\begin{aligned} \|\tilde{\gamma} - \mathbf{\Gamma}^*\theta^*\|_\infty &= \|\tilde{\gamma} - \gamma^*\|_\infty \\ &= \|\gamma(\mathbf{z}) \oslash \mathrm{vec}(\mathbb{I}_{p\times p} \otimes \mathbf{M}) - \gamma^*\|_\infty \\ &\leqslant \frac{1}{(1-\rho)^2} \|\gamma(\mathbf{z}) - \gamma_Z^*\|_\infty, \end{aligned}$$

where the first equality of the second set of equations follows from the unbiasedness of score matching, and $\mathbf{\Gamma}_Z^*$ and γ_Z^* refer to the expectations of $\mathbf{\Gamma}(\mathbf{z})$ and $\gamma(\mathbf{z})$ under the true model, respectively.

By Lemma A.4 below and a union bound, it follows that

$$\begin{aligned} \mathbb{P}\left[\max_{j,k,l} \left| \frac{1}{n} \sum_{i=1}^n \left(Z_j^{(i)}\right)^2 Z_k^{(i)} Z_l^{(i)} - \mathbb{E}[Z_j^2 Z_k Z_l] \right| > \epsilon_1 \right] \\ \leqslant \; c_2 \exp\left\{ -c_1 \left(\epsilon_1 \sqrt{\frac{n(1-\rho)^3}{v_1}} \right)^{1/4} + 3\log p \right\}, \end{aligned}$$

for $j, k, l \in \{1, 2, \ldots, p\}$ not necessarily distinct. Similarly,

$$\mathbb{P}\left[\max_{j,k}\left|\frac{1}{n}\sum_{i=1}^{n} Z_j^{(i)} Z_k^{(i)} - \mathbb{E}[Z_j Z_k]\right| > \epsilon_2\right] \leqslant c_2 \exp\left\{-c_1\left(\epsilon_2\sqrt{\frac{n(1-\rho)^2}{v_2}}\right)^{1/2} + 2\log p\right\}.$$

For our choice of ϵ_1 and ϵ_2 from (B.22) and (B.23), we have that both (3.16) and (3.17) hold with probability at least $1 - c_1 \exp(-c_2 \log p)$.

By Theorem 2, our choice of λ should be

$$\begin{aligned}\lambda &= c'''|||\mathbf{\Theta}^*|||_1 \frac{1}{(1-\rho)^{9/2}}\sqrt{v_1\frac{(\log p)^8}{n}} + c^{(4)}\frac{1}{(1-\rho)^{5/2}}\sqrt{v_2\frac{(\log p)^4}{n}} \\ &< c^{(5)}|||\mathbf{\Theta}^*|||_1 \frac{1}{(1-\rho)^{9/2}}\sqrt{\max\{v_1, v_2\}\frac{(\log p)^8}{n}},\end{aligned}$$

and with our lower bound in (3.22), the theorem follows. The proof proceeds similarly when using the surrogates given in (3.8) except we would use a combination of a result analogous to that presented in Lemma A.5 and Lemma A.3 to prove that the deviation bounds hold with high probability.

B.3 Support recovery guarantees

In this part of the Appendix, we prove support recovery guarantees for the regularized score matching framework for missing data proposed in Chapter 3, for completeness. While we rely on fairly classical proof techniques (i.e. primal-dual witness), we need to be a bit more careful here, as the optimization problems of interest are non-convex. We feel it is necessary to state that some of the mathematical arguments were inspired by those used in a recent work by [CITE].

We consider the 'block' problem (3.11), which as one may recall, is characterized by

$$\hat{\theta} \in \arg\min_{\|\theta_j\| \leqslant R \ \forall j} \frac{1}{2}\theta^T \check{\mathbf{\Gamma}}\theta - \check{\gamma}^T\theta + \lambda\|\theta\|_1.$$

Let $\bar{\theta}$ be any stationary point of the above optimization problem in the feasible region. Then, $\bar{\theta}$ must satisfy the first-order necessary conditions, given by

$$\langle \check{\mathbf{\Gamma}}\bar{\theta} - \check{\gamma} + \lambda\bar{\omega}, \theta - \bar{\theta}\rangle \geqslant 0 \quad \text{for all feasible } \theta \in \mathbb{R}^{p^2}. \tag{B.24}$$

where $\bar{\omega} \in \partial\|\bar{\theta}\|$. As before, define $S = \{(j,k) : j = k, \theta^*_{jk} \neq 0\}$. In the rest of this chapter, we show that for n sufficiently large, tuning parameters (λ, R) appropriately chosen, and $\min_{1\leqslant j\leqslant k\leqslant p} |\theta^*_{jk}|$ sufficiently large, the problem (3.11) has a *unique* stationary point, $\hat{\theta}$ (i.e. $\bar{\theta}$ is unique and equal to $\hat{\theta}$). Additionally, $\hat{S} = S$ with high probability, where $\hat{S} = \{(j,k) : \hat{\theta}_{jk} \neq 0\}$

Noting that $\tilde{\mathbf{\Gamma}}$ is block-diagonal, the 'block' problem decomposes into p sub-problems. As a refresher, the jth sub-problem (3.13) is

$$\hat{\theta}_j \in \arg\min_{\|\theta_j\|\leqslant R} \frac{1}{2}\theta_j^T \check{\mathbf{\Gamma}}_j \theta_j - \check{\gamma}_j^T \theta_j + \lambda\|\theta_j\|_1.$$

Define $S_j = \{k : \theta^*_{j,k} \neq 0\}$ (and likewise its analogue, $\hat{S}_j$) and $d = \max_j |S_j|$. To prove our claim, i.e. uniqueness and sparsistency of $\hat{\theta}$, it suffices to demonstrate that $\hat{\theta}_j$ is the unique stationary point the jth sub-problem and that $\hat{S}_j = S$ for all $j \in \{1, \ldots, p\}$. Our proof studies a single jth problem; we then proceed to show that these results hold jointly.

Primal-dual witness (PDW), see Section A.2.1, forms the foundation of this proof. We briefly review the key steps:

1. Optimize the restricted problem, given by

$$\tilde{\theta}_j \in \arg\min_{\beta\in\mathbb{R}^p : \theta_{j,S_j^c}=0,\ \|\theta_j\|_1\leqslant R} \frac{1}{2}\theta_j^T \check{\mathbf{\Gamma}}_j \theta_j - \tilde{\gamma}_j^T \theta_j + \lambda\|\theta_j\|_1. \tag{B.25}$$

   We then demonstrate that $\tilde{\theta}_j$ lies in the interior of the feasible set, i.e. $\|\tilde{\theta}_j\| < R$.

2. Set $\tilde{\omega}_{j,S_j} \in \partial\|\tilde{\theta}_{j,S_j}\|_1$. Additionally, choose $\tilde{\omega}_{j,S_j^c}$ such that the

$$\check{\mathbf{\Gamma}}_j \tilde{\theta}_j - \check{\gamma}_j + \lambda\tilde{\omega}_j = 0, \tag{B.26}$$

   which are the zero sub-gradient conditions. Demonstrate the strict dual feasibility of $\tilde{\omega}_{j,S_j^c}$ by showing that $\|\tilde{\omega}_{j,S_j^c}\|_1 < 1$.

3. Show that $\tilde{\theta}_j$ is the unique minimum in the full jth sub-problem (3.13) and that all stationary points $\bar{\theta}_j$ must be equal to $\tilde{\theta}_j$.

Assumption 7. *There exists an $\alpha \in (0, 1]$ such that*

$$\max_j \left\| \mathbf{\Gamma}^*_{j,S_j^c S_j} (\mathbf{\Gamma}^*_{j,S_j S_j})^{-1} \right\|_\infty \leqslant 1 - \alpha. \tag{B.27}$$

This is analogous to the mutual incoherence assumption needed to prove sparsistency in Chapter 2.

Theorem 9. *Suppose Assumption 7 holds and that additionally,*

(a) The following deviation conditions hold for $\check{\mathbf{\Gamma}}$ and $\check{\gamma}$, as before, for some $\epsilon_1, \epsilon_2 > 0$:

$$\|\check{\mathbf{\Gamma}} - \mathbf{\Gamma}^*\|_\infty \leqslant \epsilon_1, \tag{B.28}$$

$$\|\check{\gamma} - \gamma^*\|_\infty \leqslant \epsilon_2, \tag{B.29}$$

for some $\epsilon_1, \epsilon_2 > 0$

(b) For the ϵ_1, ϵ_2 from (a), the tuning parameters (λ, R) are chosen to satisfy:

$$\lambda > 4(\epsilon_1 R + \epsilon_2) \tag{B.30}$$

$$R \geqslant \max_j \|\theta_j^*\|_1. \tag{B.31}$$

(c) For some $\varepsilon \in \left(\frac{2\epsilon_1 R}{\lambda}, 1\right]$, we have

$$\|\tilde{\omega}_{j,S^c}\|_\infty < 1 - \varepsilon. \tag{B.32}$$

Then if

$$d\epsilon_1 \left(\max\left\{ 16, \frac{4(3-\alpha)}{\alpha} \right\} + 2 \right)^2 \leqslant \frac{\nu_{\min}(\mathbf{\Gamma}^*)}{2}, \tag{B.33}$$

then the full problem (3.13) has a unique stationary point, given by $\tilde{\theta}_j$, the solution to the restricted problem (B.25).

The proof to Theorem 9 is comprised of two separate lemmas, proven below. Lemma A.6 claims that the restricted sub-problems, characterized by (B.25), are strictly convex, which implies that $\tilde{\theta}_j$'s are unique. In Lemma A.7, we claim that all stationary points to the full problem (3.13) are equal to $\tilde{\theta}_j$; thus, there is a unique stationary point.

