Journal of Machine Learning Research 19 (2018) 1-59 Submitted 9/15; Revised 10/18; Published 10/18

On Semiparametric Exponential Family Graphical Models

Zhuoran Yang ZY6@PRINCETON.EDU
Department of Operations Research and Financial Engineering
Princeton University
Princeton, NJ 08544, USA

Yang Ning YN265@CORNELL.EDU
Department of Statistical Science
Cornell University
Ithaca, NY 14853, USA

Han Liu HANLIU@NORTHWESTERN.EDU
Department of Electrical Engineering and Computer Science
Northwestern University
Evanston, IL 60208, USA

Editor: David Blei

Abstract

We propose a new class of semiparametric exponential family graphical models for the analysis of high dimensional mixed data. Different from the existing mixed graphical models, we allow the nodewise conditional distributions to be semiparametric generalized linear models with unspecified base measure functions. Thus, one advantage of our method is that it is unnecessary to specify the type of each node and the method is more convenient to apply in practice. Under the proposed model, we consider both problems of parameter estimation and hypothesis testing in high dimensions. In particular, we propose a symmetric pairwise score test for the presence of a single edge in the graph. Compared to the existing methods for hypothesis tests, our approach takes into account of the symmetry of the parameters, such that the inferential results are invariant with respect to the different parametrizations of the same edge. Thorough numerical simulations and a real data example are provided to back up our theoretical results.

Keywords: Graphical Models, Exponential Family, High Dimensional Inference

1. Introduction

Given a d-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^T$, inferring the conditional independence among $\boldsymbol{X}$ and quantifying its uncertainty are important tasks in statistics. We propose a unified framework for modeling, estimation, and uncertainty assessment for a new type of graphical model, named as semiparametric exponential family graphical model. Let $G = (V, E)$ be an undirected graph with node set $V = \{1, 2, \ldots, d\}$ and edge set $E \subseteq \{(j,k)\colon 1 \le j < k \le d\}$. The semiparametric exponential family graphical model specifies the joint distribution of $\boldsymbol{X}$ such that for each $j \in V$, the conditional distribution of X_j given $\boldsymbol{X}_{\setminus j} := (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_d)^T$ is of the form

$$p(x_j \mid \boldsymbol{x}_{\setminus j}) = \exp\bigl[\eta_j(\boldsymbol{x}_{\setminus j}) \cdot x_j + f_j(x_j) - b_j(\eta_j, f_j)\bigr], \tag{1}$$

where $\boldsymbol{x}_{\backslash j} = (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_d)$, $\eta_j(\boldsymbol{x}_{\backslash j}) = \alpha_j + \sum_{k \neq j} \beta_{jk} x_k$ is the canonical parameter, $f_j(\cdot)$ is an unknown base measure function, and $b_j(\cdot, \cdot)$ is the log-partition function. Besides, we assume $\beta_{jk} = \beta_{kj}$ for all $j \neq k$. By definition, the unknown parameter contains $\{(\alpha_j, \beta_{jk}, f_j) \colon 1 \leq j < k \leq d\}$. To make the model identifiable, we set $\alpha_j = 0$ and absorb the term $\alpha_j x_j$ into $f_j(x_j)$. By the Hammersley-Clifford theorem (Besag, 1974), we have $\beta_{jk} \neq 0$ if and only if X_j and X_k are conditionally independent given $\{X_\ell \colon \ell \neq j, k\}$. Therefore, we set $(j,k) \in E$ if and only if $\beta_{jk} \neq 0$. The graph G thus characterizes the conditional independence relationship among the high dimensional distribution of $\boldsymbol{X}$. The key feature of the proposed model is that (1) it is a general semiparametric model and (2) it can be used to handle mixed data, which means that $\boldsymbol{X}$ may contain both continuous and discrete random variables. Unlike the existing mixed graphical models, we allow the nodewise conditional distributions to be semiparametric generalized linear models with unspecified base measure functions. Thus, our method does not need to specify the type of each node and is more convenient to apply in practice. In addition to the proposed new model, our paper has the following two novel contributions.

First, for the purpose of estimating β_{jk}, we extend the multistage relaxation algorithm (Zhang, 2010) and conduct a localized analysis for a more sophisticated loss function obtained by a statistical chromatography method (Liang and Qin, 2000; Diao et al., 2012; Chan, 2012; Ning et al., 2017b). The gradient and Hessian matrix of the loss function are nonlinear U-statistics with unbounded kernel functions. This makes our technical analysis more challenging than that in Zhang (2010). Under the assumption that the sparse eigenvalue condition holds locally, we prove the same optimal statistical rates for parameter estimation as in high dimensional linear models.

Second, we propose a symmetric pairwise score test for the null hypothesis $H_0 \colon \beta_{jk} = 0$. This is equivalent to testing whether X_j and X_k are conditionally independent given $\{X_\ell \colon \ell \neq j, k\}$. Compared with Ning et al. (2017b), the novelty of our method is that we consider a more sophisticated cross type inference which incorporates the symmetry of the parameter, i.e., $\beta_{jk} = \beta_{kj}$. By considering this unique structure of the graphical model, our proposed method achieves the invariance property of the inferential results. That means the same p-values are obtained for testing $\beta_{jk} = 0$ and $\beta_{kj} = 0$. In contrast, the asymmetric method in Ning et al. (2017b) may lead to different conclusions for testing these two equivalent null hypotheses.

1.1. Related Works

There is a huge literature on estimating undirected graphical models (Lauritzen, 1996; Edwards, 2000; Whittaker, 2009). For modeling continuous data, the most commonly used methods are Gaussian graphical models (Yuan and Lin, 2007; Banerjee et al., 2008; Friedman et al., 2008; Ravikumar et al., 2011; Rothman et al., 2008; Lam and Fan, 2009; Shen et al., 2012; Yuan, 2010; Cai et al., 2011; Sun and Zhang, 2013; Guo et al., 2011; Danaher et al., 2014; Mohan et al., 2014; Meinshausen and Bühlmann, 2006; Peng et al., 2009; Friedman et al., 2010). To relax the Gaussian assumption, Liu et al. (2009); Xue et al. (2012b); Liu et al. (2012); Ning and Liu (2013) propose the Gaussian copula model and Voorman et al. (2014) study the joint additive models for graph estimation. For modeling binary data, the Ising graphical model is considered by Lee et al. (2006); Höfling and Tibshirani (2009);

Ravikumar et al. (2010); Xue et al. (2012a); Cheng et al. (2014). In addition to binary data, Allen and Liu (2012) and Yang et al. (2013b) consider the Poisson data and Guo et al. (2015) consider the ordinal data. Moreover, Yang et al. (2013a) propose exponential family graphical models, and Tan et al. (2014) propose a general framework for graphical models with hubs.

Recently, modeling the mixed data attracts increasing interests (Lee and Hastie, 2015; Fellinghauer et al., 2013; Cheng et al., 2017; Chen et al., 2015; Fan et al., 2017; Yang et al., 2014). Compared with Lee and Hastie (2015); Cheng et al. (2017); Chen et al. (2015); Yang et al. (2014), our model has the following two main advantages. First, it is a semiparametric model, which does not need to specify the parametric conditional distribution for each node. Therefore, it provides a more flexible modeling framework than the existing ones. Second, under our proposed model, the estimation and inference methods are easier to implement. Unlike these existing methods, we propose a unified estimation and inference procedure, which does not need to distinguish whether the node satisfies the Gaussian distribution or the Bernoulli distribution. In addition, our estimation and inference methods are more efficient than the nonparametric approach in Fellinghauer et al. (2013). Finally, our method is more convenient for modeling the count data than the latent Gaussian copula approach in Fan et al. (2017).

Though significant progress has been made towards developing new graph estimation procedures, the research on uncertainty assessment of the estimated graph lags behind. In low dimensions, Drton et al. (2007); Drton and Perlman (2008) establish confidence subgraph of Gaussian graphical models. In high dimensions, Ren et al. (2015); Janková and van de Geer (2015); Gu et al. (2015) study the confidence interval for a single edge under Gaussian (copula) graphical models and Liu et al. (2013) study the false discovery rate control. However, all these methods rely on the Gaussian or sub-Gaussian assumption and cannot be easily applied to the discrete data and more generally the mixed data in high dimensions.

1.2. Notation

We adopt the following notation throughout this paper. For any vector $\mathbf{v} = (v_1, \ldots, v_d)^T \in \mathbb{R}^d$, we define its support as $\mathrm{supp}(\mathbf{v}) = \{t \colon v_t \neq 0\}$. We define its ℓ_0-norm, ℓ_p-norm, and ℓ_∞-norm as $\|\mathbf{v}\|_0 = |\mathrm{supp}(\mathbf{v})|$, $\|\mathbf{v}\|_p = (\sum_{j \in [d]} |v_j|^p)^{1/p}$ and $\|\mathbf{v}\|_\infty = \max_{j \in [d]} |v_j|$, respectively, where $p > 1$. Let $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}^T$ be the Kronecker product of a vector $\mathbf{v}$ and itself. We write $\mathbf{v} \circ \mathbf{u} = (v_1 u_1, \ldots, v_d u_d)^T$ as the Hadamard product of two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$. In addition, we use $|\mathbf{v}| = (|v_1|, \ldots, |v_d|)^T$ to denote the elementwise absolute value of vector $\mathbf{v}$ and define $\|\mathbf{v}\|_{\min} = \min_{j \in [d]} |v_j|$. For any matrix $\mathbf{A} = [a_{jk}] \in \mathbb{R}^{d_1 \times d_2}$, let $\mathbf{A}_{S_1 S_2} = [a_{jk}]_{j \in S_1, k \in S_2}$ be the submatrix of $\mathbf{A}$ with indices in $S_1 \times S_2$; let $\mathbf{A}_{j \setminus j} = [a_{jk}]_{k \neq j}$. Besides, let $\|\mathbf{A}\|_2$, $\|\mathbf{A}\|_1$, $\|\mathbf{A}\|_\infty$, $\|\mathbf{A}\|_{\ell_p}$ be the spectral norm, elementwise ℓ_1-norm, elementwise ℓ_∞-norm, and operator ℓ_p-norm of $\mathbf{A}$, respectively. Furthermore, for two matrices $\mathbf{A}_1$ and $\mathbf{A}_2$, we write $\mathbf{A}_1 \preceq \mathbf{A}_2$ if $\mathbf{A}_2 - \mathbf{A}_1$ is positive semidefinite and write $\mathbf{A}_1 \leq \mathbf{A}_2$ if every entry of $\mathbf{A}_2 - \mathbf{A}_1$ is nonnegative. For a function $f(\boldsymbol{x}) \colon \mathbb{R}^d \to \mathbb{R}$, we write $\nabla f(\boldsymbol{x})$, $\nabla_S f(\boldsymbol{x})$, $\nabla^2 f(\boldsymbol{x})$ and $\partial f(\boldsymbol{x})$ as the gradient of $f(\boldsymbol{x})$, the gradient of $f(\boldsymbol{x})$ with respect to $\boldsymbol{x}_S$, the Hessian of $f(\boldsymbol{x})$, and the subgradient of $f(\boldsymbol{x})$, respectively. Moreover, we write $\{1, 2, \ldots, d\}$ as $[d]$. For a sequence of random vectors $\{\boldsymbol{Y}_i\}_{i \geq 1}$ and a random vector $\boldsymbol{Y}$, we write $\boldsymbol{Y}_i \rightsquigarrow \boldsymbol{Y}$ if $\{\boldsymbol{Y}_i\}_{i \geq 1}$ converges to $\boldsymbol{Y}$ in distribution.

Finally, for functions $f(n)$ and $g(n)$, we write $f(n) \lesssim g(n)$ to denote that $f(n) \leq cg(n)$ for a universal constant $c \in (0,+\infty)$ and we write $f(n) \asymp g(n)$ when $f(n) \lesssim g(n)$ and $g(n) \lesssim f(n)$ hold simultaneously.

1.3. Paper Organization

The rest of this paper is organized as follows. In §2 we introduce the semiparametric exponential family graphical models. In §3 we present our methods for graph estimation and uncertainty assessment. In §4 we lay out the assumptions and main theoretical results. We study the finite-sample performance of our method on both simulated and real-world datasets in §5 and conclude the paper in §6 with some discussion.

2. Semiparametric Exponential Family Graphical Models

The semiparametric exponential family graphical models are defined by specifying the conditional distribution of each variable X_j given the rest of the variables $\{X_k \colon k \neq j\}$.

Definition 1 (Semiparametric exponential family graphical model) *A d-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^T \in \mathbb{R}^d$ follows a semiparametric exponential graphical model with graph $G = (V, E)$ if for any node $j \in V$, the conditional density of X_j given $\boldsymbol{X}_{\backslash j}$ satisfies*

$$p(x_j \,|\, \boldsymbol{x}_{\backslash j}) = \exp\bigl[x_j(\boldsymbol{\beta}_j^T \boldsymbol{x}_{\backslash j}) + f_j(x_j) - b_j(\boldsymbol{\beta}_j, f_j)\bigr], \tag{2}$$

where $f_j(\cdot)$ is an unknown base measure function and $b_j(\cdot,\cdot)$ is a known log-partition function. In particular, $(j,k) \in E$ if and only if $\beta_{jk} \neq 0$.

This model is semiparametric since we treat both $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jj-1}, \beta_{jj+1}, \ldots, \beta_{jd})^T \in \mathbb{R}^{d-1}$ and the univariate function $f_j(\cdot)$ as parameters, where $\boldsymbol{\beta}_j$ and $f_j(\cdot)$ are the parametric and nonparametric components, respectively. Because the model in Definition 1 is only specified by the conditional distributions of each variable, it is important to understand the conditions under which a valid joint distribution of $\boldsymbol{X}$ exists. This problem has been addressed by Chen et al. (2015). As shown in their Proposition 1, one sufficient condition for the existence of joint distribution of $\boldsymbol{X}$ is that, (i) $\beta_{jk} = \beta_{kj}$ for $1 \leq j, k \leq d$ and (ii) $g(\boldsymbol{x}) := \exp\bigl[\sum_{j<k} \beta_{jk} x_j x_k + \sum_{j=1}^d f_j(x_j)\bigr]$ is integrable.