Lemma A.6. *Under the assumptions of Theorem 9, the restricted problem (B.25) is strictly convex, i.e.* $\tilde{\mathbf{\Gamma}}_{j,S_jS_j}$ *is positive definite. Note that this holds for all* j*.*

Proof. It suffices to show that for the sub-matrix $\check{\mathbf{\Gamma}}_{j,S_jS_j}$ is positive definite. This is equivalent to proving that for any $u \in \{u \in \mathbb{R}^p : \|u\|_2 = 1\}$ that

$$u^T \check{\mathbf{\Gamma}}_j u > 0. \tag{B.34}$$

Additionally, recall that under assumption (B.28), $\check{\mathbf{\Gamma}}_j$ satisfies the lower-RE condition (B.12) with curvature $s_{1,L} = \nu_{\min}(\mathbf{\Gamma}^*)$ and tolerance $s_2(n,p) = \epsilon_1$, so

$$u^T \check{\mathbf{\Gamma}}_j \hat{u} \geqslant \nu_{\min}(\mathbf{\Gamma}^*)\|u\|_2^2 - \epsilon_1 \|u\|_1^2. \tag{B.35}$$

Since $u_{S^c} = 0$, $\|u\|_1 \leqslant \sqrt{d}\|a\|_2$; hence, (B.35) implies

$$u^T \check{\mathbf{\Gamma}}_j u \geqslant \nu_{\min}(\mathbf{\Gamma}^*)\|u\|_2^2 - d\epsilon_1 \|u\|_2^2. \tag{B.36}$$

By assumption (B.33), $d\epsilon_1 \leqslant \nu_{\min}(\mathbf{\Gamma}^*)/2$, so the right-hand side of (B.35) is upper-bounded by 0, as we have assumed $\nu_{\min}(\mathbf{\Gamma}^*) > 0$, and we have proven the claim that the solution to the restricted problem (B.25), the $\tilde{\theta}_j$, is unique. □

Lemma A.7. *Under the assumptions of Theorem 9, any stationary point of the full problem (3.13),* $\bar{\theta}_j$*, satisfies* $\bar{S}_j \subseteq \tilde{S}$*. More importantly,* $\bar{\theta}_j = \tilde{\theta}_j$*.*

Proof. Again, any stationary solution $\bar{\theta}_j$ satisfies the first-order optimality condition (B.24), and since the solution to the restricted problem (B.25) $\tilde{\theta}_j$ is feasible, by construction, we have, by the first-order optimality conditions,

$$-\langle \check{\mathbf{\Gamma}}_j \bar{\theta}_j + \lambda \bar{\omega}_j, \bar{\Delta}_j \rangle \geqslant 0, \tag{B.37}$$

where we have defined $\bar{\Delta}_j = \bar{\theta}_j - \hat{\theta}_j$. By construction, $\tilde{\theta}_j$ satisfies the zero-subgradient condition (B.26): $\langle \check{\mathbf{\Gamma}}_j \tilde{\theta}_j + \lambda \check{\omega}_j, \bar{\theta}_j - \check{\theta}_j \rangle = 0$. Combining (B.37) with (B.26) yields

$$0 \leqslant \langle \check{\mathbf{\Gamma}}_j \hat{\theta}_j - \check{\mathbf{\Gamma}}_j \bar{\theta}_j, \hat{\Delta}_j \rangle + \lambda \langle \hat{\omega}_j, \bar{\theta}_j \rangle + \lambda \langle \bar{\omega}_j, \tilde{\theta}_j \rangle - \lambda \|\tilde{\theta}_j\|_1 - \lambda \|\bar{\theta}_j\|_1. \tag{B.38}$$

Additionally, one can show, using the same arguments used to derive (B.35) that

$$\langle \check{\mathbf{\Gamma}}_j \bar{\theta}_j - \check{\mathbf{\Gamma}}_j \tilde{\theta}_j, \bar{\Delta}_j \rangle \geqslant \nu_{\min}(\mathbf{\Gamma}^*) \|\bar{\Delta}_j\|_2^2 - \epsilon_1 \|\bar{\Delta}_j\|_1^2. \tag{B.39}$$

Then, a simple rearrangement of (B.38) yields

$$\begin{aligned} \lambda \|\bar{\theta}_j\|_1 - \lambda \langle \tilde{\omega}_j, \bar{\theta}_j \rangle &\leqslant \langle \check{\mathbf{\Gamma}}_j \tilde{\theta}_j - \check{\mathbf{\Gamma}}_j \bar{\theta}_j, \bar{\Delta}_j \rangle + \lambda \langle \bar{\omega}_j, \tilde{\theta}_j \rangle - \lambda \|\tilde{\theta}_j\|_1 \\ &\leqslant \langle \check{\mathbf{\Gamma}}_j \tilde{\theta}_j - \check{\mathbf{\Gamma}}_j \bar{\theta}_j, \bar{\Delta}_j \rangle + \lambda \|\bar{\omega}_j\|_\infty \|\tilde{\theta}_j\|_1 - \lambda \|\tilde{\theta}_j\|_1 \\ &\leqslant \langle \check{\mathbf{\Gamma}}_j \tilde{\theta}_j - \check{\mathbf{\Gamma}}_j \bar{\theta}_j, \bar{\Delta}_j \rangle \\ &\leqslant \epsilon_1 \|\bar{\Delta}_j\|_1^2 - \nu_{\min}(\mathbf{\Gamma}^*) \|\bar{\Delta}_j\|_2^2. \end{aligned} \tag{B.40}$$

Combining (B.38) with (B.39) also results in the following inequality

$$\nu_{\min}(\mathbf{\Gamma}^*) \|\bar{\Delta}_j\|_2^2 - \epsilon_1 \|\bar{\Delta}_j\|_1^2 \leqslant \langle \check{\mathbf{\Gamma}}_j \bar{\theta}_j - \check{\mathbf{\Gamma}}_j \tilde{\theta}_j \rangle \leqslant \lambda \langle \hat{\omega}_j, \bar{\theta}_j \rangle + \lambda \langle \bar{\omega}_j, \tilde{\theta}_j \rangle - \lambda \|\bar{\theta}_j\|_1. \tag{B.41}$$

Since $\tilde{\theta}_j$ is the solution to the restricted problem (B.25), we are guaranteed $\tilde{S} \subseteq S$, which in turn, implies that

$$\lambda \langle \bar{\omega}_j, \tilde{\theta}_j \rangle - \lambda \|\bar{\theta}_j\|_1 \leqslant \lambda \|\tilde{\theta}_j\| - \lambda \|\bar{\theta}_j\|_1 = \lambda \left(\|\tilde{\theta}_{j,S_j}\|_1 - \|\bar{\theta}_{j,S_j}\|_1 - \|\tilde{\theta}_{j,S_j^c}\|_1 \right) \leqslant \lambda \left(\|\bar{\Delta}_{j,S_j}\|_1 - \|\bar{\Delta}_{j,S^c}\|_1 \right). \tag{B.42}$$

In addition, we can show that

$$\begin{aligned} \lambda \langle \tilde{\omega}_j, \bar{\theta} \rangle &= \lambda (\langle \tilde{\omega}_{j,S}, \bar{\theta}_{j,S} \rangle + \langle \tilde{\omega}_{j,S^c}, \bar{\theta}_{j,S_j^c} \rangle) \\ &\leqslant \lambda (\|\tilde{\omega}_{j,S_j}\|_\infty \|\bar{\Delta}_{j,S_j}\|_1 + \|\tilde{\omega}_{j,S_j^c}\|_\infty \|\bar{\Delta}_{j,S_j^c}\|_1) \\ &\leqslant \lambda (\|\bar{\Delta}_{j,S_j}\|_1 + (1-\varepsilon) \|\bar{\Delta}_{j,S_j^c}\|_1). \end{aligned} \tag{B.43}$$

Inequalities (B.41), (B.42), and (B.43) lead to

$$-\epsilon_1 \|\bar{\Delta}_{j,S_j}\|_1^2 \leqslant \nu_{\min}(\mathbf{\Gamma}_j^*) \|\bar{\Delta}\|_2^2 - \epsilon_1 \|\bar{\Delta}_{j,S_j^c}\|_1^2 \leqslant \lambda (2 \|\bar{\Delta}_{j,S_j}\|_1 - \varepsilon \|\bar{\Delta}_{S_j^c}\|_1). \tag{B.44}$$

By (B.32), we have $\lambda \geqslant 2\epsilon_1 R/\varepsilon$, which implies that

$$-\frac{\lambda\varepsilon}{2}\|\bar{\Delta}_j\|_1 \leqslant \lambda(2\|\bar{\Delta}_{j,S_j}\|_1 - \varepsilon\|\bar{\Delta}_{j,S_j^c}\|_1).$$

which, when rearranged, becomes

$$\frac{\varepsilon}{2}\|\bar{\Delta}_{j,S_j^c}\|_1 \leqslant \left(2 + \frac{\varepsilon}{2}\right)\|\bar{\Delta}_{j,S_j}\|_1. \tag{B.45}$$

Using (B.45), one can show that

$$\|\bar{\Delta}_{j,S_j}\|_1 = \|\bar{\Delta}_{j,S_j}\|_1 + \|\bar{\Delta}_{j,S_j^c}\|_1 \leqslant \left(\frac{4}{\varepsilon} + 2\right)\|\bar{\Delta}_{j,S_j}\|_1 \leqslant \left(\frac{4}{\varepsilon} + 2\right)\sqrt{d}\|\bar{\Delta}_j\|_2,$$

which, combined with (B.40), yields

$$\lambda\|\bar{\theta}_j\|_1 - \lambda\langle\tilde{\omega}_j, \bar{\theta}_j\rangle \leqslant d\epsilon_1\left(\frac{4}{\varepsilon} + 2\right)^2\|\bar{\Delta}_j\|_2^2 - \nu_{\min}(\mathbf{\Gamma}_j^*)\|\bar{\Delta}_j\|_2^2. \tag{B.46}$$