Hereafter, we assume that the above two conditions hold. Thus, there exists a joint probability distribution for the model defined in (2), whose density has the form of

$$p(\boldsymbol{x}) = \exp\biggl[\sum_{k<\ell} \beta_{k\ell} x_k x_\ell + \sum_{j=1}^d f_j(x_j) - A\bigl(\{\boldsymbol{\beta}_i, f_i\}_{i\in[d]}\bigr)\biggr], \tag{3}$$

where $\beta_{k\ell} \neq 0$ if and only if $(k,\ell) \in E$. Here $A(\cdot)$ is the log-partition function given by

$$A\bigl(\{\boldsymbol{\beta}_i, f_i\}_{i\in[d]}\bigr) := \log\biggl\{\int_{\mathbb{R}^d} \exp\biggl[\sum_{k<\ell} \beta_{k\ell} x_k x_\ell + \sum_{j=1}^d f_j(x_j)\biggr] \nu(\mathrm{d}\boldsymbol{x})\biggr\}, \tag{4}$$

where $\nu(\cdot)$ is a product measure satisfying $\nu(\mathrm{d}\boldsymbol{x}) = \prod_{j\in[d]} \nu_j(\mathrm{d}x_j)$, and each ν_j is either a Lebesgue or a counting measure on the domain of X_j, depending whether X_j is discrete or continuous. Since $\beta_{k\ell} = \beta_{\ell k}$ for all pairs of nodes (k, ℓ), in the sequel, we will use $\beta_{k\ell}$ and $\beta_{\ell k}$ interchangeably for notational simplicity.

Furthermore, we remark that, without the knowledge of $\{f_j\}_{j\in[d]}$, estimating parameters $\{\boldsymbol{\beta}_j\}_{j\in[d]}$ is insufficient to learn the distribution of $\boldsymbol{X}$. In this paper, we focus on the statistical inference of the underlying conditional independence graph specified by $\{\boldsymbol{\beta}_j\}_{j\in[d]}$. In the next section, by adopting a loss function for $\{\boldsymbol{\beta}_j\}_{j\in[d]}$ that is free of the base measures, we obtain estimators of these parameters, which are used to construct an estimator of the underlying graph. Moreover, by further considering the hypothesis testing problem for each β_{jk}, we are able to assess the uncertainty of the estimated graph.

2.1. Examples

We provide some widely used parametric examples in the class of semiparametric exponential family graphical models.

Gaussian Graphical Models: The Gaussian graphical models assume that $\boldsymbol{X} \in \mathbb{R}^d$ follows a multivariate Gaussian distribution $N(\mathbf{0}, \boldsymbol{\Theta}^{-1})$, where $\boldsymbol{\Theta} \in \mathbb{R}^{d\times d}$ is the precision matrix satisfying $\boldsymbol{\Theta}_{jj} = 1$ for $j \in [d]$. The conditional distribution of X_j given $\boldsymbol{X}_{\backslash j}$ satisfies

$$X_j \,|\, \boldsymbol{X}_{\backslash j} = \boldsymbol{\alpha}_j^T \boldsymbol{X}_{\backslash j} + \epsilon_j \quad \text{with} \quad \epsilon_j \sim N(0,1),$$

where $\boldsymbol{\alpha}_j = \boldsymbol{\Theta}_{\backslash j,j}$. The conditional density is given by

$$p(x_j \,|\, \boldsymbol{x}_{\backslash j}) = \sqrt{1/(2\pi)}\exp\bigl[-x_j(\boldsymbol{\Theta}_{\backslash j,j}^T \boldsymbol{x}_{\backslash j}) - 1/2\cdot x_j^2 - 1/2\cdot(\boldsymbol{\Theta}_{\backslash j,j}^T \boldsymbol{x}_{\backslash j})^2\bigr].$$

Compared with (2), we obtain $\boldsymbol{\beta}_j = -\boldsymbol{\Theta}_{\backslash j,j}$, $f_j(x) = -x^2/2$ and $b_j(\boldsymbol{\beta}_j, f_j) = (\boldsymbol{\beta}_j^T \boldsymbol{x}_{\backslash j})^2/2 + \log(2\pi)/2$.

Ising Models: In an Ising model with no external field, $\boldsymbol{X}$ takes value in $\{0,1\}^d$ and the joint probability mass function $p(\boldsymbol{x}) \propto \exp(\sum_{j<k}\theta_{jk}x_jx_k)$. Let $\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{j,j-1}, \theta_{j,j+1}, \ldots, \theta_{jd})^T$. The conditional distribution of X_j given $\boldsymbol{X}_{\backslash j}$ is of the form

$$p(x_j \,|\, \boldsymbol{x}_{\backslash j}) = \frac{\exp\bigl(\sum_{k<\ell}\theta_{k\ell}x_kx_\ell\bigr)}{\sum_{x_j\in\{0,1\}}\exp\bigl(\sum_{k<\ell}\theta_{k\ell}x_kx_\ell\bigr)} = \exp\Bigl\{x_j\bigl(\boldsymbol{\theta}_j^T\boldsymbol{x}_{\backslash j}\bigr) - \log\bigl[1+\exp(\boldsymbol{\theta}_j^T\boldsymbol{x}_{\backslash j})\bigr]\Bigr\}.$$

Therefore, in this case we have $\boldsymbol{\beta}_j = \boldsymbol{\theta}_j$, $f_j(x) = 0$ and $b_j(\boldsymbol{\beta}_j, f_j) = \log[1+\exp(\boldsymbol{\beta}_j^T\boldsymbol{x}_{\backslash j})]$.

Exponential Graphical Models: For exponential graphical models, $\boldsymbol{X}$ takes values in $[0,+\infty)^d$ and the joint probability density satisfies $p(\boldsymbol{x}) \propto \exp(-\sum_{j=1}^d \phi_jx_j - \sum_{k<\ell}\theta_{k\ell}x_kx_\ell)$. In order to ensure that this probability distribution is normalizable, we require that $\phi_j > 0, \theta_{jk} \geq 0$ for all $j,k \in [d]$. Then we obtain the following conditional probability density of X_j given $\boldsymbol{X}_{\backslash j}$:

$$\begin{aligned} p(x_j \,|\, \boldsymbol{x}_{\backslash j}) &= \exp\Bigl(-\sum_{k=1}^d \phi_kx_k - \sum_{k<\ell}\theta_{k\ell}x_kx_\ell\Bigr)\Big/\int_{x_j\geq 0}\exp\Bigl(-\sum_{k=1}^d\phi_kx_k - \sum_{k<\ell}\theta_{k\ell}x_kx_\ell\Bigr)\mathrm{d}x_j \\ &= \exp\bigl[-x_j\bigl(\phi_j + \boldsymbol{\theta}_j^T\boldsymbol{x}_{\backslash j}\bigr) - \log\bigl(\phi_j + \boldsymbol{\theta}_j^T\boldsymbol{x}_{\backslash j}\bigr)\bigr]. \end{aligned}$$

Thus, we have $\boldsymbol{\beta}_j = -\boldsymbol{\theta}_j$, $f_j(x) = -\phi_jx$ and $b_j(\boldsymbol{\beta}_j, f_j) = \log(\boldsymbol{\beta}_j^T\boldsymbol{x}_{\backslash j} + \phi_j)$.

Poisson Graphical Models: In a Poisson graphical model, every node X_j is a discrete random variable taking values in $\mathbb{N} = \{0, 1, 2, \ldots\}$. The joint probability mass function is given by

$$p(\boldsymbol{x}) \propto \exp\biggl[\sum_{j=1}^{d} \phi_j x_j - \sum_{j=1}^{d} \log(x_j!) + \sum_{k<\ell} \theta_{k\ell} x_k x_\ell\biggr].$$

Similar to the exponential graphical models, we also need to impose some restrictions on the parameters so that the probability mass function is normalizable. Here we require that $\theta_{jk} \le 0$ for all $j, k \in [d]$. By direct computation, the conditional probability mass function of X_j given $\boldsymbol{X}_{\backslash j}$ is given by

$$p(x_j \,|\, \boldsymbol{x}_{\backslash j}) = \exp\bigl[x_j\bigl(\boldsymbol{\theta}_j^T \boldsymbol{x}_{\backslash j}\bigr) + \phi_j x_j - \log(x_j!) - b_j(\boldsymbol{\theta}_j, f_j)\bigr],$$

where we have $\boldsymbol{\beta}_j = \boldsymbol{\theta}_j$, $f_j(x) = \phi_j x - \log(x!)$ and $b_j(\boldsymbol{\beta}_j, f_j) = \log\bigl\{\sum_{y=0}^{\infty} \exp\bigl[y(\boldsymbol{\beta}_j^T \boldsymbol{x}_{\backslash j}) + f_j(y)\bigr]\bigr\}$.

3. Graph Estimation and Uncertainty Assessment

In this section, we lay out the procedures for graph estimation and uncertainty assessment. Throughout our analysis, we use $\{\boldsymbol{\beta}_i^*, f_i^*\}_{i\in[d]}$ to denote the true parameters, and $\mathbb{E}(\cdot)$ to denote the expectation with respect to the joint density in (3) with the true parameters. We first introduce a pseudo-likelihood loss function for the parametric components $\{\boldsymbol{\beta}_j\}_{j=1}^d$ that is invariant to the nuisance parameters $\{f_j\}_{j\in[d]}$. Based on such a loss function, we present an Adaptive Multi-stage Convex Relaxation algorithm to estimate each $\boldsymbol{\beta}_j^*$ by minimizing the loss function regularized by a nonconvex penalty function. We then proceed to introduce the inferential procedure for accessing the uncertainty of a given edge in the graph.

3.1. A Nuisance-Free Loss Function

For graph estimation, we treat $\boldsymbol{\beta}_j$ as the parameter of interest and the base measures $f_j(\cdot)$ as nuisance parameter. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be n i.i.d. copies of $\boldsymbol{X}$. Due to the presence of $f_j(\cdot)$, finding the conditional maximum likelihood estimator of $\boldsymbol{\beta}_j$ is intractable. To solve this problem, we exploit a pseudo-likelihood loss function proposed in Ning et al. (2017b) that is invariant to the nuisance parameters $\{f_j\}_{j\in[d]}$. This pseudo-likelihood loss is based on pairwise local order statistics, which have been previously studied in Liang and Qin (2000); Diao et al. (2012); Chan (2012) for semiparametric regression models. More details are presented as follows.

Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ be n data points that are realizations of $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$. For any $1 \le i < i' \le n$, let

$$\mathcal{A}_{ii'}^j := \bigl\{(X_{ij}, X_{i'j}) = (x_{ij}, x_{i'j}), \boldsymbol{X}_{i\backslash j} = \boldsymbol{x}_{i\backslash j}, \boldsymbol{X}_{i'\backslash j} = \boldsymbol{x}_{i'\backslash j}\bigr\}$$

be the event that we observe $\boldsymbol{X}_{i\backslash j} = \boldsymbol{x}_{i\backslash j}$ and $\boldsymbol{X}_{i'\backslash j} = \boldsymbol{x}_{i'\backslash j}$ and the order statistics of X_{ij} and $X_{i'j}$ (but not the relative ranks of X_{ij} and $X_{i'j}$). More specifically, we denote $\max\{X_{ij}, X_{i'j}\}$ and $\min\{X_{ij}, X_{i'j}\}$ by O_1 and O_2, and let o_1 and o_2 be the observed values of O_1 and O_2. Then $\mathcal{A}_{ii'}^j$ can be equivalently written as $\{O_1 = o_1, O_2 = o_2, \boldsymbol{X}_{i\backslash j} = \boldsymbol{x}_{i\backslash j}, \boldsymbol{X}_{i'\backslash j} = \boldsymbol{x}_{i'\backslash j}\}$.

Let $R \in \{(1,2),(2,1)\}$ be the relative rank of X_{ij} and $X_{i'j}$, and r be the observed value. Then, by definition, we have

$$\begin{aligned}&\mathbb{P}\big(X_{ij} = x_{ij}, X_{i'j} = x_{i'j} \,\big|\, \boldsymbol{X}_{i\backslash j} = \boldsymbol{x}_{i\backslash j}, \boldsymbol{X}_{i'\backslash j} = \boldsymbol{x}_{i'\backslash j}\big)\\ &\quad = \mathbb{P}\big(O_1 = o_1, O_2 = o_2 \,\big|\, \boldsymbol{X}_{i\backslash j} = \boldsymbol{x}_{i\backslash j}, \boldsymbol{X}_{i'\backslash j} = \boldsymbol{x}_{i'\backslash j}\big) \cdot \mathbb{P}\big(R = r \,\big|\, \mathcal{A}^j_{ii'}\big).\end{aligned}$$

Furthermore, we have

$$\begin{aligned}\mathbb{P}\big(R = r \,\big|\, \mathcal{A}^j_{ii'}\big) &= \left[1 + \frac{\mathbb{P}(X_{ij} = x_{i'j}, X_{i'j} = x_{ij} \,|\, \mathcal{A}^j_{ii'})}{\mathbb{P}(X_{ij} = x_{ij}, X_{i'j} = x_{i'j} \,|\, \mathcal{A}^j_{ii'})}\right]^{-1}\\ = &\left[1 + \frac{\mathbb{P}(X_{ij} = x_{i'j}, X_{i'j} = x_{ij} \,|\, \boldsymbol{X}_{i\backslash j} = \boldsymbol{x}_{i\backslash j}, \boldsymbol{X}_{i'\backslash j} = \boldsymbol{x}_{i'\backslash j})}{\mathbb{P}(X_{ij} = x_{ij}, X_{i'j} = x_{i'j} \,|\, \boldsymbol{X}_{i\backslash j} = \boldsymbol{x}_{i\backslash j}, \boldsymbol{X}_{i'\backslash j} = \boldsymbol{x}_{i'\backslash j})}\right]^{-1} = \big[1 + R^j_{ii'}(\boldsymbol{\beta}_j)\big]^{-1},\end{aligned} \tag{5}$$

where $R^j_{ii'}(\boldsymbol{\beta}_j) := \exp[-(x_{ij} - x_{i'j})\boldsymbol{\beta}_j{}^T(\boldsymbol{x}_{i\backslash j} - \boldsymbol{x}_{i'\backslash j})]$. Based on the conditional likelihood in (5), we construct the following pseudo-likelihood loss function for $\boldsymbol{\beta}_j$:

$$L_j(\boldsymbol{\beta}_j) := \frac{2}{n(n-1)} \sum_{1 \le i < i' \le n} \log\big[1 + R^j_{ii'}(\boldsymbol{\beta}_j)\big]. \tag{6}$$

Obviously, $L_j(\cdot)$ only involves $\boldsymbol{\beta}_j$. Since its form resembles the logistic loss, to find a minimizer of this loss function, we could readily apply any logistic regression solver.