By assumption, $d\epsilon_1\left(4/\varepsilon + 2\right)^2 \leqslant \nu_{\min}(\mathbf{\Gamma}_j^*)/2$; thus,

$$\lambda\|\bar{\theta}_j\|_1 - \lambda\langle\tilde{\omega}_j, \bar{\theta}_j\rangle \leqslant \nu_{\min}(\mathbf{\Gamma}_j^*)\|\bar{\Delta}_j\|_2^2/2 \leqslant 0. \tag{B.47}$$

By Hölder's inequality, $\lambda\langle\tilde{\omega}_j, \bar{\theta}_j\rangle \leqslant \lambda\|\bar{\theta}_j\|_1$. Since the left-hand side of (B.47) is lower bounded by 0, we must have $\lambda\|\bar{\theta}_j\|_1 = \lambda\langle\tilde{\omega}_j, \bar{\theta}_j\rangle$. On the other hand, since we have assumed $\|\tilde{\omega}_{j,S_j^c}\| < 1 - \varepsilon$, we must have $\bar{\theta}_{j,S_j^c} = 0$, which implies that $\bar{S}_j \subseteq S_j$. More importantly, we have shown that $\hat{\Delta}_j$ must be 0; therefore, $\bar{\theta}_j = \tilde{\theta}_j$, and the stationary point must be unique. Our proof is then complete. □

Proposition 10. *Suppose the assumptions in Theorem 9 hold, except here, λ is chosen to satisfy*

$$\lambda > \max\left\{4, \frac{3-\alpha}{\alpha}\right\}(\epsilon_1 R + \epsilon_2), \tag{B.48}$$

where α is the incoherence parameter from Assumption 7. Then strict duality holds provided that

$$d\epsilon_1 \leqslant \min\left\{\frac{\nu_{\min}(\mathbf{\Gamma}^*)}{\left(\max\left\{16, \frac{4(3-\alpha)}{\alpha}\right\} + 2\right)^2}, \frac{\alpha}{c_{\mathbf{\Gamma}^*}(6-2\alpha)}\right\}. \tag{B.49}$$

Proof. This proof is structured very similarly to that of Theorem 1 in Chapter 2.

The zero sub-gradient conditions (B.26) can be expanded into

$$\begin{bmatrix} \mathbf{\Gamma}^*_{j,S_jS_j} & \mathbf{\Gamma}^*_{j,S_jS_j^c} \\ \mathbf{\Gamma}^*_{j,S_j^cS_j} & \mathbf{\Gamma}^*_{j,S_j^cS_j^c} \end{bmatrix} \begin{bmatrix} \tilde{\Delta}_{j,S} \\ \tilde{\Delta}_{j,S_j^c} \end{bmatrix} + \begin{bmatrix} \check{\mathbf{R}}_{1,j,S_jS_j} & \check{\mathbf{R}}_{1,j,S_jS_j^c} \\ \check{\mathbf{R}}_{1,j,S^cS} & \check{\mathbf{R}}_{1,j,S_j^cS_j^c} \end{bmatrix} \begin{bmatrix} \tilde{\theta}_{j,S_j} \\ \tilde{\theta}_{j,S_j^c} \end{bmatrix} + \begin{bmatrix} \check{r}_{2,j,S} \\ \check{r}_{2,j,S^c} \end{bmatrix} + \lambda \begin{bmatrix} \hat{\omega}_{j,S} \\ \hat{\omega}_{j,S^c} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tag{B.50}$$

where we have defined $\tilde{\Delta} = \tilde{\theta} - \theta^*$, $\check{\mathbf{R}}_1 = \check{\mathbf{\Gamma}} - \mathbf{\Gamma}^*$ (and $\check{\mathbf{R}}_{1,j} = \check{\mathbf{\Gamma}}_j - \mathbf{\Gamma}^*_j$), and $\check{r}_2 = \tilde{\gamma} - \gamma^*$ (likewise, $\tilde{r}_{2,j} = \tilde{\gamma}_j - \gamma^*_j$). We have, by the PDW construction, that

$$\tilde{\omega}_{j,S_j^c} = \frac{1}{\lambda}\left[-\mathbf{\Gamma}^*_{j,S_j^cS_j}(\mathbf{\Gamma}^*_{j,S_jS_j})^{-1}\left(\check{\mathbf{R}}_{j,1,S_jS_j}\tilde{\theta}_{j,S_j} + \check{r}_{j,2,S_j}\right) + \check{\mathbf{R}}_{j,1,S_j^cS_j}\hat{\theta}_{j,S_j} + \check{r}_{j,2,S_j^c} + \lambda\mathbf{\Gamma}^*_{j,S_j^cS_j}(\mathbf{\Gamma}^*_{S_jS_j})^{-1}\tilde{\omega}_{j,S_j}\right] \tag{B.51}$$

It follows from the triangle inequality and Hölder's inequality that

$$\begin{aligned} \|\tilde{\omega}_{j,S^c}\|_\infty \leqslant \frac{2-\alpha}{\lambda}\|\check{\mathbf{R}}_{1,j}\|_\infty\|\theta^*_j\|_1 + \frac{(2-\alpha)}{\lambda}\left|\!\left|\!\left|\check{\mathbf{R}}_{1,j,\cdot S}\right|\!\right|\!\right|_\infty\|\tilde{\Delta}_{j,S}\|_\infty + \frac{(2-\alpha)}{\lambda}\|\check{r}_{2,j}\|_\infty \\ + 1 - \alpha \end{aligned} \tag{B.52}$$

Applying assumptions (B.28), (B.29) and (B.31), (B.52) implies

$$\|\tilde{\omega}_{j,S^c}\|_\infty \leqslant \frac{2-\alpha}{\lambda}(\epsilon_1 R + \epsilon_2) + \frac{(2-\alpha)}{\lambda}d\epsilon_1\|\Delta_{j,S}\|_\infty. \tag{B.53}$$

From Lemma A.6, $\tilde{\mathbf{\Gamma}}_{j,S_jS_j}$ is invertible. Then, one can derive an explicit expression for Δ_{j,S_j} based on the zero sub-gradient conditions:

$$\tilde{\Delta}_{j,S_j} = (\check{\mathbf{\Gamma}}_{j,S_jS_j})^{-1}(-\mathbf{R}_{j,1,S_jS_j}\theta^*_{j,S_j} - r_{2,j} - \lambda\hat{\omega}_{j,S_j}). \tag{B.54}$$

Another application of the triangle's inequality yields

$$\|\tilde{\Delta}_{j,S_j}\|_\infty \leqslant \left|\!\left|\!\left|(\check{\mathbf{\Gamma}}_{j,S_jS_j})^{-1}\right|\!\right|\!\right|_\infty\left[\|\mathbf{R}_{j,\check{1},S_jS_j}\theta^*_{j,S}\|_\infty + \|\check{r}_{2,j}\|_\infty + \lambda\|\tilde{\omega}_{j,S_j}\|_\infty\right] \tag{B.55}$$

$$\leqslant \left|\!\left|\!\left|\check{\mathbf{\Gamma}}_{j,S_jS_j})^{-1}\right|\!\right|\!\right|_\infty(\epsilon_1 R + \epsilon_2 + \lambda). \tag{B.56}$$

Using the same set of arguments as in Section A.2.1 to derive (A.16), we arrive at the

following bound for $\left|\left|\left|\check{\mathbf{\Gamma}}_{j,S_jS_j})^{-1}\right|\right|\right|_\infty$:

$$\begin{aligned}\left|\left|\left|(\check{\mathbf{\Gamma}}_{j,S_jS_j})^{-1}\right|\right|\right|_\infty &\leqslant \left|\left|\left|(\mathbf{\Gamma}^*_{j,S_jS_j})^{-1}\right|\right|\right|_\infty + \left|\left|\left|(\check{\mathbf{\Gamma}}_{j,S_jS_j})^{-1} - (\mathbf{\Gamma}^*_{j,S_jS_j})^{-1}\right|\right|\right|_\infty \\ &\leqslant \left|\left|\left|(\mathbf{\Gamma}^*_{j,S_jS_j})^{-1}\right|\right|\right|_\infty \times \frac{1}{1 - \left|\left|\left|(\mathbf{\Gamma}^*_{j,S_jS_j})^{-1}\right|\right|\right|_\infty \left|\left|\left|\check{\mathbf{R}}_{1,j,S_jS_j}\right|\right|\right|_\infty} \qquad \text{(B.57)} \\ &\leqslant \frac{c_{\mathbf{\Gamma}^*}}{1 - dc_{\mathbf{\Gamma}^*}\epsilon_1}, \qquad \text{(B.58)}\end{aligned}$$

which, when plugged into (B.55), leads to

$$\|\tilde{\Delta}_{j,S_j}\|_\infty \leqslant \frac{c_{\mathbf{\Gamma}^*}}{1 - dc_{\mathbf{\Gamma}^*}\epsilon_1}(\epsilon_1 R + \epsilon_2 + \lambda), \qquad \text{(B.59)}$$

which, returning to (B.53), implies that

$$\|\tilde{\omega}_{j,S^c}\|_\infty \leqslant \frac{(2-\alpha)}{\lambda}\left[1 + \frac{dc_{\mathbf{\Gamma}^*}\epsilon_1}{1 - dc_{\mathbf{\Gamma}^*}\epsilon_1}\right](\epsilon_1 R + \epsilon_2) + (2-\alpha)\frac{dc_{\mathbf{\Gamma}^*}\epsilon_1}{1 - dc_{\mathbf{\Gamma}^*}\epsilon_1}. \qquad \text{(B.60)}$$

We proceed to show that under theorem assumptions, the first and second terms on the right-hand side of (B.60) are each bounded by $\alpha/3$. To do so, we simply refer to the assumption that $dc_{\mathbf{\Gamma}^*}\epsilon_1 \leqslant \alpha/(6-2\alpha)$ (B.49) and our choice of λ (B.48): in particular, $\lambda > (6-2\alpha)(\epsilon_1 R + \epsilon_2)/2\alpha$. Together, these imply that

$$\begin{aligned}\frac{(2-\alpha)}{\lambda}\left[1 + \frac{dc_{\mathbf{\Gamma}^*}\epsilon_1}{1 - dc_{\mathbf{\Gamma}^*}\epsilon_1}\right](\epsilon_1 R + \epsilon_2) &\leqslant \frac{(2-\alpha)}{\lambda} \times \frac{(2 - \frac{2\alpha}{3})}{(2-\alpha)} \times \frac{2\alpha\lambda}{(6-2\alpha)} < \frac{2\alpha}{3} \\ (2-\alpha)\frac{dc_{\mathbf{\Gamma}^*}\epsilon_1}{1 - dc_{\mathbf{\Gamma}^*}\epsilon_1} &\leqslant (2-\alpha) \times \frac{\frac{\alpha}{(6-2\alpha)}}{1 - \frac{\alpha}{(6-2\alpha)}} < \frac{\alpha}{3},\end{aligned}$$

and our proof is complete. □

Corollary 8. *Suppose* $\mathbf{x} \in \mathbb{R}^{n\times p}$ *is generated from a normal distribution* $N(0, \mathbf{\Sigma}^*)$, $\mathbf{\Sigma}^*$ *positive definite, and let* $\mathbf{z} \in \mathbb{R}^{n\times p}$ *be the observed data matrix generated according to (3.1) with parameter* $\rho \in [0, 1)$. *Then, for the choice of surrogates given by (3.3), if the tuning parameters* λ *and* R *are chosen according to (B.48) and (B.31), respectively, with* ϵ_1 *set to (B.17),* ϵ_2 *set to 0, and*

$$n \gtrsim \max\left\{\frac{(\max\{16, 4(3-\alpha)/\alpha\} + 2)}{\nu_{\min}(\mathbf{\Gamma}^*)}, \frac{c_{\mathbf{\Gamma}^*}(6-2\alpha)}{\alpha}\right\}\left(\frac{\max_j \sigma^{*2}_{jj}}{(1-\rho)^6}\right) d^2 \log p,$$

we have $\hat{S} \subseteq S$ *and*

$$\|\hat{\theta} - \theta^*\|_\infty < c(\rho, \mathbf{\Gamma}^*, R)\sqrt{\frac{\log p}{n}}.$$

with probability $1 - c_1 exp(-c_2 \log p)$*, for some universal* $c_1, c_2 > 0$*.*

Proof. In Corollary 3, we prove that the deviation conditions (B.28) and (B.29) hold for these values of ϵ_1 and ϵ_2 with probability $1 - c_1 \exp(-c_2 \log p)$. Plugging these listed values into the proofs of Theorem 9 and Proposition 10, we obtain the desired result. □

Corollary 9. *Suppose* $\mathbf{x} \in \mathbb{R}^{n \times p}$ *is generated from a non-negative Gaussian distribution with parameter* $\mathbf{K}^*$*, i.e.* $N(0, (\mathbf{K}^*)^{-1})$ *truncated at* $\mathbb{R}^p_+$*,* $\mathbf{K}^*$ *positive definite, and let* $\mathbf{z} \in \mathbb{R}^{n \times p}$ *be the observed data matrix generated according to (3.1) with parameter* $\rho \in [0, 1)$*. Define* $v_1 = \max_{j,k,l} Var[X_j^2 X_k X_l]$ *and* $v_2 = \max_{j,k} Var[X_j X_k]$*. Then, for the set of surrogates given either by (3.6) or (3.8), if* λ *and* R *are chosen according to (B.48) and (B.31), respectively, with* ϵ_1 *set to (B.22),* ϵ_2 *set to (B.23), and*

$$n \gtrsim \max\left\{\frac{(\max\{16, 4(3-\alpha)/\alpha\} + 2)}{\nu_{\min}(\mathbf{\Gamma}^*)}, \frac{c_{\mathbf{\Gamma}^*}(6 - 2\alpha)}{\alpha}\right\}\left(\frac{v_1}{(1-\rho)^9}\right) d^2 (\log p)^8,$$

we have $\hat{S} \subseteq S$ *and*

$$\|\hat{\theta} - \theta^*\|_\infty < c(\rho, \mathbf{\Gamma}^*, R)\sqrt{\frac{\log p}{n}}.$$

with probability $1 - c_1 exp(-c_2 \log p)$*, for some universal* $c_1, c_2 > 0$*.*

Proof. See proof of Corollary 8. □

Appendix C

SUPPLEMENT TO CHAPTER 4

C.1 Proofs for Section 4.3

C.1.1 Proof of Proposition 4

This was proven in Shao and Deng [2012] (see Proof of Theorem 1). Define $\mathbf{\Gamma} = \begin{bmatrix} \mathbf{\Gamma}' & (\mathbf{\Gamma})_\perp \end{bmatrix}$; $\mathbf{\Gamma}'$ is orthogonal, i.e. $\mathbf{\Gamma}'^T\mathbf{\Gamma}' = \mathbf{\Gamma}'\mathbf{\Gamma}'^T = \mathbf{I}_{p\times p}$. By definition (4.5), we have

$$\begin{aligned}
\mathbb{E}[\hat{\beta}] - \theta^* &= \frac{1}{N}(\hat{\mathbf{\Sigma}} + \lambda \mathbf{I}_{p\times p})^{-1}\mathbf{X}^T\mathbf{X}\theta^* - \theta_* \\
&= -(\lambda^{-1}N^{-1}\mathbf{X}^T\mathbf{X} + \mathbf{I}_{p\times p})^{-1}\theta^* \\
&= -\mathbf{\Gamma}'(\lambda^{-1}N^{-1}\mathbf{\Gamma}'^T\mathbf{X}^T\mathbf{X}\mathbf{\Gamma}' + \mathbf{I}_{p\times p})^{-1}\mathbf{\Gamma}'^T\mathbf{\Gamma}\mathbf{\Gamma}^T\theta^* \\
&= -\mathbf{\Gamma}(\lambda^{-1}N^{-1}\mathbf{D}^2 + \mathbf{I}_{R\times R})^{-1}\mathbf{\Gamma}^T\theta^*.
\end{aligned}$$

Observing that the diagonal entries to $\mathbf{D}$ are positive, one obtains

$$(\lambda^{-1}N^{-1}\mathbf{D}^2 + \mathbf{I}_{R\times R})^{-1} \preceq \frac{\lambda^{-1}/\nu_{\min\neq 0}(\hat{\mathbf{\Sigma}})}{1 + \lambda^{-1}/\nu_{\min\neq 0}(\hat{\mathbf{\Sigma}})}\mathbf{I}_{R\times R}, \tag{C.1}$$

which, combined with the fact that $\mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{I}_{R\times R}$, we obtain

$$\max_{j\in\{1,\ldots,p\}} |\mathbb{E}[\hat{\beta}_j] - \theta_j^*| \leqslant \lambda\|\theta^*\|_2\nu_{\min\neq 0}(\hat{\mathbf{\Sigma}})^{-1},$$

as desired. The bound on the variance follows directly from (4.6)

C.1.2 Proof of Theorems 7 and 8

Because there is an overlap in the lemmas used to prove Theorems 7 and 8, we present them together. Define $u^* = \|\mathbf{X}^T(y - \mathbf{X}\beta^*)\|_\infty/N$.

Lemma A.8. *Suppose we take* λ_L *to be*

$$\lambda_L = \frac{(\xi+1)}{(\xi-1)}\sqrt{\frac{2(\sigma^{*2}+\tau^{*2}qn)(\log(p)-\log(\varepsilon/2))}{N}}. \tag{C.2}$$

Under the model given by (4.3), the event $u^* \leqslant \lambda_L(\xi-1)/(\xi+1)$ *occurs with probability greater than* $1-\varepsilon$.

Proof. Define $u_j = x_j^T(y - \mathbf{X}\beta^*)/N$. Then $u^* = \max_j |u_j|$. Under model (4.3), we observe that,

$$u_j \sim N(0, x_j^T \mathbf{V}(\sigma^{*2}, \tau^{*2})x_j)$$

It follows from the sub-Gaussianity of u_j (since Gaussianity implies sub-Gaussianity) that

$$\mathbb{P}[|u_j| > \lambda_L(\xi-1)/(\xi+1)] \leqslant 2e^{-\frac{\lambda_L^2(\xi-1)^2/(\xi+1)^2}{2x_j^T\mathbf{V}(\sigma^{*2},\tau^{*2})x_j}} \leqslant 2e^{-\frac{\lambda_L^2(\xi-1)^2/(\xi+1)^2}{2N\nu_{\max}(\mathbf{V}(\sigma^{*2},\tau^{*2}))}} \leqslant \frac{\varepsilon}{p}. \tag{C.3}$$

The second inequality follows from the fact that the columns of $\mathbf{Z}$ are standardized such that $\|z_j\|_2^2 = n \;\forall\; j$, which implies that the largest eigenvalue of $\mathbf{V}(\sigma^{*2}, \tau^{*2})$ satisfies $\nu_{\max}(\mathbf{V}(\sigma^*, \tau^*)) \leqslant \sigma^{*2} + \tau^{*2}qn$. The third inequality in (C.3) is obtained by plugging in our choice of λ_L (4.29). Employing a simple union bound, we have

$$\mathbb{P}[u^* \leqslant \lambda_L] \geqslant 1 - \sum_{j=1}^{p} \mathbb{P}[|u_j| > \lambda_L] \geqslant 1 - \varepsilon,$$