3.2. Adaptive Multi-stage Convex Relaxation Algorithm

Now we are ready to present the algorithm for parameter estimation. For high dimensional sparse estimation, to promote sparsity, we minimize the sum of the loss functions $L_j(\boldsymbol{\beta}_j)$ and some penalty function. Two of the most prevalent methods are the LASSO (ℓ_1-penalization) (Tibshirani, 1996) and the folded concave penalization (Fan et al., 2014). Although the ℓ_1-penalization enjoys good computational properties as a convex optimization problem, it is known to incur significant estimation bias for parameters with large absolute values (Zhang and Huang, 2008). In contrast, nonconvex penalties such as smoothly clipped absolute deviation (SCAD) penalty, minimax concave penalty (MCP) and capped-ℓ_1 penalty can eliminate such bias and attain improved rates of convergence. Therefore, we consider the nonconvex optimization problem

$$\widehat{\boldsymbol{\beta}}_j = \underset{\mathbb{R}^{d-1}}{\operatorname{argmin}} \left\{L_j(\boldsymbol{\beta}_j) + \sum_{k \neq j} p_\lambda(|\beta_{jk}|)\right\}, \tag{7}$$

where $\lambda > 0$ is a regularization parameter and $p_\lambda(\cdot) : [0, +\infty) \to [0, +\infty)$ is a penalty function satisfying the following three conditions:

(C.1) The penalty function $p_\lambda(u)$ is continuously nondecreasing and concave with $p_\lambda(0) = 0$.

(C.2) The right-hand derivative at $u = 0$ satisfies $p'_\lambda(0) = p'_\lambda(0+) = \lambda$.

(C.3) There exist constants $c_1 \in [0,1]$ and $c_2 \in (0,+\infty)$ such that $p'_\lambda(u+) \geq c_1\lambda$ for $u \in [0, c_2\lambda]$.

Note that we only require the penalty function to be right-differentiable. In what follows, we denote by $p'_\lambda(u)$ the right-hand derivative. By (C.1), $p'_\lambda(u)$ is nonincreasing and nonnegative in $[0,\infty)$. It is easy to verify that SCAD, MCP and capped-ℓ_1 penalty all satisfy (C.1)–(C.3).

Due to the penalty term, the optimization problem in (7) is nonconvex and may have multiple local solutions. To overcome such difficulty, we exploit the local linear approximation algorithm (Zou and Li, 2008; Fan et al., 2014) or equivalently, the multi-stage convex relaxation (Zhang, 2010; Zhang et al., 2013; Fan et al., 2018) to attain an estimator of $\boldsymbol{\beta}_j^*$. Compared with previous works that mainly focus on sparse linear regression, our loss function $L_j(\boldsymbol{\beta}_j)$ is a U-statistics based logistic loss, which requires nontrivial extensions of the existing theoretical analysis.

We present the proposed adaptive multi-stage convex relaxation method in Algorithm 1. Our algorithm solves a sequence of convex optimization problems corresponding to finer and finer convex relaxations of the original nonconvex optimization problem. More specifically, for each $j = 1, \ldots, d$, in the first iteration, step 4 of Algorithm 1 is equivalent to a ℓ_1-regularized optimization problem and we obtain the first-step solution $\widehat{\boldsymbol{\beta}}_j^{(1)}$. Then, in each subsequent iteration, we solve an adaptive ℓ_1-regularized optimization problem where the weights of the penalty depend on the solution of the previous step. For example, in the ℓ-th iteration, the regularization parameter $\lambda_{jk}^{(\ell-1)}$ in (8) is updated using the $(\ell-1)$-th step estimator $\widehat{\boldsymbol{\beta}}_j^{(\ell-1)}$. Note that $p'_\lambda\big(|\beta_{jk}^{(\ell)}|\big)$ is the right-hand derivative of $p_\lambda(u)$ evaluated at $u = |\beta_{jk}^{(\ell)}|$.

Since the optimization problem in step 4 is convex, our method is computationally efficient. Besides, note that (8) with $\ell = 1$ corresponds to the ℓ_1-regularized problem. Hence, our approach can be viewed as a refinement of LASSO. As we will show in §4.1, the estimator $\widehat{\boldsymbol{\beta}}_j$ of $\boldsymbol{\beta}_j^*$ constructed by Algorithm 1 attains the optimal statistical rates of convergence for parameter estimation.

Algorithm 1 Adaptive Multi-stage Convex Relaxation algorithm for parameter estimation

```
1: Initialize $\lambda_{jk}^{(0)} = \lambda$ for $1 \leq j, k \leq d$.
2: for j= 1,2,...,d do
3:   for $\ell = 1, 2, \ldots,$ until convergence do
4:     Solve the convex optimization problem
           $$\widehat{\boldsymbol{\beta}}_j^{(\ell)} = \underset{\mathbb{R}^{d-1}}{\mathrm{argmin}} \Big\{ L_j(\boldsymbol{\beta}_j) + \sum_{k \neq j} \lambda_{jk}^{(\ell-1)} |\beta_{jk}| \Big\}. \qquad (8)$$
5:     Update $\lambda_{jk}^{(\ell)}$ by $\lambda_{jk}^{(\ell)} = p'_\lambda(|\widehat{\beta}_{jk}^{(\ell)}|)$ for $1 \leq k \leq d, k \neq j$.
6:   end for
7:   Output $\widehat{\boldsymbol{\beta}}_j = \widehat{\boldsymbol{\beta}}_j^{(\ell)}$, where $\ell$ is the number of iterations until convergence is attained.
8: end for
```

3.3. Graph Inference: Composite Pairwise Score Test

For any given $1 \leq j < k \leq d$, we are interested in testing if $(j, k) \in E$, i.e., we consider the hypothesis testing problem $H_0 : \beta^*_{jk} = 0$ versus $H_1 : \beta^*_{jk} \neq 0$. To simplify the notation, we write $\boldsymbol{\beta}_{j\setminus k} = (\beta_{j1}, \ldots, \beta_{jj-1}, \beta_{jj+1}, \ldots, \beta_{jk-1}, \beta_{jk+1}, \ldots, \beta_{jd})^T \in \mathbb{R}^{d-2}$ and denote the parameters associated with node j and node k by $\boldsymbol{\beta}_{j\vee k} := \big(\beta_{jk}; \boldsymbol{\beta}^T_{j\setminus k}, \boldsymbol{\beta}^T_{k\setminus j}\big)^T \in \mathbb{R}^{2d-3}$. In addition, let $\mathbf{H}^j := \mathbb{E}\big[\nabla^2 L_j(\boldsymbol{\beta}^*_j)\big]$ be the expected Hessian of $L_j(\boldsymbol{\beta}_j)$ evaluated at $\boldsymbol{\beta}^*_j$. We define two submatrices $\mathbf{H}^j_{jk,j\setminus k}$ and $\mathbf{H}^j_{j\setminus k,j\setminus k}$ of $\mathbf{H}^j$ as

$$\mathbf{H}^j_{jk,j\setminus k} := \left[\mathbb{E}\frac{\partial^2 L_j(\boldsymbol{\beta}^*_j)}{\partial\beta_{jk}\partial\beta_{jv}}\right]_{v\neq k} \in \mathbb{R}^{d-2} \text{ and } \mathbf{H}^j_{j\setminus k,j\setminus k} := \left[\mathbb{E}\frac{\partial^2 L_j(\boldsymbol{\beta}^*_j)}{\partial\beta_{ju}\partial\beta_{jv}}\right]_{u\neq k,v\neq k} \in \mathbb{R}^{(d-2)\times(d-2)},$$

and we define $\mathbf{H}^k_{jk,k\setminus j}$ and $\mathbf{H}^k_{k\setminus j,k\setminus j}$ similarly. Furthermore, we define

$$\mathbf{w}^*_{j,k} = \mathbf{H}^j_{jk,j\setminus k}\big[\mathbf{H}^j_{j\setminus k,j\setminus k}\big]^{-1} \quad \text{and} \quad \mathbf{w}^*_{k,j} = \mathbf{H}^k_{jk,k\setminus j}\big[\mathbf{H}^k_{k\setminus j,k\setminus j}\big]^{-1}. \tag{9}$$

Following the general approach in Ning et al. (2017a); Neykov et al. (2018), the composite pairwise score function for parameter β_{jk} is defined as

$$S_{jk}(\boldsymbol{\beta}_{j\vee k}) = \nabla_{jk}L_j(\boldsymbol{\beta}_j) + \nabla_{jk}L_k(\boldsymbol{\beta}_k) - {\mathbf{w}^*_{j,k}}^T\nabla_{j\setminus k}L_j(\boldsymbol{\beta}_j) - {\mathbf{w}^*_{k,j}}^T\nabla_{k\setminus j}L_k(\boldsymbol{\beta}_k). \tag{10}$$

where we write $\nabla_{jk}L_j(\boldsymbol{\beta}_j) = \partial L_j(\boldsymbol{\beta}_j)/\partial\beta_{jk}$ and $\nabla_{j\setminus k}L_j(\boldsymbol{\beta}_j) = \partial L_j(\boldsymbol{\beta}_j)/\partial\boldsymbol{\beta}_{j\setminus k}$. Here, the last two terms in (10) are constructed to reduce the effect of nuisance parameters $\boldsymbol{\beta}_{j\setminus k}$ and $\boldsymbol{\beta}_{k\setminus j}$ on assessing the uncertainty of β^*_{jk}, which is the parameter of interest. A key feature of $S_{jk}(\boldsymbol{\beta}_{j\vee k})$ is that the symmetry of β_{jk} and β_{kj} (i.e., $\beta_{jk} = \beta_{kj}$) is taken into account, which is distinct from the existing works such as Ren et al. (2015); Janková and van de Geer (2015); Liu et al. (2013) for Gaussian graphical models and Ning et al. (2017b) in the regression setup.

Note that both $\mathbf{w}^*_{j,k}$ and $\mathbf{w}^*_{k,j}$ are computed from $\mathbf{H}$, which is unknown. We estimate them using the Dantzig-type estimators (Candés et al., 2007). Specifically, we define the empirical versions of $\mathbf{H}^j_{jk,j\setminus k}$ and $\mathbf{H}^j_{j\setminus k,j\setminus k}$ as

$$\nabla^2_{jk,j\setminus k}L_j(\boldsymbol{\beta}_j) = \left[\frac{\partial^2 L_j(\boldsymbol{\beta}_j)}{\partial\beta_{jk}\partial\beta_{jv}}\right]_{v\neq k} \quad \text{and} \quad \nabla^2_{j\setminus k,j\setminus k}L_j(\boldsymbol{\beta}_j) = \left[\frac{\partial^2 L_j(\boldsymbol{\beta}_j)}{\partial\beta_{ju}\partial\beta_{jv}}\right]_{u\neq k,v\neq k}.$$

We also define $\nabla^2_{jk,k\setminus j}L_k(\boldsymbol{\beta}_k)$ and $\nabla^2_{k\setminus j,k\setminus j}L_k(\boldsymbol{\beta}_k)$ similarly. Then we estimate $\mathbf{w}^*_{j,k}$ by solving

$$\widehat{\mathbf{w}}_{j,k} = \operatorname{argmin}\|\mathbf{w}\|_1 \quad \text{such that} \quad \big\|\nabla^2_{jk,j\setminus k}L_j(0,\widehat{\boldsymbol{\beta}}_{j\setminus k}) - \mathbf{w}^T\nabla^2_{j\setminus k,j\setminus k}L_j(0,\widehat{\boldsymbol{\beta}}_{j\setminus k})\big\|_\infty \leq \lambda_D, \tag{11}$$

where $\widehat{\boldsymbol{\beta}}_j$ is the estimator of $\boldsymbol{\beta}^*_j$ obtained from Algorithm 1 and λ_D is a regularization parameter. An estimator $\widehat{\mathbf{w}}_{k,j}$ of $\mathbf{w}^*_{k,j}$ can be similarly obtained. Based on $\widehat{\mathbf{w}}_{j,k}$ and $\widehat{\mathbf{w}}_{k,j}$, we construct the composite pairwise score statistic for β^*_{jk} by

$$\widehat{S}_{jk} = \nabla_{jk}L_j(0,\widehat{\boldsymbol{\beta}}_{j\setminus k}) + \nabla_{jk}L_k\big(0,\widehat{\boldsymbol{\beta}}_{k\setminus j}\big) - \widehat{\mathbf{w}}^T_{j,k}\nabla_{j\setminus k}L_j\big(0,\widehat{\boldsymbol{\beta}}_{j\setminus k}\big) - \widehat{\mathbf{w}}^T_{k,j}\nabla_{k\setminus j}L_k\big(0,\widehat{\boldsymbol{\beta}}_{k\setminus j}\big). \tag{12}$$

Comparing (10) and (12), we see that $\widehat{S}_{jk}$ is obtained by replacing $\boldsymbol{\beta}_j$ and $\boldsymbol{\beta}_k$ in (10) by $(0, \widehat{\boldsymbol{\beta}}_{j\setminus k})$ and $(0, \widehat{\boldsymbol{\beta}}_{k\setminus j})$ respectively and replacing $\mathbf{w}^*_{j,k}$ and $\mathbf{w}^*_{k,j}$ in (10) by $\widehat{\mathbf{w}}_{j,k}$ and $\widehat{\mathbf{w}}_{k,j}$.

To obtain a valid hypothesis test, we need to establish the limiting distribution of $\widehat{S}_{jk}$ under the null hypothesis. Note that $\widehat{S}_{jk}$ is a linear combination of entries of $\nabla L_j(\boldsymbol{\beta}_j)$ and $\nabla L_k(\boldsymbol{\beta}_k)$, both of which are U-statistics. In the next section, we prove the asymptotic normality of $\widehat{S}_{jk}$. More specifically, under the null hypothesis, we have $\sqrt{n}\widehat{S}_{jk}/2 \rightsquigarrow N(0, \sigma^2_{jk})$, where the limiting variance can be estimated consistently by $\widehat{\sigma}^2_{jk}$ (More details will be explained in the following section). With a significance level $\alpha \in (0,1)$, the test function $\psi_{jk}(\alpha)$ is defined as

$$\psi_{jk}(\alpha) = \begin{cases} 1 & \text{if} \quad \big|\sqrt{n}\widehat{S}_{jk}/(2\widehat{\sigma}_{jk})\big| > \Phi^{-1}(1-\alpha/2) \\ 0 & \text{if} \quad \big|\sqrt{n}\widehat{S}_{jk}/(2\widehat{\sigma}_{jk})\big| \le \Phi^{-1}(1-\alpha/2) \end{cases}, \tag{13}$$

where $\Phi(t)$ is the cumulative distribution function of a standard normal random variable.