This is our desired result. □

Lemma A.9. *Suppose Assumption 2 holds and let* λ_L *be defined by (C.2) (or 4.29) for some small* $\varepsilon > 0$ *and* ξ *as in Assumption 2. In the event that* $u^* \leqslant \lambda_L(\xi-1)/(\xi+1)$,

$$\|\hat{\beta}^L - \beta^*\|_\infty \leqslant \frac{\lambda_L + u^*}{\zeta} \leqslant \frac{2\xi\lambda_L}{(\xi+1)\zeta} \tag{C.4}$$

Proof. This proof follows directly from the proof of Theorem 3 in Ye and Zhang [2010]. Suppose that $u^* \leqslant \lambda_L$. Define $h = \beta^L - \beta^*$. The Karuhn-Kush-Tucker (KKT) optimality conditions for Lasso is given by

$$\begin{cases} \frac{x_j^T(y-\mathbf{X}\hat{\beta}^L)}{N} = \lambda_L \text{sign}(\hat{\beta}_j^L), & \hat{\beta}_j^L \neq 0 \\ \frac{x_j^T(y-\mathbf{X}\hat{\beta}^L)}{N} \in \lambda_L[-1,+1], & \hat{\beta}_j^L = 0 \end{cases}.$$

With some rearrangement, the KKT conditions can be rewritten as

$$\frac{\mathbf{X}^T\epsilon - \hat{\mathbf{\Sigma}}h}{N} = \lambda_L\hat{\kappa} \tag{C.5}$$

with $\hat{\kappa} \in \mathbb{R}^p$ and $\kappa_j = \text{sign}(\hat{\beta}_j^L)$ if $j \in \tilde{S}$ and $\kappa_j \in [-1, +1]$ otherwise: the subdifferential which arises from $\|\beta\|_1$. Rearranging (C.5) and observing that $\text{sign}(\hat{\beta}_j^L) = \text{sign}(h_j)$ for $j \notin S$ yields

$$h'^T\hat{\mathbf{\Sigma}}h \leqslant (u^* + \lambda_L)\|h'_S\|_1 + (u^* - \lambda_L)\|h'_{S^c}\|_1$$

for all vectors h' with $\text{sign}(h'_{S^c}) = \text{sign}(h_{S^c})$. If we take $h' = h$, one can see that $h \in \mathcal{C}(\xi, S)$:

$$0 \leqslant h^T\hat{\mathbf{\Sigma}}h \leqslant (u^* + \lambda_L)\|h'_S\|_1 + (u^* - \lambda_L)\|h'_{S^c}\|_1 \implies \|h'_{S^c}\|_1 \leqslant \frac{(u^* + \lambda_L)}{(\lambda_L - u^*)}\|h'_S\|_1 \leqslant \xi\|h'_S\|_1.$$

On the other hand, setting h' to be any vector so that for some $j \in S^c$, $h'_j = h_j$ and 0 elsewhere gives

$$h_j\hat{\mathbf{\Sigma}}_{j,\cdot}h \leqslant (u^* - \lambda_L)|h_j| \leqslant 0,$$

which implies that $h \in \mathcal{C}_-(\xi, S)$. The KKT conditions (C.5) also tell us that

$$\|\hat{\mathbf{\Sigma}}h\|_\infty \leqslant u^* + \lambda_L,$$

which, when combined with the definition of ζ (4.28) yields

$$\|h\|_\infty \leqslant \frac{u^* + \lambda_L}{\zeta},$$

which is the desired result. In the event that $u^* \leqslant \lambda_L(\xi - 1)/(\xi + 1)$, we have

$$\|h\|_\infty \leqslant \frac{2\lambda_L}{\zeta(\xi + 1)}.$$

□

Lemma A.10. *Suppose that Assumption 3 holds, and λ_L is defined according to (4.29). In the event that $u^* \leqslant \lambda_L(\xi - 1)/(\xi + 1)$, $|\hat{S}\backslash S| < N'$.*

Proof. The proof is largely adapted from that of Theorem 3 in Sun and Zhang [2012]. By construction, $\hat{\beta}^L$ satisfies the KKT conditions, given by (C.5) which implies that

$$\begin{aligned}\frac{|x_j^T X(\hat{\beta}^L - \beta^*)|}{N} &= \frac{|x_j^T(y - \mathbf{X}\hat{\beta}^L - \epsilon)|}{N} \\ &\geqslant \frac{|x_j^T(y - \mathbf{X}\hat{\beta}^L - \epsilon)|}{N} - \frac{|x_j^T\epsilon|}{N} \\ &\geqslant \lambda_L - u^*\end{aligned}$$

For $\mathcal{A} \subseteq \hat{S}\backslash S$, such that $|\mathcal{A}| \leqslant N'$, the previous inequality implies

$$\begin{aligned}(\lambda_L - u_\infty^*)^2|\mathcal{A}| &\leqslant \frac{\sum_{j\in A}|x_j^T\mathbf{X}(\hat{\beta}^L - \beta^*)|^2}{N^2} \\ &= \frac{\sum_{j\in A}(\mathbf{X}h)^T x_j x_j^T(\mathbf{X}h)}{N^2} \leqslant \frac{\kappa_+(N', S)\|\mathbf{X}h\|_2^2}{N}.\end{aligned} \tag{C.6}$$

Going back to the KKT conditions (C.5), we have, for arbitrary $h' \in \mathbb{R}^p$,

$$\frac{(\mathbf{X}\hat{\beta}^L - \mathbf{X}h')^T\mathbf{X}h}{N} \leqslant \lambda_L(\|h'\|_1 - \|\hat{\beta}^L\|_1) + U^*\|h' - \hat{\beta}^L\|_1,$$

which, when combined with the fact that

$$2(\mathbf{X}\hat{\beta}^L - \mathbf{X}h')^T\mathbf{X}h = \|\mathbf{X}\hat{\beta}^L - \mathbf{X}h'\|_2^2 + \|\mathbf{X}h\|_2^2 - \|\mathbf{X}\beta^* - \mathbf{X}h'\|_2^2$$

gives the inequality

$$\begin{aligned}\frac{\|\mathbf{X}h\|_2^2}{N} &\leqslant \lambda_L(\|\beta^*\|_1 - \|\hat{\beta}^L\|_1) + u^*\|h\|_1 \\ &\leqslant (\lambda_L + u^*)\|h_S\|_1,\end{aligned} \tag{C.7}$$

and thus h lies in the cone (4.27) in the event that $u^* \leqslant \lambda_L(\xi - 1)/(\xi + 1)$ (by noting that the left-hand side is lower bounded by 0). By definition of $\kappa(\xi, S)$ from (4.31),

$$\frac{\|\mathbf{X}h\|_2^2}{N} \leqslant \frac{(\lambda_L + u^*)^2 d}{\kappa^2(\xi, S)},$$

which, when combined with (C.6) implies

$$|\mathcal{A}| \leqslant \frac{\kappa_+(N', S)\xi^2 d}{\kappa^2(\xi, S)} < N',$$

by Assumption 3. □

Proof of Theorem 7. Suppose that $u^* \leqslant \lambda_L(\xi-1)/(\xi+1)$. Then by Lemmas A.9 and A.10 and the referenced assumptions within, we have

$$\|\hat{\beta}^L - \beta^*\|_\infty \leqslant \frac{2\xi\lambda_L}{(\xi+1)\zeta} \implies \|\beta_j^*\| \leqslant \frac{4\xi\lambda_L}{(\xi+1)\zeta} \quad \text{for all } j \in S\backslash\hat{S} \tag{C.8}$$

$$|\hat{S}\backslash S| \leqslant N' \implies |\hat{S}| \leqslant N' + d \lesssim d. \tag{C.9}$$

Denote $\hat{\mathbf{X}} = \begin{bmatrix}\mathbf{X}_{\hat{S}} & \mathbf{Z}\end{bmatrix}$. Under candidate model (4.20), our variance component estimators (via Henderson's Method III) are given by

$$\hat{\sigma}^2 = \frac{y^T\left(\mathbf{I}_{N\times N} - \mathbf{P}_{\hat{\mathbf{X}}}\right)y}{N - \operatorname{rank}(\hat{\mathbf{X}})},$$

$$\hat{\tau}^2 = \frac{y^T\left(\mathbf{P}_{\hat{\mathbf{X}}} - \mathbf{P}_{\mathbf{X}_{\hat{S}}}\right)y - \hat{\sigma}^2[\operatorname{rank}(\hat{\mathbf{X}}) - \operatorname{rank}(\mathbf{X}_{\hat{S}})]}{\operatorname{tr}\left[\mathbf{Z}^T\left(\mathbf{I}_{N\times N} - \mathbf{P}_{\mathbf{X}_{\hat{S}}}\right)\mathbf{Z}\right]}.$$

See (4.25) and (4.26).