In sum, the composite pairwise score test for the null hypothesis $H_0\colon \beta^*_{jk} = 0$ consists of the following four steps: (i) Calculate $\widehat{\boldsymbol{\beta}}_j$ and $\widehat{\boldsymbol{\beta}}_k$ from Algorithm 1; (ii) Obtain $\widehat{\mathbf{w}}_{j,k}$ and $\widehat{\mathbf{w}}_{k,j}$ by solving two Dantzig-type problems defined in (11); (iii) Compute the limiting variance $\widehat{\sigma}^2_{jk}$; (iv) Evaluate the test function (13).

4. Theoretical Properties

In this section, we present our theoretical results. We first prove that the proposed procedure attains the optimal rate of convergence for parameter estimation. Then, we provide theory for the composite pairwise score test.

4.1. Theoretical Results for Parameter Estimation

We first establish the rates of convergence of the adaptive multi-stage convex relaxation estimator. We begin by listing several required assumptions. The first is about moment conditions of $\{X_j\}$ and the local smoothness of the log-partition function $A(\cdot)$ defined in (4). This assumption also appears in Yang et al. (2013a) and Chen et al. (2015) as a pivotal technical condition for theoretical analysis.

Assumption 2 *For all $j \in [d]$, we assume that the first two moments of X_j are bounded. That is, there exist two constants κ_m and κ_v such that $|\mathbb{E}(X_j)| \le \kappa_m$ and $\mathbb{E}(X_j^2) \le \kappa_v$. Denote the true parameters by $\{\boldsymbol{\beta}^*_j, f^*_j\}_{j\in[d]}$ and define d univariate functions $\bar{A}_j(\cdot)\colon \mathbb{R} \to \mathbb{R}$ as*

$$\bar{A}_j(u) := \log\bigg\{ \int_{\mathbb{R}^d} \exp\bigg[u x_j + \sum_{k<\ell} \beta^*_{k\ell} x_k x_\ell + \sum_{i=1}^d f^*_i(x_i) \bigg] \mathrm{d}\nu(\boldsymbol{x}) \bigg\}, \quad j \in [d].$$

We assume that there exists a constant κ_h such that $\max_{u\colon |u|\le 1} \bar{A}''_j(u) \le \kappa_h$ *for all $j \in [d]$.*

Unlike the Ising graphical models, $\{X_j\}_{j\in[d]}$ are not bounded in general for semiparametric exponential family graphical models. Instead, we impose mild conditions as in

Assumption 2 to obtain a loose control of the tail behaviors of the distribution of $\boldsymbol{X}$. As shown in Yang et al. (2013a), Assumption 2 implies that for all $j \in [d]$,

$$\max\bigl\{\log \mathbb{E}[\exp(X_j)], \log \mathbb{E}[\exp(-X_j)]\bigr\} \le \kappa_m + \kappa_h/2.$$

Markov inequality implies for any $x > 0$,

$$\mathbb{P}\bigl(|X_j| \ge x\bigr) \le 2\exp(\kappa_m + \kappa_h/2) \cdot \exp(-x). \tag{14}$$

Thus, by setting $x = C \log d$ in (14) with constant C sufficiently large, we have $\|\boldsymbol{X}\|_\infty \le C \log d$ with high probability. In addition to Assumption 2, we also impose conditions to control the curvature of function $L_j(\cdot)$.

Definition 3 (Sparse eigenvalue condition) *For any $j, s \in [d]$, we define the s-sparse eigenvalues of $\mathbb{E}[\nabla^2 L_j(\boldsymbol{\beta}_j^*)]$ as*

$$\begin{aligned}\rho_{j+}^*(s) &:= \sup_{\mathbf{v}\in\mathbb{R}^{d-1}} \bigl\{\mathbf{v}^T \mathbb{E}\bigl[\nabla^2 L_j(\boldsymbol{\beta}_j^*)\bigr]\mathbf{v} \colon \|\mathbf{v}\|_0 \le s, \|\mathbf{v}\|_2 = 1\bigr\};\\ \rho_{j-}^*(s) &:= \inf_{\mathbf{v}\in\mathbb{R}^{d-1}} \bigl\{\mathbf{v}^T \mathbb{E}\bigl[\nabla^2 L_j(\boldsymbol{\beta}_j^*)\bigr]\mathbf{v} \colon \|\mathbf{v}\|_0 \le s, \|\mathbf{v}\|_2 = 1\bigr\}.\end{aligned}$$

Assumption 4 *Let $s^* = \max_{j\in[d]} \|\boldsymbol{\beta}_j^*\|_0$. We assume that for any $j \in [d]$, there exist an integer $k^* \ge 2s^*$ satisfying $\lim\limits_{n\to\infty} k^*(\log^9 d/n)^{1/2} = 0$ and a positive number ρ_* such that the sparse eigenvalues of $\mathbb{E}[\nabla^2 L_j(\boldsymbol{\beta}_j^*)]$ satisfy*

$$\begin{aligned}&0 < \rho_* \le \rho_{j-}^*(2s^* + 2k^*) < \rho_{j+}^*(k^*) < +\infty \quad \textit{and}\\ &\rho_{j+}^*(k^*)\big/\rho_{j-}^*(2s^* + 2k^*) \le 1 + 0.2k^*/s^* \quad \textit{for any } j \in [d].\end{aligned}$$

The condition $\rho_{j+}^*(k^*)/\rho_{j-}^*(2s^*+2k^*) \le 1+0.2k^*/s^*$ requires the eigenvalue ratio $\rho_{j+}^*(k)/\rho_{j-}^*(2k+2s^*)$ to grow sub-linearly in k. Assumption 4 is commonly referred to as sparse eigenvalue condition, which is standard for sparse estimation problems and has been studied by Bickel et al. (2009); Raskutti et al. (2010); Zhang (2010); Negahban et al. (2012); Xiao and Zhang (2013); Loh and Wainwright (2015) and Wang et al. (2014). Our assumption is similar to that in Zhang (2010) and is weaker than the restricted isometry property (RIP) proposed in Candés and Tao (2005). We claim that this assumption is true in general and will be verified for Gaussian graphical models in the appendix.

Now we are ready to present the main theorem of this section. Recall that the penalty function $p_\lambda(u)$ satisfies conditions (C.1)–(C.3) in §3.2. We use $p'_\lambda(u)$ to denote its right-hand derivative. For convenience, we will set $p'_\lambda(u) = 1$ when $u < 0$.

Theorem 5 (ℓ_2- and ℓ_1-rates of convergence) *For all $j \in [d]$, we define the support of $\boldsymbol{\beta}_j^*$ as $S_j := \bigl\{(j,k)\colon \beta_{jk}^* \neq 0, k \in [d]\bigr\}$ and let $s^* = \max_{j\in[d]} \|\boldsymbol{\beta}_j^*\|_0$. Let $\rho_* > 0$ be defined in Assumption 4. Under Assumptions 2 and 4, there exists an absolute constant $K > 0$ such that $\|\nabla L_j(\boldsymbol{\beta}_j^*)\|_\infty \le K\sqrt{\log d/n}, \forall j \in [d]$ with probability at least $1 - (2d)^{-1}$. Moreover, the penalty function $p_\lambda(\cdot)$ in (7) satisfies (C.1)–(C.3) listed in §3.2 with $c_1 = 0.91$ and $c_2 \ge 24/\rho_*$ for condition (C.3). We set the regulization parameter $\lambda = C\sqrt{\log d/n}$ with $C \ge 25K$. We denote constants $\varrho = c_2(c_2\rho_* - 11)^{-1}$, $A_1 = 22\varrho$, $A_2 = 2.2c_2$, $B_1 = 32\varrho$,*

$B_2 = 3.2c_2$, $\gamma = 11c_2^{-1}\rho_*^{-1} < 1$, and define $\Upsilon_j := [\sum_{(j,k)\in S_j} p'_\lambda(|\beta^*_{jk}| - c_2\lambda)^2]^{1/2}$. Then, with probability at least $1-d^{-1}$, we have the following statistical rates of convergence:

$$\big\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}_j^*\big\|_2 \le A_1\big[\big\|\nabla_{S_j} L_j(\boldsymbol{\beta}_j^*)\big\|_2 + \Upsilon_j\big] + A_2\sqrt{s^*}\lambda\gamma^\ell \quad \textit{and} \tag{15}$$

$$\big\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}_j^*\big\|_1 \le B_1\sqrt{s^*}\big[\big\|\nabla_{S_j} L_j(\boldsymbol{\beta}_j^*)\big\|_2 + \Upsilon_j\big] + B_2 s^*\lambda\gamma^\ell, \forall j \in [d]. \tag{16}$$

By Theorem 5, the statistical rates are dominated by the second term if $p'_\lambda(|\beta^*_{jk}| - c_2\lambda)$ is not negligible. If the signal strength is large enough such that $p'_\lambda(\beta - c_2\lambda) = 0$ where $\beta = \min_{(j,k)\in S_j} |\beta^*_{jk}|$, after sufficient number of iterations, the statistical rates will be of the order

$$\big\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}_j^*\big\|_2 = \mathcal{O}_{\mathbb{P}}\big(\big\|\nabla_{S_j} L_j(\boldsymbol{\beta}_j^*)\big\|_2\big) \quad \text{and} \quad \big\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}_j^*\big\|_1 = \mathcal{O}_{\mathbb{P}}\big(\sqrt{s^*}\big\|\nabla_{S_j} L_j(\boldsymbol{\beta}_j^*)\big\|_2\big).$$

However, if the signals are uniformly small such that $p'_\lambda\big(|\beta^*_{jk}| - c_2\lambda\big) > 0$ for all $(j,k) \in S_j$, the rates of convergence will be of the order

$$\big\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}_j^*\big\|_2 = \mathcal{O}_{\mathbb{P}}\big(\sqrt{s^*}\lambda\big) \quad \text{and} \quad \big\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}_j^*\big\|_1 = \mathcal{O}_{\mathbb{P}}\big(s^*\lambda\big),$$

which are identical to the ℓ_2- and ℓ_1-rates of the LASSO estimator, respectively (Ning et al., 2017b). Thus $c_2\lambda$ can be viewed as the threshold of signal strength. Therefore, after sufficient numbers of iterations, the final estimator $\widehat{\boldsymbol{\beta}}_j$ obtained by Algorithm 1 attains the following more refined rates of convergence:

$$\big\|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\big\|_2 = \mathcal{O}_{\mathbb{P}}\Big(\big\|\nabla_{S_j} L_j(\boldsymbol{\beta}_j^*)\big\|_2 + \Upsilon_j\Big) \text{ and } \big\|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\big\|_1 = \mathcal{O}_{\mathbb{P}}\Big(\sqrt{s^*}\big[\big\|\nabla_{S_j} L_j(\boldsymbol{\beta}_j^*)\big\|_2 + \Upsilon_j\big]\Big).$$

These statistical rates of convergence are optimal in the sense that they cannot be improved in terms of the order.

Finally, we comment that the sparsity level s^* in (15) and (16) can be replaced by the sparsity level of each $\boldsymbol{\beta}_j^*$. Let $s_j^* = \|\boldsymbol{\beta}_j^*\|_0$ be the sparsity level of $\boldsymbol{\beta}_j^*$ and λ_j be the regularization parameter for optimization problem (7) such that $\lambda_j \asymp \|\nabla L_j(\boldsymbol{\beta}_j^*)\|_\infty$. The statistical rates of convergence for each $\widehat{\boldsymbol{\beta}}_j^{(\ell)}$ can be improved to

$$\big\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}_j^*\big\|_2 = \mathcal{O}_{\mathbb{P}}\big(\sqrt{s_j^*}\lambda_j\big) \quad \text{and} \quad \big\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}_j^*\big\|_1 = \mathcal{O}_{\mathbb{P}}\big(s_j^*\lambda_j\big).$$

We use the uniform sparsity level $s^* = \max_{j\in[d]} s_j^*$ and the same regularization parameter λ for simplicity, but the proof can be easily adapted to individual s_j^* and λ_j for each $j \in [d]$.

4.2. Theoretical Results for Composite Pairwise Score Test

In the composite pairwise score test for the null hypothesis $H_0 : \beta^*_{jk} = 0$, we construct the test statistic by combining the loss functions $L_j(\cdot)$ and $L_k(\cdot)$ together because β_{jk} appears in both $L_j(\boldsymbol{\beta}_j)$ and $L_k(\boldsymbol{\beta}_k)$ (recall that we use β_{jk} and β_{kj} interchangeably). In the sequel, we present the theoretical results that guarantee the validity of the proposed inferential method.