Consider the scenario where $|S\backslash\hat{S}| > 0$. We first prove, under the given assumptions, that $|\hat{\sigma}^2 - \sigma^{*2}| = o_P(1)$. Write $S_O = S\backslash\hat{S}$, 'O' for omitted. Then,

$$\begin{aligned}\hat{\sigma}^2 &= \frac{(\hat{\mathbf{X}}\beta^*_{\hat{S}} + \mathbf{X}_{SO}\beta^*_{SO} + \epsilon)^T(\mathbf{I}_{N\times N} - \mathbf{P}_{\hat{\mathbf{X}}})(\hat{\mathbf{X}}\beta_{\hat{S}} + \mathbf{X}_{SO}\beta^*_{SO} + \epsilon)}{N - \operatorname{rank}(\hat{\mathbf{X}})} \\ &= \frac{\beta^{*T}_{SO}\mathbf{X}^T_{SO}(\mathbf{I}_{N\times N} - \mathbf{P}_{\hat{\mathbf{X}}})\mathbf{X}_{SO}\beta^*_{SO}}{N - \operatorname{rank}(\hat{\mathbf{X}})} + \frac{2\beta^{*T}_{SO}\mathbf{X}^T_{SO}(\mathbf{I}_{N\times N} - \mathbf{P}_{\hat{\mathbf{X}}})\epsilon}{N - \operatorname{rank}(\hat{\mathbf{X}})} + \frac{\epsilon^T(\mathbf{I}_{N\times N} - \mathbf{P}_{\hat{\mathbf{X}}})\epsilon}{N - \operatorname{rank}(\hat{\mathbf{X}})}. \end{aligned} \tag{C.10}$$

We proceed to show that the three parts to (C.10) satisfy

$$\left|\frac{\epsilon^T(\mathbf{I}_{N\times N} - \mathbf{P}_{\hat{\mathbf{X}}})\epsilon}{N - \operatorname{rank}(\hat{\mathbf{X}})} - \sigma^{*2}\right| = o_P(1), \tag{C.11}$$

$$\left|\frac{2\beta^{*T}_{SO}\mathbf{X}^T_{SO}(\mathbf{I}_{N\times N} - \mathbf{P}_{\hat{\mathbf{X}}})\epsilon}{N - \operatorname{rank}(\hat{\mathbf{X}})}\right| = o_P(1), \tag{C.12}$$

$$\text{and} \quad \underbrace{\frac{\beta^{*T}_{SO}\mathbf{X}^T_{SO}(\mathbf{I}_{N\times N} - \mathbf{P}_{\hat{\mathbf{X}}})\mathbf{X}_{SO}\beta^*_{SO}}{N - \operatorname{rank}(\hat{\mathbf{X}})}}_{\text{Bias}(\hat{\sigma}^2) = \mathbb{E}[\hat{\sigma}^2] - \sigma^{*2}} = o(1), \tag{C.13}$$

which would suggest that $\hat{\sigma}^2$ is indeed consistent for σ^{*2}.

1. *Proving (C.12)*: Let $\boldsymbol{\Gamma}_{\hat{\mathbf{X}}_\perp}\mathbf{D}_{\hat{\mathbf{X}}_\perp}\boldsymbol{\Gamma}^T_{\hat{\mathbf{X}}_\perp}$ represent the eigendecomposition of $\mathbf{I}_{N\times N} - \mathbf{P}_{\hat{\mathbf{X}}}$. We note that the latter is idempotent, implying that the diagonal matrix $\mathbf{D}_{\hat{\mathbf{X}}_\perp}$, which is of

rank $N - \text{rank}(\hat{\mathbf{X}})$, has only 0 and 1s as its eigenvalues. It is straightforward to show that

$$\begin{aligned}
\mathbb{E}\left[\frac{2\beta_{SO}^{*T}\mathbf{X}_{SO}^{T}(\mathbf{I}_{N\times N}-\mathbf{P}_{\hat{\mathbf{X}}})\epsilon}{N-\text{rank}(\hat{\mathbf{X}})}\right] &= 0 \quad \text{and} \\
\text{Var}\left[\frac{2\beta_{SO}^{*T}\mathbf{X}_{SO}^{T}(\mathbf{I}_{N\times N}-\mathbf{P}_{\hat{\mathbf{X}}})\epsilon}{N-\text{rank}(\hat{\mathbf{X}})}\right] &= \frac{4\sigma^{*2}\beta_{SO}^{*T}\mathbf{X}_{SO}^{T}\Gamma_{\hat{\mathbf{X}}_\perp}\mathbf{D}_{\hat{\mathbf{X}}_\perp}\mathbf{\Gamma}_{\hat{\mathbf{X}}_\perp}^{T}\mathbf{X}_{SO}\beta_{SO}^{*}}{[N-\text{rank}(\hat{\mathbf{X}})]^2} \\
&\leqslant \frac{4\sigma^{*2}\|\Gamma_{\hat{\mathbf{X}}_\perp}^{T}\mathbf{X}_{SO}\beta_{SO}^{*}\|_\infty^2}{N-\text{rank}(\hat{\mathbf{X}})} \\
&\leqslant \frac{4\sigma^{*2}\|\Gamma_{\hat{\mathbf{X}}_\perp}^{T}\mathbf{X}_{SO}\|_\infty^2}{N-\text{rank}(\hat{\mathbf{X}})} \times \frac{d^2 q\log p}{M} = o(1).
\end{aligned}$$

Statement (C.12) then follows from Chebyshev's inequality.

2. *Proving (C.11)*: let $\chi_i^2(1)$, $i = 1, \ldots, N - \text{rank}(\hat{\mathbf{X}})$, be i.i.d random variables following a χ^2 distribution of degrees of freedom 1. Observe that,

$$\frac{\epsilon^T[\mathbf{I}_{N\times N}-\mathbf{P}_{\hat{\mathbf{X}}}]\epsilon}{N-\text{rank}(\hat{\mathbf{X}})} =_d \frac{\epsilon^T\mathbf{D}_{\hat{\mathbf{X}}_\perp}\epsilon}{N-\text{rank}(\hat{\mathbf{X}})} =_d \frac{\sigma^{*2}\sum_{i=1}^{N-\text{rank}(\hat{\mathbf{X}})}\chi_i^2(1)}{N-\text{rank}(\hat{\mathbf{X}})} \tag{C.14}$$

$$\left|\frac{\sigma^{*2}\sum_{i=1}^{N-\text{rank}(\hat{\mathbf{X}})}\chi_i^2(1)}{N-\text{rank}(\hat{\mathbf{X}})} - \sigma^{*2}\right| = o_P(1). \tag{C.15}$$

(C.15) follows from Assumption 3, which implies that $N - \text{rank}(\hat{\mathbf{X}}) \to \infty$ as $M \to \infty$. Applying the Strong Law of Large Numbers (SLLN) for i.i.d. random variables, we arrive at (C.11).

3. *Proving (C.13)*: We observe that

$$\begin{aligned}
\text{Bias}(\hat{\sigma}^2) &\leqslant \|\mathbf{\Gamma}_{\hat{\mathbf{X}}_\perp}^{T}\mathbf{X}_{S_O}\|_\infty^2 \times \|\beta_{S_O}^{*}\|_\infty^2 \times d^2 \\
&\lesssim \frac{d^2 q\log(p)}{M} = o(1),
\end{aligned}$$

and we have completed our proof that $\hat{\sigma}^2$ is consistent under the stated assumptions.

We now demonstrate that the same claim holds for $\hat{\tau}^2$. Expanding out y, we obtain, after some algebraic manipulation,

$$\begin{aligned}\hat{\tau}^2 = &\frac{\beta_{S_O}^{*T}\mathbf{X}_{S_O}^T(\mathbf{P}_{\hat{\mathbf{X}}}-\mathbf{P}_{\mathbf{X}_{\hat{S}}})\mathbf{X}_{S_O}\beta_{S_O}^*}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}+\frac{2\beta_{S_O}^{*T}\mathbf{X}_{S_O}^T(\mathbf{P}_{\hat{\mathbf{X}}}-\mathbf{P}_{\mathbf{X}_{\hat{S}}})\epsilon}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}+\frac{\epsilon^T(\mathbf{P}_{\hat{\mathbf{X}}}-\mathbf{P}_{\mathbf{X}_{\hat{S}}})\epsilon}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}\\ &+\frac{2\beta_{S_O}^{*T}\mathbf{X}_{S_O}^T(\mathbf{I}_{N\times N}-\mathbf{P}_{\mathbf{X}_{\hat{S}}})\mathbf{Z}v}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}+\frac{v^T\mathbf{Z}'^T\mathbf{Z}'v}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}+\frac{2v^T\mathbf{Z}^T(\mathbf{I}_{N\times N}-\mathbf{P}_{\mathbf{X}_{\hat{S}}})\epsilon}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}\\ &-\frac{\sigma^{*2}[\mathrm{rank}(\hat{\mathbf{X}})-\mathrm{rank}(\mathbf{X}_{\hat{S}})]}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}-\frac{\mathrm{Bias}(\hat{\sigma}^2)[\mathrm{rank}(\hat{\mathbf{X}})-\mathrm{rank}(\mathbf{X}_{\hat{S}})]}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}.\end{aligned} \tag{C.16}$$

where we have defined $\mathbf{Z}' = (\mathbf{I}_{N\times N}-\mathbf{P}_{\mathbf{X}_{\hat{S}}})\mathbf{Z}$. We set out to prove that the terms in (C.16) satisfy

$$\left|\frac{2\beta_{S_O}^{*T}\mathbf{X}_{S_O}^T(\mathbf{P}_{\hat{\mathbf{X}}}-\mathbf{P}_{\mathbf{X}_{\hat{S}}})\epsilon}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}\right| = o_P(1) \tag{C.17}$$

$$\left|\frac{\epsilon^T(\mathbf{P}_{\hat{\mathbf{X}}}-\mathbf{P}_{\mathbf{X}_{\hat{S}}})\epsilon}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}-\frac{\sigma^{*2}[\mathrm{rank}(\hat{\mathbf{X}})-\mathrm{rank}(\mathbf{X}_{\hat{S}})]}{\mathrm{tr}\,[\mathbf{Z}'^T\mathbf{Z}']}\right| = o_P(1) \tag{C.18}$$

$$\left|\frac{2\beta_{S_O}^{*T}\mathbf{X}_{S_O}^T(\mathbf{I}_{N\times N}-\mathbf{P}_{\mathbf{X}_{\hat{S}}})\mathbf{Z}v}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}\right| = o_P(1) \tag{C.19}$$

$$\left|\frac{v^T\mathbf{Z}^T(\mathbf{I}_{N\times N}-\mathbf{P}_{\mathbf{X}_{\hat{S}}})\mathbf{Z}v}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}-\tau^{*2}\right| = o_P(1) \tag{C.20}$$

$$\left|\frac{2v^T\mathbf{Z}^T(\mathbf{I}_{N\times N}-\mathbf{P}_{\mathbf{X}_{\hat{S}}})\epsilon}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}\right| = o_P(1) \tag{C.21}$$

$$\text{and}\quad \underbrace{\frac{\beta_{S_O}^{*T}\mathbf{X}_{S_O}^T(\mathbf{P}_{\hat{\mathbf{X}}}-\mathbf{P}_{\mathbf{X}_{\hat{S}}})\mathbf{X}_{S_O}\beta_{S_O}^*}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}-\frac{\mathrm{Bias}(\hat{\sigma}^2)[\mathrm{rank}(\hat{\mathbf{X}})-\mathrm{rank}(\mathbf{X}_{\hat{S}})]}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}}_{\mathrm{Bias}[\hat{\tau}^2]=\mathbb{E}[\hat{\tau}^2]-\tau^{*2}} = o(1). \tag{C.22}$$