Recall that we define the pairwise score function $S_{jk}(\boldsymbol{\beta}_{j\vee k})$ and the pairwise score statistic $\widehat{S}_{jk}$ in (10) and (12) respectively. According to a fixed pair of nodes (j,k), entries in

$\boldsymbol{\beta}_j$ and $\boldsymbol{\beta}_k$ can be categorized into three types: (i) β_{jk}, (ii) $\boldsymbol{\beta}_{j\backslash k} = (\beta_{j\ell}; \ell \neq k)^T$, and (iii) $\boldsymbol{\beta}_{k\backslash j} = (\beta_{k\ell}; \ell \neq j)^T$. Recall that we write $\boldsymbol{\beta}_{j\vee k} = (\beta_{jk}, \boldsymbol{\beta}_{j\backslash k}^T, \boldsymbol{\beta}_{k\backslash j}^T)^T$ for notational simplicity. Moreover, letting $L_{jk}\big(\boldsymbol{\beta}_{j\vee k}\big) := L_j(\boldsymbol{\beta}_j) + L_k(\boldsymbol{\beta}_j)$, the entries of $\nabla L_{jk}\big(\boldsymbol{\beta}_{j\vee k}\big)$ are given by

$$\begin{aligned}&\nabla_{jk} L_{jk}(\boldsymbol{\beta}_{j\vee k}) = \nabla_{jk} L_j(\boldsymbol{\beta}_j) + \nabla_{kj} L_k(\boldsymbol{\beta}_k); \quad \nabla_{j\backslash k} L_{jk}(\boldsymbol{\beta}_{j\vee k}) = \nabla_{j\backslash k} L_j(\boldsymbol{\beta}_j), \quad \text{and}\\ &\nabla_{k\backslash j} L_{jk}(\boldsymbol{\beta}_{j\vee k}) = \nabla_{k\backslash j} L_k(\boldsymbol{\beta}_k).\end{aligned}$$

Let $\widehat{\boldsymbol{\beta}}_j$ and $\widehat{\boldsymbol{\beta}}_k$ be the estimators of $\boldsymbol{\beta}_j^*$ and $\boldsymbol{\beta}_k^*$ obtained from Algorithm 1. Note that we can write the pairwise score function $S_{jk}(\cdot)$ and the test statistic $\widehat{S}_{jk}$ as

$$S_{jk}(\boldsymbol{\beta}_{j\vee k}) = \nabla_{jk} L_{jk}\big(\boldsymbol{\beta}_{j\vee k}\big) - \mathbf{w}_{j,k}^{*T} \nabla_{j\backslash k} L_{jk}\big(\boldsymbol{\beta}_{j\vee k}\big) - \mathbf{w}_{k,j}^{*}{}^{T} \nabla_{k\backslash j} L_{jk}\big(\boldsymbol{\beta}_{j\vee k}\big) \quad \text{and} \tag{17}$$

$$\widehat{S}_{jk} = \nabla_{jk} L_{jk}\big(\widehat{\boldsymbol{\beta}}'_{j\vee k}\big) - \widehat{\mathbf{w}}_{j,k}^{T} \nabla_{j\backslash k} L_{jk}\big(\widehat{\boldsymbol{\beta}}'_{j\vee k}\big) - \widehat{\mathbf{w}}_{k,j}^{T} \nabla_{k\backslash j} L_{jk}\big(\widehat{\boldsymbol{\beta}}'_{j\vee k}\big), \tag{18}$$

where we write $\widehat{\boldsymbol{\beta}}'_{j\vee k} := \big(0, \widehat{\boldsymbol{\beta}}_{j\backslash k}^T, \widehat{\boldsymbol{\beta}}_{k\backslash j}^T\big)^T$, $\mathbf{w}_{j,k}^*$ and $\mathbf{w}_{k,j}^*$ are defined in (9), $\widehat{\mathbf{w}}_{j,k}$ is obtained from the Dantzig-type problem in (11), and $\widehat{\mathbf{w}}_{k,j}$ can be obtained similarly. To derive the asymptotic distribution of $\widehat{S}_{jk}$ under the null hypothesis, we first show that $\sqrt{n}\big[\widehat{S}_{jk} - S_{jk}(\boldsymbol{\beta}_{j\vee k}^*)\big] = o_{\mathbb{P}}(1)$. Then the problem is reduced to finding the limiting distribution of $S_{jk}(\boldsymbol{\beta}_{j\vee k}^*)$ under H_0. Thanks to its structure of being a U-statistics, we can characterize the limiting distribution of $S_{jk}(\boldsymbol{\beta}_{j\vee k}^*)$ using the method of Hájek projection (Van der Vaart, 2000), which approximates a U-statistic with a sum of independent random variables.

To begin with, we denote the kernel functions of $\nabla L_j(\boldsymbol{\beta}_j)$, $\nabla L_k(\boldsymbol{\beta}_k)$ and $\nabla L_{jk}(\boldsymbol{\beta}_{j\vee k})$ as $\mathbf{h}_{ii'}^{j}(\boldsymbol{\beta}_j)$, $\mathbf{h}_{ii'}^{k}(\boldsymbol{\beta}_k)$ and $\mathbf{h}_{ii'}^{jk}(\boldsymbol{\beta}_{j\vee k})$ respectively. It can be shown that $\mathbb{E}[\mathbf{h}_{ii'}^{j}(\boldsymbol{\beta}_j^*)] = \mathbb{E}[\mathbf{h}_{ii'}^{k}(\boldsymbol{\beta}_k^*)] = 0$; hence $\mathbf{h}_{ii'}^{jk}(\boldsymbol{\beta}_{j\vee k}^*)$ is also centered. We define

$$\mathbf{g}_{jk}(\boldsymbol{X}_i) := n/2 \cdot \mathbb{E}\big[\nabla L_{jk}\big(\boldsymbol{\beta}_{j\vee k}^*\big)\big|\boldsymbol{X}_i\big] = \mathbb{E}\big[\mathbf{h}_{ii'}^{jk}\big(\boldsymbol{\beta}_{j\vee k}^*\big)\big|\boldsymbol{X}_i\big] \quad \text{and} \tag{19}$$

$$\mathbf{U}_{jk} := \frac{2}{n} \sum_{i=1}^{n} \mathbf{g}_{jk}(\boldsymbol{X}_i) = \sum_{i=1}^{n} \mathbb{E}\big[\nabla L_{jk}\big(\boldsymbol{\beta}_{j\vee k}^*\big)\big|\boldsymbol{X}_i\big]. \tag{20}$$

Thus $2/n \cdot \mathbf{g}_{jk}\big(\boldsymbol{X}_i\big)$ is the projection of $\nabla L_{jk}\big(\boldsymbol{\beta}_{j\vee k}^*\big)$ onto the σ-filed generated by $\boldsymbol{X}_i$ and $\mathbf{U}_{jk}$ is the Hájek projection of $\nabla L_{jk}\big(\boldsymbol{\beta}_{j\vee k}^*\big)$. Under mild conditions, $\mathbf{U}_{jk}$ in (20) is a good approximation of $\nabla L_{jk}\big(\boldsymbol{\beta}_{j\vee k}^*\big)$, which enables us to characterize the limiting distribution of $S_{jk}(\boldsymbol{\beta}_{j\vee k}^*)$. We present the following assumption that guarantees the non-degeneracy of $\mathbf{g}_{jk}\big(\boldsymbol{X}_i\big)$.

Assumption 6 *Under Assumption 2, for $\mathbf{g}_{jk}(\boldsymbol{X}_i)$ defined in* (19)*, we denote the covariance matrix of $\mathbf{g}_{jk}(\boldsymbol{X}_i)$ as $\boldsymbol{\Sigma}^{jk} := \mathbb{E}[\mathbf{g}_{jk}(\boldsymbol{X}_i)\mathbf{g}_{jk}(\boldsymbol{X}_i)^T]$. We assume that there exists a constant $c_{\Sigma} > 0$ such that $\lambda_{\min}(\boldsymbol{\Sigma}^{jk}) \geq c_{\Sigma}$ for all $1 \leq j < k \leq d$.*

Assumption 6 requires the minimum eigenvalue of $\boldsymbol{\Sigma}^{jk}$ to be bounded away from 0, which implies $\mathrm{Var}(\mathbf{v}^T\mathbf{U}_{jk}) \geq 4c_{\Sigma}$ for all $\mathbf{v} \in \mathbb{R}^{2d-3}$ with $\|\mathbf{v}\|_2 = 1$. Thus, this assumption guarantees the asymptotic variance of $\sqrt{n}S_{jk}(\boldsymbol{\beta}_{j\vee k}^*)$ is bounded away from 0. We also present the following assumption that specifies the scaling of the Dantzig selector problem in (11).

Assumption 7 *We assume that $\mathbf{H}^j$ is invertible for all $j \in [d]$. In addition, we assume that there exist an integer $s_0^\star$ and a positive number w_0 such that $\|\mathbf{w}_{j,k}^*\|_0 \le s_0^\star - 1$ and $\|\mathbf{w}_{j,k}^*\|_1 \le w_0$. Besides, the regularization parameter λ_D in (11) satisfies $\lambda_D \asymp \max\{1, w_0\} s^* \lambda \log^2 d$. Moreover, we assume that*

$$\lim_{n\to\infty} (1+w_0+w_0^2) s^* \lambda \log^2 d = 0, \quad \lim_{n\to\infty} (1+w_0) s_0^\star \lambda_D = 0, \quad \text{and} \lim_{n\to\infty} \sqrt{n}(s^*+s_0^\star)\lambda\lambda_D = 0. \tag{21}$$

In addition, recall that we denote the s-sparse eigenvalues of $\mathbb{E}[\nabla^2 L_j(\boldsymbol{\beta}_j^*)]$ by $\rho_{j-}^*(s)$ and $\rho_{j+}^*(s)$. We further assume that there exist an integer $k_0^\star \ge s_0^\star$ and a positive number ν_* such that

$$\lim_{n\to\infty} k_0^\star \big(\log^9 d/n\big)^{1/2} = 0, \quad 0 < \nu_* \le \rho_{j-}^*(s_0^\star + k_0^\star) < \rho_{j+}^*(k_0^\star) \le \big(1 + 0.5 k_0^\star / s_0^\star\big)\nu_*, \quad 1 \le j \le d.$$

If we can treat w_0 as a constant, and k^* and $k_0^\star$ is of the same order of s^* and $s_0^\star$, respectively, Assumption 7 is reduced to $\lambda_D \asymp s^*\lambda \log^2 d$, $s_0^\star \lambda_D = o(1)$, $s^*\lambda\log^2 d = o(1)$, and $(s^*{+}s^\star)\lambda\lambda_D = o(n^{1/2})$. Since $\lambda \asymp \sqrt{\log d/n}$, we can choose $\lambda_D = Cs^*(\log^5 d/n)^{1/2}$ with a sufficiently large C, provided $(s^* + s_0^\star)(\log^9 d/n)^{1/2} = o(1)$, $s_0^\star s^*(\log^5 d/n)^{1/2} = o(1)$, and $(s^* + s_0^\star)s^* \log^3 d/n = o(n^{-1/2})$. Hence this condition is fulfilled if

$$\log d = o\Big(\min\big\{(\sqrt{n}/s^*)^{2/9}, (\sqrt{n}/s_0^\star)^{2/9}, (\sqrt{n}/s^{*2})^{1/3}, (\sqrt{n}/s^* s^\star)^{1/3}\big\}\Big).$$

Now we are ready to present the main theorem of composite pairwise score test.

Theorem 8 *Under the Assumptions 2, 4, 6 and 7, it holds uniformly for all $j \ne k$ and $j, k \in [d]$ that $\sqrt{n}\widehat{S}_{jk} = \sqrt{n} S_{jk}\big(\boldsymbol{\beta}_{j\vee k}^*\big) + o_{\mathbb{P}}(1)$. Furthermore, we let $\widehat{\boldsymbol{\beta}}'_{j\vee k} = (0, \widehat{\boldsymbol{\beta}}_{j\setminus k}^T, \widehat{\boldsymbol{\beta}}_{k\setminus j}^T)^T$ and define $\widehat{\boldsymbol{\Sigma}}^{jk} := n^{-1}\sum_{i=1}^n \big\{(n-1)^{-1}\sum_{i'\ne i} \mathbf{h}_{ii'}^{jk}(\widehat{\boldsymbol{\beta}}'_{j\vee k})\big\}^{\otimes 2}$, where $\mathbf{h}_{ii'}^{jk}\big(\boldsymbol{\beta}_{j\vee k}\big)$ is the kernel function of the second-order U-statistic $\nabla L_{jk}(\boldsymbol{\beta}_{j\vee k})$. In addition, we define $\widehat{\sigma}_{jk}$ by*

$$\widehat{\sigma}_{jk}^2 := \widehat{\boldsymbol{\Sigma}}_{jk,jk}^{jk} - 2\widehat{\boldsymbol{\Sigma}}_{jk,j\setminus k}^{jk}\widehat{\mathbf{w}}_{j,k} - 2\widehat{\boldsymbol{\Sigma}}_{jk,k\setminus j}^{jk}\widehat{\mathbf{w}}_{k,j} + \widehat{\mathbf{w}}_{j,k}^T\widehat{\boldsymbol{\Sigma}}_{j\setminus k,j\setminus k}^{jk}\widehat{\mathbf{w}}_{j,k} + \widehat{\mathbf{w}}_{k,j}^T\widehat{\boldsymbol{\Sigma}}_{k\setminus j,k\setminus j}^{jk}\widehat{\mathbf{w}}_{k,j}.$$

Then, under the null hypothesis $H_0\colon \beta_{jk}^* = 0$, we have $\sqrt{n}\widehat{S}_{jk}\big/(2\widehat{\sigma}_{jk}) \rightsquigarrow N(0,1)$.

By Theorem 8, to test the null hypothesis $H_0\colon \beta_{jk}^* = 0$ against the alternative hypothesis $H_1\colon \beta_{jk}^* \ne 0$, we reject H_0 if the studentized test statistic $\sqrt{n}\widehat{S}_{jk}\big/(2\widehat{\sigma}_{jk})$ is too extreme. Recall that the test function of the composite pairwise score test with significance level α is deboted by $\psi_{jk}(\alpha)$ in (13). The associated p-value is defined as $p_\psi^{jk} := 2\big[1 - \Phi\big(\big|\sqrt{n}\widehat{S}_{jk}\big/(2\widehat{\sigma}_{jk})\big|\big)\big]$. By Theorem 8, under H_0, we have

$$\lim_{n\to\infty} \mathbb{P}\big(\psi_{jk}(\alpha) = 1 \mid H_0\big) = \alpha \quad \text{and} \quad p_\psi^{jk} \rightsquigarrow \text{Unif}[0,1] \ \text{ under } H_0,$$

where Unif[0, 1] is the uniform distribution over [0, 1].

We note that our inferential approach is still valid if we replace $\widehat{\boldsymbol{\beta}}'_{j\vee k}$ in (18) by other estimators of $\boldsymbol{\beta}_{j\vee k}^*$, provided such an estimator converges to $\boldsymbol{\beta}_{j\vee k}^*$ at an appropriate statistical rate. Our theory still holds after simple modification on the proof when controlling the order of the remainder terms.

Remark 9 *There are a number of recent works on the uncertainty assessment for high dimensional linear models or generalized linear models with ℓ_1-penalty; see Lee et al. (2016); Lockhart et al. (2014); Belloni et al. (2012, 2013); Zhang and Zhang (2014); Javanmard and Montanari (2014); van de Geer et al. (2014). These works utilize the convexity and the Karush-Kuhn-Tuker conditions of the LASSO problem. Compared with these works, our pairwise score test is constructed using a nonconvex penalty function and is applicable to a larger model class. Ning et al. (2017b) consider the score test for ℓ_1-penalized semiparametric generalized linear models in the regression setting. Compared with this work, we adopt a composite score test with a nonconvex penalty and relax many technical assumptions including the bounded covariate assumption. For nonconvex penalizations, Fan and Lv (2011); Bradic et al. (2011) establish the asymptotic normality for the low dimensional and nonzero parameters in high dimensional models based on the oracle properties. However, their approach depends on the minimal signal strength assumption, which is not needed in our approach.*

5. Numerical Results

In this section we study the finite-sample performance of the proposed graph inference methods on both simulated and real-world datasets.

5.1. Simulation Studies

We first examine the numerical performance of the proposed pairwise score tests for the null hypothesis $H_0\colon \beta^*_{jk} = 0$. We simulate data from the following three settings:

(i) Gaussian graphical model. We set $n = 100$ and $d = 200$. The graph structure is a 4-nearest-neighbor graph, that is, for $j, k \in [d]$, $j \neq k$, node j is connected with node k if $|j - k| = 1, 2, d - 2, d - 1$. More specifically, we sample $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ from a Gaussian distribution $N_d(\mathbf{0}, \boldsymbol{\Sigma})$. For the precision matrix $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$, we set $\boldsymbol{\Theta}_{jj} = 1$, $\big|\boldsymbol{\Theta}_{jk}\big| = \mu \in [0, 0.25)$ for $|j - k| = 1, 2, d-2, d-1$ and $\boldsymbol{\Theta}_{jk} = 0$ for $2 \leq |j - k| \leq d - 2$. Note that μ denotes the signal strength of the graph inference problem and $\mu \leq 0.25$ ensures that $\boldsymbol{\Theta}$ is diagonal dominant and invertible.

(ii) Ising graphical model. We set $n = 100$ and $d = 200$. The graph structure is a 10×20 grid with the sparsity level $s^* = 4$. We use Markov Chain Monte Carlo method (MCMC) to simulate n data from an Ising model with joint distribution $p(\boldsymbol{x}) \propto \exp\big(\sum_{j \neq k} \beta^*_{jk} x_j x_k\big)$ (using the package `IsingSampler` (Epskamp, 2015)). We set $|\beta^*_{jk}| = \mu \in [0, 1]$ if there exists an edge connecting node j and node k, and $\beta^*_{jk} = 0$ otherwise.

(iii) Mixed graphical model. We set $n = 100$ and $d = 200$. The graph structure is a $10 \times 10 \times 2$ grid with the sparsity level $s^* = 5$. We set the nodes in the first layer to be binomial and nodes in the second layer to be Gaussian. We set $|\beta^*_{jk}| = \mu \in [0, 1]$ if there exists an edge connecting node j and node k, and $\beta^*_{jk} = 0$ otherwise. We refer to Lee and Hastie (2015) for details.

We denote the true parameters of the graphical models as $\{\beta^*_{jk}, j \neq k\}$. We also denote $\boldsymbol{\beta}^*_j = (\beta^*_{j1}, \dots, \beta^*_{jd})^T$. For the Gaussian graphical model, we have $\beta^*_{jk} = \boldsymbol{\Theta}_{jk}$. We first obtain a point estimate of $\boldsymbol{\beta}_j^*$ by solving (7) using Algorithm 1 with the capped-ℓ_1 penalty $p_\lambda(u) = \lambda \min\{u, \lambda\}$. The parameter λ is chosen by 10-fold cross validation as suggested by Ning et al. (2017b).

Recall that the form of the loss function $L_j(\boldsymbol{\beta}_j)$ is exactly the loss function for logistic regression, where we use Rademacher random variables $y_{ii'}$ as response and $y_{ii'}(x_{ij} - x_{i'j})\boldsymbol{\beta}_j^T(\boldsymbol{x}_{i\setminus j} - \boldsymbol{x}_{i'\setminus j})$ as covariates, Algorithm 1 can be easily implemented by using the ℓ_1-regularized logistic regression such as the `PICASSO` package (Ge et al., 2017). In particular, the algorithm converges quickly after a few iterations, indicating that it attains a good balance between computational efficiency and statistical accuracy. Once $\widehat{\boldsymbol{\beta}}_j$ is obtained, we solve the Dantzig-type problem (11) using $\widehat{\boldsymbol{\beta}}_j$ as input. We set the regularization parameter λ_D to be 1. In practice, the performance of the proposed method is not very sensitive to the choice of λ_D.

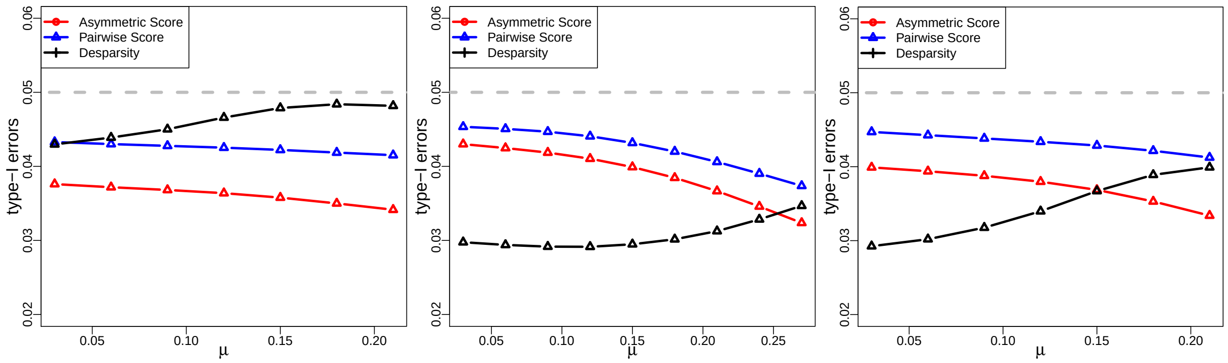
To examine the performance of our semiparametric modeling approach, we compare the pairwise score test with the desparsity method in van de Geer et al. (2014). Although this method is proposed for hypothesis tests in generalized linear models (GLMs), it can be adapted for graphical models by performing nodewise regression, assuming the base measures $\{f_j\}_{j\in[d]}$ are correctly specified. When testing $H_0\colon \beta_{jk}^* = 0$ with $j < k$, we apply the desparsity method with X_j and $\boldsymbol{X}_{\setminus j}$ being the response and covariates, respectively. Furthermore, to show that combining both $L_j(\boldsymbol{\beta}_j)$ and $L_k(\boldsymbol{\beta}_k)$ is beneficial for inferring β_{jk}^*, we also compare our method with the asymmetric score test, which constructs a score test statistic similar to that in (12) based solely on $L_j(\boldsymbol{\beta}_j)$.

To examine the validity of our method, we test $H_0\colon \beta_{jk}^* = 0$ versus $H_1\colon \beta_{jk}^* \neq 0$ for all (j, k). Recall that $\beta_{jk}^* = \mu$ when there is an edge. Here, we let μ increase from 0 to a sufficiently large number. We calculate the type I errors and powers as

$$\text{Type I error} = \frac{\text{the number of rejected hypotheses when there is no edge}}{d(d-1)/2 - \text{the total number of edges}},$$

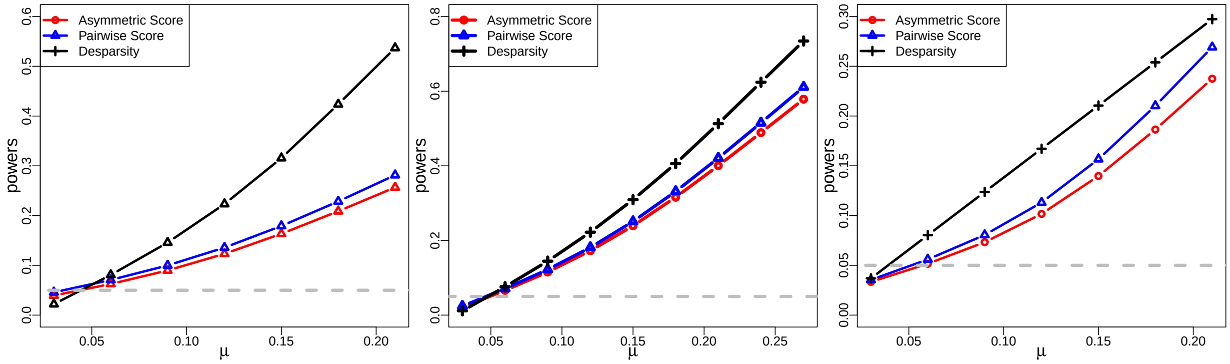
$$\text{Power} = \frac{\text{the number of rejected hypotheses when there is an edge}}{\text{the total number of edges}}.$$

We report the type I errors and powers of the hypothesis tests at the 0.05 significance level in Figure 1 and Figure 2, respectively. The simulation is repeated 100 times. As revealed in Figure 1, both the asymmetric and the pairwise score test achieve accurate type I errors, which is comparable to the desparsity method. Moreover, in terms of the power of the test, in Figure 2, the two score tests based on the loss function defined in (6) are less powerful than the desparsity method, which shows the loss of efficiency by only considering the relative rank. However, as shown in Figure 2-(b) and (c), the two score tests are nearly as powerful as the desparsity method in the Ising and mixed graphical models. In addition, we emphasize that for mixed graphical models the desparsity method needs to know the type (or distribution) of each nodes as a priori. Such phenomenon suggests that we may sacrifice little efficiency for model generality/robustness. Furthermore, comparing the performances of these two score tests, we see that the pairwise score test achieves uniformly higher power than the asymmetric one, which perfectly illustrates that taking into consideration of the symmetry of β_{jk}^* and β_{kj}^* may improve the inference accuracy.



(a). Gaussian graphical model. (b). Ising model. (c). Mixed graphical model.

Figure 1: Type-I errors of the composite pairwise score test, asymmetric score test, and the desparsity method for the three graphical models at the 0.05 significance level. These figures are based on 100 independent simulations.



(a). Gaussian graphical model. (b). Ising model. (c). Mixed graphical model.

Figure 2: Powers of the composite pairwise score test, asymmetric score test, and the desparsity method for the three graphical models at the 0.05 significance level. These figures are based on 100 independent simulations.

5.2. Real Data Analysis

We then apply the proposed methods to analyze a publicly available dataset named `Computer Audition Lab 500-Song (CAL500)` dataset (Turnbull et al., 2008). The data can be obtained from the `Mulan` database (Tsoumakas et al., 2011). The `CAL500` dataset consists of 502 popular music tracks each of which is annotated by at least three listeners. The attributes of this dataset include two subsets: (i) continuous numerical features extracted from the time series of the audio signal and (ii) discrete binary labels assigned by human listeners to give semantic descriptions of the song. For each music track, short time Fourier transform is implemented for a sequence of half-overlapping 23ms time windows over the song's digital audio file. This procedure generates four types of continuous features: *spectral centroids, spectral flux, zero crossings* and a time series of Mel-frequency cepstral coefficient

(MFCC). For the MFCC vectors, every consecutive 502 short time windows are grouped together as a block window to produce the following four types of features: (i) overall mean of MFCC vectors in each block window, (ii) mean of standard deviations of MFCC vectors in each block window, (iii) standard deviation of the means of MFCC vectors in each block window, and (iv) standard deviation of the standard deviations of MFCC vectors in each block window. More details on the feature extraction can be found in Tzanetakis and Cook (2002). In addition to these continuous variables, binary variables in the `CAL500` dataset include a 174-dimensional array indicating the existence of each annotation. These 174 annotations can be grouped into six categories: emotions (36 variables), instruments (33), usages (15), genres (47), song characteristics (27) and vocal types (16). Our goal is to infer the association between these different types of variables using graphical models. This dataset has been analyzed in Cheng et al. (2017) where they exploit a nodewise group-LASSO regression to estimate the graph structure. In what follows, we use the proposed pairwise score test to examine the graph structure.

Similar to Turnbull et al. (2008) and Cheng et al. (2017), we only keep the MFCC features because they can be interpreted as the amplitude of the audio signal and the other continuous features are not readily interpretable. Unlike Cheng et al. (2017), we keep all the binary labels. Thus the processed dataset has $n = 502$ data points of dimension $d = 226$ with 52 continuous variables and 174 binary variables. We apply the pairwise score test to each pair of variables to determine the presence of an edge between them. The p-values for the null hypothesis that two variables are conditionally independence given the rest of variables are calculated. We then apply the Bonferroni correction to control the familywise error rate at 0.05. We set the nonconvex penalty function in optimization problem (7) to be capped-ℓ_1 penalty $p_\lambda(u) = \lambda \min\{u, \lambda\}$ with the regularization parameter λ selected by 10-fold cross-validation as in the previous section.

We compare the pairwise score test with the desparsity method and the asymmetric score test, which are constructed in the same way as in the simulation. We present the fitted graphs obtained by these three methods in Figure 3-(a)–(c), where we plot the connected components and omit the singletons. Moreover, in Figure 3-(d), we plot the intersection of these three graphs. To better display the graphical structure, we use a square to represent each type of 13 MFCC features respectively. If a node is connected to any node within the group of variables in a MFCC node, then we draw an edge. We use circles to represent the binary variables and use different colors to indicate their categories. The obtained graphs have some interesting properties. While all three tests create different graphs, the graphs obtained by the pairwise score test and the asymmetric score test have more common edges, which agrees with our simulation results. Indeed, our test can correct the inconsistency of the asymmetric score test, in the sense that the asymmetric score tests for $\beta^*_{jk} = 0$ and $\beta^*_{kj} = 0$ may yield different test results. To show this inconsistency problem, we also plot the graph obtained by the asymmetric score test based on the loss function $L_k(\boldsymbol{\beta}_k)$ in Figure 4 in the appendix. Comparing with Figure 3-(b), we can see that the asymmetric score test indeed leads to many contradictory edges.