Let $\mathbf{Q}_{\mathbf{Z}'}\mathbf{D}_{\mathbf{Z}'}\mathbf{\Gamma}_{\mathbf{Z}'}^T$ represent the singular value decomposition of $\mathbf{Z}'$, with $\mathbf{Q}_{\mathbf{Z}'}$ and $\mathbf{\Gamma}_{\mathbf{Z}'}$ of dimensions $N\times qM$ and $qM\times qM$, respectively. Additionally, write the eigendecompositions of $\mathbf{P}_{\hat{\mathbf{X}}}-\mathbf{P}_{\mathbf{X}_{\hat{S}}}$ and $\mathbf{I}_{N\times N}-\mathbf{P}_{\mathbf{X}_{\hat{S}}}$ as $\mathbf{\Gamma}_{\hat{\mathbf{X}}\perp\mathbf{X}_{\hat{S}}}\mathbf{D}_{\hat{\mathbf{X}}\perp\mathbf{X}_{\hat{S}}}\mathbf{\Gamma}_{\hat{\mathbf{X}}\perp\mathbf{X}_{\hat{S}}}^T$ and $\mathbf{\Gamma}_{\mathbf{X}_{\hat{S}}\perp}\mathbf{D}_{\mathbf{X}_{\hat{S}}\perp}\mathbf{\Gamma}_{\mathbf{X}_{\hat{S}}\perp}^T$, respectively.

To avoid repetition, some of the proofs are presented in abbreviated form.

1. *Proving (C.17)*: Clearly,

$$\begin{aligned}
\mathbb{E}\left[\frac{2\beta_{S_O}^{*T}\mathbf{X}_{S_O}^T(\mathbf{P}_{\hat{\mathbf{X}}}-\mathbf{P}_{\mathbf{X}_{\hat{S}}})\epsilon}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}\right] &= 0 \\
\mathrm{Var}\left[\frac{2\beta_{S_O}^{*T}\mathbf{X}_{S_O}^T(\mathbf{P}_{\hat{\mathbf{X}}}-\mathbf{P}_{\mathbf{X}_{\hat{S}}})\epsilon}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}\right] &= \frac{4\beta_{S_O}^{*T}\mathbf{X}_{S_O}^T(\mathbf{P}_{\hat{\mathbf{X}}}-\mathbf{P}_{\mathbf{X}_{\hat{S}}})\mathbf{X}_{S_O}\beta_{S_O}^*}{\mathrm{tr}\left[\mathbf{Z}'^T\mathbf{Z}'\right]^2} \\
&= \frac{4\beta_{S_O}^{*T}\mathbf{X}_{S_O}^T\boldsymbol{\Gamma}_{\hat{\mathbf{X}}\perp\mathbf{X}_{\hat{S}}}\mathbf{D}_{\hat{\mathbf{X}}\perp\mathbf{X}_{\hat{S}}}\boldsymbol{\Gamma}_{\hat{\mathbf{X}}\perp\mathbf{X}_{\hat{S}}}^T\mathbf{X}_{S_O}\beta_{S_O}^*}{\mathrm{tr}\left[\mathbf{Z}'^T\mathbf{Z}'\right]^2} \\
&\leqslant \frac{4\times d^2\times qM\times\|\boldsymbol{\Gamma}_{\hat{\mathbf{X}}\perp\mathbf{X}_{\hat{S}}}\mathbf{X}_{S_O}\|_\infty\times\|\beta_{S_O}^*\|_\infty^2}{\mathrm{tr}\left[\mathbf{Z}'^T\mathbf{Z}'\right]^2} \\
&= o(1),
\end{aligned}$$

   following from (4.33) in Assumption 4 and Assumption 5.

2. *Proving (C.18)*: Orthogonality of $\boldsymbol{\Gamma}_{\hat{\mathbf{X}}\perp\mathbf{X}_{\hat{S}}}$ implies that $\epsilon^T(\mathbf{P}_{\hat{\mathbf{X}}}-\mathbf{P}_{\mathbf{X}_{\hat{S}}})\epsilon =_d \epsilon^T\mathbf{D}_{\hat{\mathbf{X}}\perp\mathbf{X}_{\hat{S}}}\epsilon$, so

$$\mathbb{E}\left[\frac{\epsilon^T(\mathbf{P}_{\hat{\mathbf{X}}}-\mathbf{P}_{\mathbf{X}_{\hat{S}}})\epsilon}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}\right] = \mathbb{E}\left[\frac{\epsilon^T\mathbf{D}_{\hat{\mathbf{X}}\perp\mathbf{X}_{\hat{S}}}\epsilon}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}\right] = \frac{\sigma^{*2}[\mathrm{rank}(\hat{\mathbf{X}})-\mathrm{rank}(\mathbf{X}_{\hat{S}})]}{\mathrm{tr}\left[\mathbf{Z}'^T\mathbf{Z}'\right]}$$

   and, using properties of quadratic forms, we have

$$\mathrm{Var}\left[\frac{\epsilon^T(\mathbf{P}_{\hat{\mathbf{X}}}-\mathbf{P}_{\mathbf{X}_{\hat{S}}})\epsilon}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}\right] = \mathrm{Var}\left[\frac{\epsilon^T\mathbf{D}_{\hat{\mathbf{X}}\perp\mathbf{X}_{\hat{S}}}\epsilon}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}\right] = \frac{2\sigma^{*2}\mathrm{tr}(\mathbf{D}_{\hat{\mathbf{X}}})}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')^2} \lesssim \frac{1}{M} = o(1),$$

   the latter relation following from (4.34) in Assumption 4. This proves (C.18).

3. *Proving (C.19)*: Proof is similar to that of (C.17), as we note that

$$\begin{aligned}
\mathbb{E}\left[\frac{2\beta_{S_O}^{*T}\mathbf{X}_{S_O}^T(\mathbf{I}_{N\times N}-\mathbf{P}_{\mathbf{X}_{\hat{S}}})\mathbf{Z}v}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}\right] &= 0 \\
\mathrm{Var}\left[\frac{2\beta_{S_O}^{*T}\mathbf{X}_{S_O}^T(\mathbf{I}_{N\times N}-\mathbf{P}_{\mathbf{X}_{\hat{S}}})\mathbf{Z}v}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}\right] &= \frac{4\tau^{*2}\beta_{S_O}^{*T}\mathbf{X}_{S_O}^T\mathbf{Q}_{\mathbf{Z}'}\mathbf{D}_{\mathbf{Z}'}^2\mathbf{Q}_{\mathbf{Z}'}^T\mathbf{X}_{S_O}\beta_{S_O}^*}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')^2} \\
&\leqslant \frac{4\tau^{*2}\mathrm{tr}(\mathbf{Z}'^T\mathbf{Z})\|\mathbf{Q}_{\mathbf{Z}'}^T\mathbf{X}_{S_O}\beta_{S_O}^*\|_\infty^2}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')^2} \\
&= o(1),
\end{aligned}$$

   having applied Assumptions 4 and 5 here.

4. *Proving (C.20)*: Similar to as in (C.18), and again using properties of quadratic forms, we can show that

$$\mathbb{E}\left[\frac{v^T\mathbf{Z}^T(\mathbf{I}_{N\times N}-\mathbf{P}_{\mathbf{X}_{\hat{S}}})\mathbf{Z}v}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}\right]=\tau^{*2}$$
$$\mathrm{Var}\left[\frac{v^T\mathbf{Z}^T(\mathbf{I}_{N\times N}-\mathbf{P}_{\mathbf{X}_{\hat{S}}})\mathbf{Z}v}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}\right]=\frac{2\tau^{*4}\mathrm{tr}(\mathbf{D}^4_{\mathbf{Z}'})}{\mathrm{tr}(\mathbf{D}^2_{\mathbf{Z}'})^2}=o(1),$$

the last relation the result of (4.33) from Assumption 4.

5. *Proving (C.21)*: We can rewrite

$$\frac{2v^T\mathbf{Z}'^T\epsilon}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}=\frac{2v^T\mathbf{\Gamma}_{\mathbf{Z}'}\mathbf{D}_{\mathbf{Z}'}\mathbf{Q}^T_{\mathbf{Z}'}\epsilon}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}=_d\frac{\sigma^*\tau^*\sum_{i=1}^{qM}s_iB_i}{\sum_{i=1}^{qM}s_i^2}.$$

where B_i, $i=1,\dots\mathrm{rank}(\mathbf{Z}')$ are random variables formed as the product of two independent $N(0,1)$ random variables. Then,

$$\mathrm{Var}\left(\sum_{i=1}^{qM}s_iB_i\right)=\sum_{i=1}^{qM}s_i^2,$$

which implies that

$$\mathrm{Var}\left(\frac{\sigma^*\tau^*\sum_{i=1}^{qM}s_iB_i}{\sum_{i=1}^{qM}s_i^2}\right)=\frac{1}{\sum_{i=1}^{qM}s_i^2}$$

and since $\sum_{i=1}^{qM}s_i^2\asymp qM$ by (4.33) in Assumption 4, we have proven our claim (C.21).

6. *Proving (C.22)*: By the definition of $\mathrm{Bias}[\hat{\tau}^2]$, it is clear that

$$\begin{aligned}|\mathrm{Bias}(\hat{\tau}^2)|&\leqslant\frac{\beta^{*T}_{S_O}\mathbf{X}^T_{S_O}(\mathbf{P}_{\hat{\mathbf{X}}}-\mathbf{P}_{\mathbf{X}_{\hat{S}}})\mathbf{X}_{S_O}\beta^*_{S_O}}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}+\frac{\mathrm{Bias}(\hat{\sigma}^2)[\mathrm{rank}(\hat{\mathbf{X}})-\mathrm{rank}(\mathbf{X}_{\hat{S}})]}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}\\&\leqslant\frac{\beta^{*T}_{S_O}\mathbf{X}^T_{S_O}\mathbf{P}_{\mathbf{Z}}\mathbf{X}_{S_O}\beta^*_{S_O}}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}+\frac{qM\times\mathrm{Bias}(\hat{\sigma}^2)}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}\\&\leqslant\frac{qM\times d\times\|\mathbf{\Gamma}_{\mathbf{Z}}\mathbf{X}_{S^O}\|^2_\infty\times\|\beta^*_{S_O}\|^2_\infty}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}+\frac{qM\times\mathrm{Bias}(\hat{\sigma}^2)}{\mathrm{tr}\,(\mathbf{Z}'^T\mathbf{Z}')}\\&\lesssim\frac{d^2q\log(p)}{M}=o(1),\end{aligned}$$

where the second last relation follows from the proven claim that $\mathrm{Bias}(\hat{\sigma}^2)$ is $o(1)$, (4.33) in Assumption 4 and Assumption 5.

Since the event $u^* \leqslant \lambda_L(\xi-1)/(\xi+1)$ occurs with probability greater than $1-1/p \to 1$ as $p \to \infty$,

$$|\hat{\sigma}^2 - \sigma^{*2}| = o_P(1)$$
$$|\hat{\tau}^2 - \tau^{*2}| = o_P(1)$$

as claimed.

□

Proof of Theorem 8. Suppose that $u^* \leqslant \lambda_L(\xi-1)/(\xi+1)$, and write $S_O = S \backslash \hat{S}$. The OLS fit $\hat{\beta}^{\text{init}}$ has a simple closed-form expression:

$$\begin{aligned}\hat{\beta}^{\text{init}}_{\hat{S}} &= (\mathbf{X}_{\hat{S}}^T \mathbf{X}_{\hat{S}})^{-1} \mathbf{X}_{\hat{S}}^T y \\ &= \beta^*_{\hat{S}} + (\mathbf{X}_{\hat{S}}^T \mathbf{X}_{\hat{S}})^{-1} \mathbf{X}_{\hat{S}}^T \mathbf{X}_{S_O} \beta^*_{S_O} + (\mathbf{X}_{\hat{S}}^T \mathbf{X}_{\hat{S}})^{-1} \mathbf{X}_{\hat{S}}^T \epsilon.\end{aligned}$$

and $\hat{\beta}^{\text{init}}_{\hat{S}^c} = 0$. Thus, by triangle inequality,

$$\|\hat{\beta}^{\text{init}}_{\hat{S}} - \beta^*_{\hat{S}}\|_1 \leqslant \|(\mathbf{X}_{\hat{S}}^T \mathbf{X}_{\hat{S}})^{-1} \mathbf{X}_{\hat{S}}^T \mathbf{X}_{S_O} \beta^*_{S_O}\|_1 + \|(\mathbf{X}_{\hat{S}}^T \mathbf{X}_{\hat{S}})^{-1} \mathbf{X}_{\hat{S}}^T \epsilon\|_1. \tag{C.23}$$

We proceed by first bounding the first term on the right-hand side of (C.23). By Assumption 6,

$$\begin{aligned}\|(\mathbf{X}_{\hat{S}}^T \mathbf{X}_{\hat{S}})^{-1} \mathbf{X}_{\hat{S}}^T\|_2 &\leqslant \sqrt{\nu_{\max}[(\mathbf{X}_{\hat{S}}^T \mathbf{X}_{\hat{S}})^{-1} \mathbf{X}_{\hat{S}}^T \mathbf{X}_{\hat{S}} (\mathbf{X}_{\hat{S}}^T \mathbf{X}_{\hat{S}})^{-1}]} \\ &= \sqrt{\nu_{\max}[(\mathbf{X}_{\hat{S}}^T \mathbf{X}_{\hat{S}})^{-1}]} \\ &= \sqrt{\frac{1}{N} \nu_{\max}\left[(\hat{\mathbf{\Sigma}}_{\hat{\mathbf{S}}\hat{\mathbf{S}}})^{-1}\right]} \\ &= \sqrt{\frac{1}{N} \nu_{\min}\left[\hat{\mathbf{\Sigma}}_{\hat{\mathbf{S}}\hat{\mathbf{S}}}\right]} \leqslant \frac{1}{\sqrt{N \kappa_-(N', S)}} \lesssim \frac{1}{\sqrt{N}}\end{aligned}$$

This, in turn, implies that

$$\begin{aligned}\|(\mathbf{X}_{\hat{S}}^T\mathbf{X}_{\hat{S}})^{-1}\mathbf{X}_{\hat{S}}^T\mathbf{X}_{S_O}\beta^*_{S_O}\|_1 &\leqslant \sqrt{N'+d}\|(\mathbf{X}_{\hat{S}}^T\mathbf{X}_{\hat{S}})^{-1}\mathbf{X}_{\hat{S}}^T\mathbf{X}_{S_O}\beta^*_{S_O}\|_2 \\ &\leqslant \sqrt{N'+d}\|(\mathbf{X}_{\hat{S}}^T\mathbf{X}_{\hat{S}})^{-1}\mathbf{X}_{\hat{S}}^T\|_2\|\mathbf{X}_{S_O}\beta^*_{S_O}\|_2 \\ &\leqslant \sqrt{N'+d}\|(\mathbf{X}_{\tilde{S}}^T\mathbf{X}_{\tilde{S}})^{-1}\mathbf{X}_{\tilde{S}}^T\|_2\|\mathbf{X}_O\|_F\|\beta^*_O\|_\infty \\ &\leqslant \sqrt{\frac{(N'+|S)d}{\kappa_-(N',S)}}\frac{2\xi\lambda_L}{(\xi+1)\zeta} \lesssim \sqrt{\frac{d^2q\log(p)}{M}} = o(1),\end{aligned}$$

where the last relation follows from Assumption 3.

We proceed to bound the second component on the right-hand side of (C.23). We observe that

$$\begin{aligned}\|(\mathbf{X}_{\tilde{S}}^T\mathbf{X}_{\tilde{S}})^{-1}\mathbf{X}_{\tilde{S}}^T\epsilon\|_1 &\leqslant \sqrt{N'+d}\|(\mathbf{X}_{\tilde{S}}^T\mathbf{X}_{\tilde{S}})^{-1}\mathbf{X}_{\tilde{S}}^T\epsilon\|_2 \\ &= \sqrt{N'+d}\left\|(\hat{\mathbf{\Sigma}}_{\tilde{S}\tilde{S}})^{-1}\frac{\mathbf{X}_{\tilde{S}}^T\epsilon}{N}\right\|_2 \\ &\leqslant (N'+d)\left|\left|\left|(\hat{\mathbf{\Sigma}}_{\tilde{S}\tilde{S}})^{-1}\right|\right|\right|_2\left\|\frac{\mathbf{X}_{\tilde{S}}^T\epsilon}{N}\right\|_2 \\ &\leqslant (N'+d)\left|\left|\left|(\hat{\mathbf{\Sigma}}_{\tilde{S}\tilde{S}})^{-1}\right|\right|\right|_2\left\|\frac{\mathbf{X}_{\tilde{S}}^T\epsilon}{N}\right\|_\infty \\ &\leqslant \frac{N'+d}{\kappa_-(N',S)}\lambda_L \lesssim \sqrt{\frac{d^2q\log(p)}{M}} = o(1).\end{aligned}$$

From Lemma A.9,

$$\|\hat{\beta}^{\text{init}}_{S_O} - \beta^*_{S_O}\|_\infty = \|\beta^*_{S_O}\|_\infty \lesssim \sqrt{\frac{q\log(p)}{M}},$$

which implies that

$$\|\hat{\beta}^{\text{init}}_{S_O} - \beta^*_{S_O}\|_1 \lesssim \sqrt{\frac{d^2q\log(p)}{M}} = o(1).$$

By Lemma A.8, the event $u^* \leqslant \lambda_L(\xi-1)/(\xi+1)$ occurs with probability exceeding $1-1/p$. Combined, we obtain the desired result.

□

VITA

Lina Lin pursued a PhD in Statistics at the University of Washington, Seattle, after completing a MS in Statistics and a BASc in Engineering Science at the University of Toronto. She originally hails from Thornhill, Ontario, Canada, and maintains some Canadian habits. In her free time, she enjoys reading, drawing, playing card games/board games/video games, and travelling.