In Figure 3, both the pairwise score test and this asymmetric score test discover that songs that are danceable (circle 92) are suited for parties (circle 93), but such a connection is not found by the desparsity method. This is also true for the connection between the rapping vocals (circle 119) and the rap genre (circle 48) and the edge between strong vocals (circle 122) and songs with strong emotions (circle 19). Moreover, in all three graphs, the continuous features are densely connected within themselves, which is similar to the results in Cheng et al. (2017). All three tests find that the noisiness of the music (square 4) is connected with the quality of songs (circle 85). Furthermore, the common edges connecting two binary variables also display interesting patterns. For instance, we find that awakening emotions (circle 6) are connected with soothing emotions (circle 8); laid-back emotions (circle 14) are connected with songs with high energy (circle 32); sad emotions (circle 20) are connected with songs with positive feelings (circle 84); songs with female lead vocals (circle 62) are connected with those with male lead vocals (circle 66). In addition, songs using drum sets (circle 59) are connected with the electronica genre (circle 46), which is also connected with the acoustic texture (circle 88). All these edges have fairly intuitive explanations.

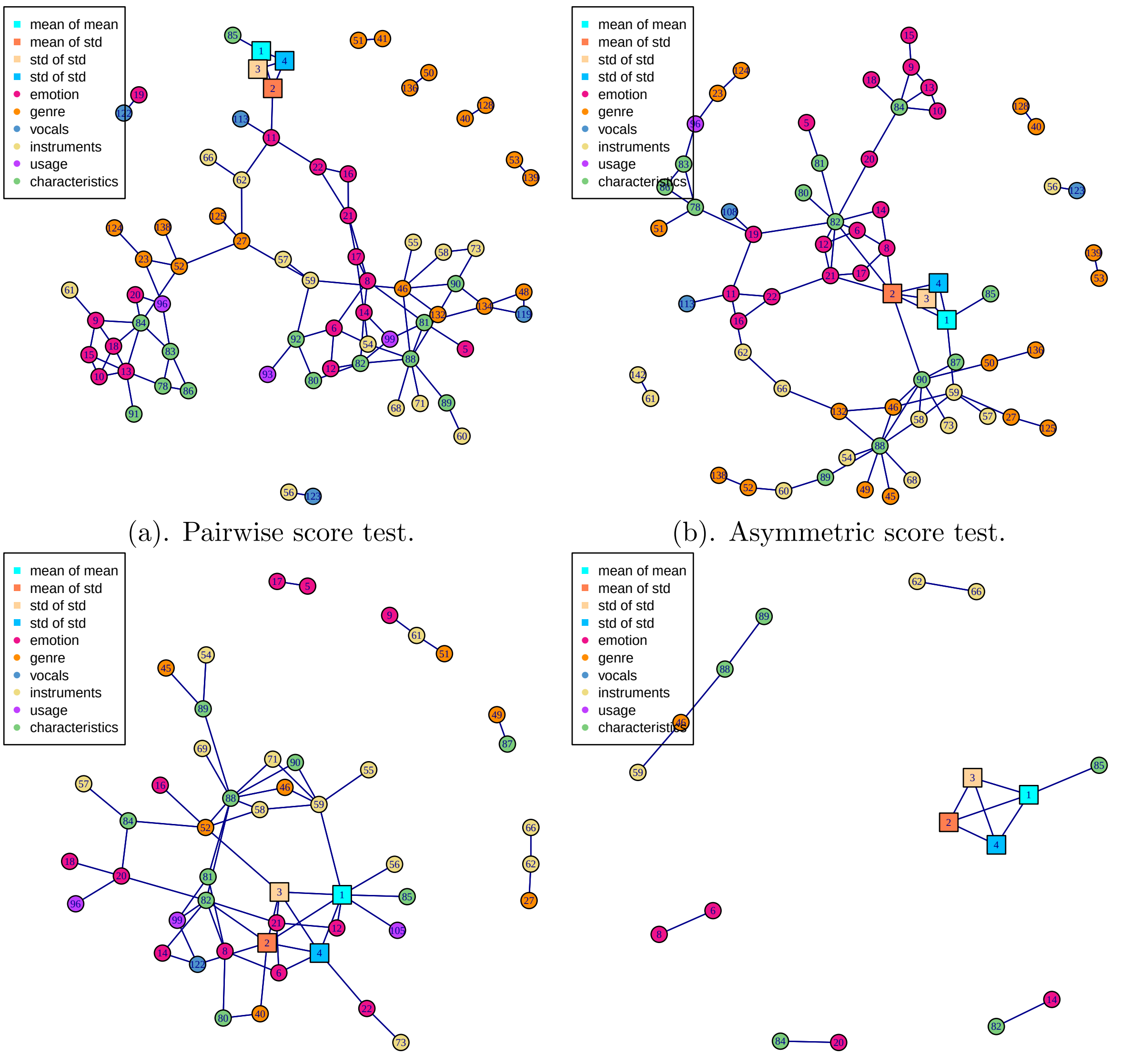
In summary, the proposed method reveals some interesting associations between these variables and can be used as a useful complement to analyze high dimensional datasets with more complex distributions.

6. Conclusion

We propose an integrated framework for uncertainty assessment of a new semiparametric exponential family graphical model. The novelty of our model is that the base measures of each nodewise conditional distribution are treated as unknown nuisance functions. Towards the goal of uncertainty assessment, we first adopt the adaptive multi-stage relaxation algorithm to perform the parameter estimation. Then we propose a composite pairwise score test of the graph structure. Our method provides a rigorous justification for the uncertainty assessment, and is further supported by extensive numerical results. In a followup paper (Tan et al., 2016), the proposed model is further extended to account for the unobserved latent variables in the graphical model.

Acknowledgments

The authors are grateful for the support of NSF CAREER Award DMS1454377, NSF IIS1408910, NSF IIS1332109, NIH R01MH102339, NIH R01GM083084, and NIH R01HG06841.



(a). Pairwise score test.

(b). Asymmetric score test.

(c). Desparsity method.

(d). The common edges.

Figure 3: Estimated graphs in the CAL500 dataset inferred by the pairwise score test, asymmetric score test, and the desparsity method. We plot the connected components of the estimated graph. In (a)-(c) we plot the graphs obtained by these three approaches, respectively, and plot the common edges in (d). For better illustration, we only plot the connected components, combine the same type of continuous variables, display them as a square and draw each binary variable as a circle. The edges of the estimated graph show the association between these variables.

Appendix A. Proof of the Main Results

In this appendix we lay out the proof of the main results. In §A.1 we prove the result of parameter estimation. The proof is based an induction argument that Algorithm 1 keeps penalizing most of the irrelevant features and gradually reduces the bias in relevant features.

A.1. Proof of Theorem 5

Proof We only need to prove the theorem for one node $j \in [d]$, the proof is identical for the rest. To begin with, we first define a few index sets that play a significant role in our analysis. For all $j \in [d]$, we let $S_j := \{(j,k)\colon \beta^*_{jk} \neq 0, k \in [d]\}$ be the support of $\boldsymbol{\beta}^*_j$. For the number of iterations $\ell = 1, 2, \ldots$, let $G_j^\ell := \big\{(j,k) \notin S_j\colon \lambda_{jk}^{(\ell-1)} \geq p'_\lambda(c_2\lambda), k \in [d]\big\}$. By condition (C.3) of the penalty function $p_\lambda(u)$ (see §3.2), we have $p'_\lambda(c_2\lambda) \geq 0.91\lambda$. In addition, we let J_j^ℓ be the largest k^* components of $\big[\widehat{\boldsymbol{\beta}}_j^{(\ell)}\big]_{G_j^\ell}$ in absolute value where k^* is defined in Assumption 4. In addition, we let $I_j^\ell = (G_j^\ell)^c \cup J_j^\ell$. Moreover, for notational simplicity, we denote $\big[\boldsymbol{\beta}_j\big]_{G_j^\ell}, \big[\boldsymbol{\beta}_j\big]_{G_j^\ell}$ and $\big[\boldsymbol{\beta}_j\big]_{I_j^\ell}$ as $\boldsymbol{\beta}_{G_j^\ell}, \boldsymbol{\beta}_{J_j^\ell}$ and $\boldsymbol{\beta}_{I_j^\ell}$ respectively when no ambiguity arises.

The key point of the proof is to show that the complement of G_j^ℓ is not too large. To be more specific, we show that $\big|(G_j^\ell)^c\big| \leq 2s^*$ for $\ell \geq 1$. Since $S_j \subset (G_j^\ell)^c$, $(G_j^\ell)^c \leq 2s^*$ implies $\big|(G_j^\ell)^c - S_j\big| \leq s^*$. Note that G_j^ℓ is the set of irrelevant features that are heavily penalized in the ℓ-th iteration of the algorithm, $(G_j^\ell)^c - S$ being a small set indicates that the most of the irrelevant features are heavily penalized in each step. We show that $\big|(G^\ell)^c\big| \leq 2s^*$ for each $\ell \geq 1$ by induction.

For $\ell = 1$, we have $G_j^1 = S_j^c$ because $\lambda_{jk}^{(0)} = \lambda$ for all $j, k \in [d]$. Hence $\big|(G_j^1)^c\big| \leq s^*$. Now we assume that $|\big(G_j^\ell\big)^c| \leq 2s^*$ for some integer ℓ and our goal is to prove that $|\big(G_j^{\ell+1}\big)^c| \leq 2s^*$. Our proof is based on three technical lemmas. The first lemma shows that the regularization parameter λ in (7) is of the same order as $\|\nabla L_j(\boldsymbol{\beta}_j^*)\|_\infty$.

Lemma 10 *Under Assumptions 2 and 4, there exists a positive constants K such that, it holds with probability at least $1 - (2d)^{-1}$ that*

$$\big\|\nabla L_j(\boldsymbol{\beta}_j^*)\big\|_\infty \leq K\sqrt{\log d/n}, \quad \forall j \in [d]. \tag{22}$$

Proof See §C.1 for a proof. ■

By this lemma, we conclude that the regularization parameter $\lambda \geq 25\big\|\nabla L_j(\boldsymbol{\beta}_j^*)\big\|_\infty$ with high probability. The following lemma bounds the ℓ_1- and ℓ_2-norms of $\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}_j^*$ by the norms of its subvector under the induction assumption that $\big|(G_j^\ell)^c\big| \leq 2s^*$.

Lemma 11 *Letting the index sets S_j, G_j^ℓ, J_j^ℓ and I_j^ℓ be defined as above, we denote $\widetilde{G}_j^\ell := \big(G_j^\ell\big)^c$. Under the assumption that $|G_j^\ell| \leq 2s^*$, we have*

$$\big\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}_j^*\big\|_2 \leq 2.2\big\|\widehat{\boldsymbol{\beta}}_{I_j^\ell}^{(\ell)} - \boldsymbol{\beta}_{I_j^\ell}^*\big\|_2 \ \ \textit{and} \ \ \big\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}_j^*\big\|_1 \leq 2.2\big\|\widehat{\boldsymbol{\beta}}_{\widetilde{G}_j^\ell}^{(\ell)} - \boldsymbol{\beta}_{\widetilde{G}_j^\ell}^*\big\|_1. \tag{23}$$

Proof See §C.2 for a detailed proof. ■

The next lemma guarantees that $\widehat{\boldsymbol{\beta}}_j^{(\ell)}$ stays in the ℓ_1-ball centered at $\boldsymbol{\beta}_j^*$ with radius r for $\ell \geq 1$ where r appears in Assumption 4. Moreover, by showing this property of Algorithm 1 , we obtain a crude rate for parameter estimation. We summarized this result in the next lemma.

Lemma 12 *For $\ell \geq 1$ and $j \in [d]$, we denote $\boldsymbol{\lambda}_{S_j}^{(\ell)} := (\lambda_{jk}^{(\ell)}, (j,k) \in S_j)^T$. Assuming that $\big|(G_j^\ell)^c\big| \leq 2s^*$, it holds with probability at least $1 - d^{-1}$ that, for all $j \in [d]$, the estimators $\widehat{\boldsymbol{\beta}}_j^{(\ell)}$ obtained in each iteration of Algorithm 1 satisfy*

$$\|\widehat{\boldsymbol{\beta}}_{I_j^\ell}^{(\ell)} - \boldsymbol{\beta}_{I_j^\ell}^*\|_2 \leq 10\rho_*^{-1}\Big[\|\nabla_{\widetilde{G}_j^\ell} L_j(\boldsymbol{\beta}_j^*)\|_2 + \|\boldsymbol{\lambda}_{S_j}^{(\ell-1)}\|_2\Big], \quad \widetilde{G}_j^\ell := (G_j^\ell)^c. \tag{24}$$

This implies the following crude rates of convergence for $\widehat{\boldsymbol{\beta}}_j^{(\ell)}$:

$$\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}_j^*\|_2 \leq 24\rho_*^{-1}\sqrt{s^*}\lambda \quad \text{and} \quad \|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}_j^*\|_1 \leq 33\rho_*^{-1}s^*\lambda. \tag{25}$$

Proof See §C.3 for a detailed proof. ■

Now we show that $\widetilde{G}_j^{\ell+1} = (G_j^{\ell+1})^c$ satisfies $|\widetilde{G}_j^{\ell+1}| \leq 2s^*$, which concludes our induction. Letting $A := (G_j^{\ell+1})^c - S_j$, by the definition of $G_j^{\ell+1}$, $(j,k) \in A$ implies that $(j,k) \notin S_j$ and $p'_\lambda\big(|\widehat{\beta}_{jk}^{(\ell)}|\big) \leq p'_\lambda(c_2\lambda)$. Hence by the concavity of $p_\lambda(\cdot)$, for any $(j,k) \in A$, $\big|\widehat{\beta}_{jk}^{(\ell)}\big| \geq c_2\lambda$. Therefore we have

$$\sqrt{|A|} \leq \|\widehat{\boldsymbol{\beta}}_A^{(\ell)}\|_2/(c_2\lambda) = \|\widehat{\boldsymbol{\beta}}_A^{(\ell)} - \boldsymbol{\beta}_A^*\|_2/(c_2\lambda) \leq 24\rho_*^{-1}\sqrt{s^*}/c_2 \leq \sqrt{s^*}, \tag{26}$$

where the first inequality follows from $|A| \leq \sum_{(j,k)\in A}\big|\widehat{\beta}_{jk}^{(\ell)}\big|^2/(c_2\lambda)^2$. Note that (26) implies that $\big|(G_j^{\ell+1})^c\big| \leq 2s^*$. Therefore by induction, $\big|(G_j^\ell)^c\big| \leq 2s^*$ for any $\ell \geq 1$.

Now we have shown that for $\ell \geq 1$ and $j \in [d]$, $\big|(G_j^\ell)^c\big| \leq 2s^*$ and the crude statistical rates (25) hold. In what follows, we derive the more refined rates (15) and (16).

A refined bound for $\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}_j^*\|_2$ and $\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}_j^*\|_1$: For notional simplicity, we let $\boldsymbol{\delta}^{(\ell)} = \widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}_j^*$ and omit subscript j in S_j, G_j^ℓ, J_j^ℓ and I_j^ℓ. We also denote $\widetilde{G}^\ell := (G^\ell)^c$. We first derive a recursive bound that links $\|\boldsymbol{\delta}_{I^\ell}^{(\ell)}\|_2$ to $\|\boldsymbol{\delta}_{I^{\ell-1}}^{(\ell-1)}\|_2$. Note that by (23), $\|\boldsymbol{\delta}^{(\ell)}\|_1 \leq 2.2\|\boldsymbol{\delta}_{\widetilde{G}^\ell}^{(\ell)}\|_1 \leq 2.2\sqrt{2s^*}\|\boldsymbol{\delta}_{\widetilde{G}^\ell}^{(\ell)}\|_2$. Hence we only need to control $\|\boldsymbol{\delta}_{I^\ell}^{(\ell)}\|_2$ to obtain the statistical rates of convergence for $\widehat{\boldsymbol{\beta}}_j^{(\ell)}$. By triangle inequality,

$$\|\nabla_{\widetilde{G}^\ell} L_j(\boldsymbol{\beta}_j^*)\|_2 \leq \|\nabla_S L_j(\boldsymbol{\beta}_j^*)\|_2 + \sqrt{|\widetilde{G}^\ell - S|}\|\nabla L_j(\boldsymbol{\beta}_j^*)\|_\infty.$$

Since $\lambda > 25\|\nabla L_j(\boldsymbol{\beta}_j^*)\|_\infty$, (26) implies that

$$\|\nabla_{\widetilde{G}^\ell} L_j(\boldsymbol{\beta}_j^*)\|_2 \leq \|\nabla_S L_j(\boldsymbol{\beta}_j^*)\|_2 + \|\boldsymbol{\delta}_A^{(\ell-1)}\|_2/(25c_2), \tag{27}$$

where $A := (G^\ell)^c - S \subset I^\ell$. Thus (27) can be written as

$$\|\nabla_{\widetilde{G}^\ell} L_j(\boldsymbol{\beta}_j^*)\|_2 \leq \|\nabla_S L_j(\boldsymbol{\beta}_j^*)\|_2 + \|\boldsymbol{\delta}_{I^\ell}^{(\ell-1)}\|_2/(25c_2). \tag{28}$$

Also notice that $\forall \beta_{jk} \in \mathbb{R}$, if $|\beta_{jk} - \beta^*_{jk}| \geq c_2\lambda$,

$$p'_\lambda(|\beta_{jk}|) \leq \lambda \leq |\beta_{jk} - \beta^*_{jk}|/c_2;$$

otherwise we have $|\beta^*_{jk}| - |\beta_{jk}| \leq |\beta_{jk} - \beta^*_{jk}| < c_2\lambda$ and thus $p'_\lambda(|\beta_{jk}|) \leq p'_\lambda\big(|\beta^*_{jk}| - c_2\lambda\big)$ by the concavity of $p_\lambda(\cdot)$. Hence the following inequality always holds:

$$p'_\lambda(|\beta_{jk}|) \leq p'_\lambda\big(|\beta^*_{jk}| - c_2\lambda\big) + |\beta_{jk} - \beta^*_{jk}|/c_2. \tag{29}$$

Applying (29) to $\widehat{\boldsymbol{\beta}}_j^{(\ell-1)}$ we have

$$\big\|\boldsymbol{\lambda}_S^{(\ell-1)}\big\|_2 \leq \Big[\sum_{(j,k)\in S} p'_\lambda\big(|\beta^*_{jk}| - c_2\lambda\big)^2\Big]^{1/2} + \Big[\sum_{(j,k)\in S} |\widehat{\beta}_{jk}^{(\ell-1)} - \beta^*_{jk}|^2\Big]^{1/2}\Big/ c_2,$$

which leads to

$$\big\|\boldsymbol{\lambda}_S^{(\ell-1)}\big\|_2 \leq \Big[\sum_{(j,k)\in S} p'_\lambda\big(|\beta^*_{jk}| - c_2\lambda\big)^2\Big]^{1/2} + \big\|\boldsymbol{\delta}_{I^{\ell-1}}^{(\ell-1)}\big\|_2/c_2. \tag{30}$$

By (24), (28) and (30) we obtain

$$\big\|\boldsymbol{\delta}_{I^\ell}^{(\ell)}\big\|_2 \leq 10\rho_*^{-1}\big[\big\|\nabla_S L_j(\boldsymbol{\beta}^*_j)\big\|_2 + \Upsilon_j\big] + \gamma\big\|\boldsymbol{\delta}_{I^{\ell-1}}^{(\ell-1)}\big\|_2,$$

where $\gamma := 11(c_2\rho_*)^{-1}$ and we define $\Upsilon_j := \big[\sum_{(j,k)\in S} p'_\lambda\big(|\beta^*_{jk}| - c_2\lambda\big)^2\big]^{1/2}$ for notational simplicity. Note that since $c_2 \geq 24\rho_*^{-1}$, we have $\gamma < 1$. By recursion we obtain

$$\big\|\boldsymbol{\delta}_{I^\ell}^{(\ell)}\big\|_2 \leq 10\varrho\big[\big\|\nabla_S L_j(\boldsymbol{\beta}^*_j)\big\|_2 + \Upsilon_j\big] + \gamma^{\ell-1}\big\|\boldsymbol{\delta}_{I^1}^{(1)}\big\|_2, \tag{31}$$

where $\varrho := \rho_*^{-1}\cdot(1-\gamma)^{-1} = c_2(c_2\rho_* - 11)^{-1}$. Using $\big\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}^*_j\big\|_2 \leq 2.2\big\|\widehat{\boldsymbol{\beta}}_{I_j^\ell}^{(\ell)} - \boldsymbol{\beta}^*_{I_j^\ell}\big\|_2$, we can bound $\big\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}^*_j\big\|_2$ by

$$\big\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}^*_j\big\|_2 \leq 22\varrho\big[\big\|\nabla_{S_j} L_j(\boldsymbol{\beta}^*_j)\big\|_2 + \Upsilon_j\big] + 2.2\gamma^{\ell-1}\big\|\boldsymbol{\delta}_{I_j^1}^{(1)}\big\|_2.$$

Note that for $\ell = 1$, by (24) we have

$$\big\|\boldsymbol{\delta}_{I_j^1}^{(1)}\big\|_2 \leq 10\rho_*^{-1}\sqrt{s^*}\big[\lambda + \sqrt{2}\big\|\nabla L_j(\boldsymbol{\beta}^*_j)\big\|_\infty\big] \leq 11\rho_*^{-1}\sqrt{s^*}\lambda = c_2\gamma\sqrt{s^*}\lambda. \tag{32}$$

then we establish the following bound for $\big\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}^*_j\big\|_2$:

$$\big\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}^*_j\big\|_2 \leq 22\varrho\big[\big\|\nabla_{S_j} L_j(\boldsymbol{\beta}^*_j)\big\|_2 + \Upsilon_j\big] + 2.2c_2\sqrt{s^*}\lambda\gamma^\ell. \tag{33}$$

Similarly, by $\big\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}^*_j\big\|_1 \leq 2.2\sqrt{2s^*}\big\|\widehat{\boldsymbol{\beta}}_{I_j^\ell}^{(\ell)} - \boldsymbol{\beta}^*_{I_j^\ell}\big\|_2$, we obtain a bound on $\big\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}^*_j\big\|_1$:

$$\big\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}^*_j\big\|_1 \leq 32\sqrt{s^*}\varrho\big[\big\|\nabla_{S_j} L_j(\boldsymbol{\beta}^*_j)\big\|_2 + \Upsilon_j\big] + 2.2\gamma^{\ell-1}\sqrt{2s^*}\big\|\boldsymbol{\delta}_{I_j^1}^{(1)}\big\|_2. \tag{34}$$

By (32) we have $2.2\sqrt{2s^*}\|\boldsymbol{\delta}^{(1)}_{I^1_j}\|_2 \le 3.2c_2\gamma s^*\lambda$, then the right-hand side of (34) can be bounded by

$$\|\widehat{\boldsymbol{\beta}}^{(\ell)}_j - \boldsymbol{\beta}^*_j\|_1 \le 32\sqrt{s^*}\varrho\big[\|\nabla_{S_j}L_j(\boldsymbol{\beta}^*_j)\|_2 + \Upsilon_j\big] + 3.2c_2 s^*\lambda\gamma^{\ell}. \tag{35}$$

Therefore (15) and (16) can be implied by (33) and (35) respectively. Moreover, by Lemma 12, we conclude that the statistical rates (33) and (35) hold for all $j \in [d]$ with probability at least $1 - d^{-1}$. ∎

A.2. Proof of Theorem 8

Proof We first remind the reader that, for $1 \le j \ne k \le d$, we denote

$$\boldsymbol{\beta}_{j\backslash k} = (\beta_{j1}, \dots, \beta_{jj-1}, \beta_{jj+1}, \dots, \beta_{jk-1}, \beta_{jk+1}, \dots, \beta_{jd})^T \in \mathbb{R}^{d-2},$$

$\boldsymbol{\beta}_{j\vee k} = (\beta_{jk}, \boldsymbol{\beta}_{j\backslash k}, \boldsymbol{\beta}_{k\backslash j})^T \in \mathbb{R}^{2d-3}$ and $\widehat{\boldsymbol{\beta}}'_{j\vee k} = (0, \widehat{\boldsymbol{\beta}}_{j\backslash k}, \widehat{\boldsymbol{\beta}}_{k\backslash j})^T$. In addition, we define $\sigma^2_{jk} = \boldsymbol{\Sigma}^{jk}_{jk,jk} - 2\boldsymbol{\Sigma}^{jk}_{jk,j\backslash k}\mathbf{w}^*_{j,k} - 2\boldsymbol{\Sigma}^{jk}_{jk,k\backslash j}\mathbf{w}^*_{k,j} + \mathbf{w}^{*T}_{j,k}\boldsymbol{\Sigma}^{jk}_{j\backslash k,j\backslash k}\mathbf{w}^*_{j,k} + \mathbf{w}^{*T}_{k,j}\boldsymbol{\Sigma}^{jk}_{k\backslash j,k\backslash j}\mathbf{w}^*_{k,j}$. To prove the theorem our goal is to prove the following two arguments:

$$\lim_{n\to\infty}\max_{j<k}\sqrt{n}\big|\widehat{S}_{jk} - S_{jk}(\boldsymbol{\beta}^*_{j\vee k})\big| = 0 \quad\text{and}\quad \lim_{n\to\infty}\max_{j<k}|\widehat{\sigma}_{jk} - \sigma_{jk}| = 0. \tag{36}$$

Note that by Lemma 14, σ^2_{jk} is the asymptotic variance of $\sqrt{n}/2\cdot S_{jk}(\boldsymbol{\beta}^*_{j\vee k})$. Thus combining (36) and Slutsky's theorem yields the theorem. By the the expression of $S_{jk}(\boldsymbol{\beta}^*_{j\vee k})$ and $\widehat{S}_{jk}$ in (17) and (18), under null hypothesis, for a fixed pair of nodes j and k, we have $\widehat{S}_{jk} - S_{jk}(\boldsymbol{\beta}^*_{j\vee k}) = I_{1j} + I_{2j} + I_{1k} + I_{2k}$ where I_{1j} and I_{2j} are defined as

$$\begin{aligned}
&I_{1j} := \big[\nabla_{jk}L_j(\widehat{\boldsymbol{\beta}}'_j) - \nabla_{jk}L_j(\boldsymbol{\beta}^*_j)\big] - \widehat{\mathbf{w}}^T_{j,k}\big[\nabla_{j\backslash k}L_j(\widehat{\boldsymbol{\beta}}'_j) - \nabla_{j\backslash k}L_j(\boldsymbol{\beta}^*_j)\big] \quad\text{and}\\
&I_{2j} := (\mathbf{w}^*_{j,k} - \widehat{\mathbf{w}}_{j,k})^T\nabla_{j\backslash k}L_j(\boldsymbol{\beta}^*_j);
\end{aligned}$$

whereas I_{1k} and I_{2k} are defined by interchanging j and k in I_{1j} and I_{2j}:

$$\begin{aligned}
&I_{1k} := \big[\nabla_{kj}L_k(\widehat{\boldsymbol{\beta}}'_k) - \nabla_{jk}L_k(\boldsymbol{\beta}^*_k)\big] - \widehat{\mathbf{w}}^T_{k,j}\big[\nabla_{k\backslash j}L_k(\widehat{\boldsymbol{\beta}}'_k) - \nabla_{k\backslash j}L_k(\boldsymbol{\beta}^*_k)\big] \quad\text{and}\\
&I_{2k} := (\mathbf{w}^*_{k,j} - \widehat{\mathbf{w}}_{k,j})^T\nabla_{k\backslash j}L_j(\boldsymbol{\beta}^*_k).
\end{aligned}$$

We first bound I_{1j}. Recall that $\widehat{\boldsymbol{\beta}}'_j = (0, \widehat{\boldsymbol{\beta}}_{j\backslash k})^T$. Note that under the null hypothesis, $\beta^*_{jk} = 0$, by the Mean-Value Theorem, there exists a $\widetilde{\boldsymbol{\beta}}_{j\backslash k} \in \mathbb{R}^{d-2}$ in the line segment between $\widehat{\boldsymbol{\beta}}_{j\backslash k}$ and $\boldsymbol{\beta}^*_{j\backslash k}$ such that

$$I_{1j} = \big[\widetilde{\boldsymbol{\Lambda}}_{jk,j\backslash k} - \widehat{\mathbf{w}}^T_{j,k}\widetilde{\boldsymbol{\Lambda}}_{j\backslash k,j\backslash k}\big]\big(\widehat{\boldsymbol{\beta}}_{j\backslash k} - \boldsymbol{\beta}^*_{j\backslash k}\big),$$

where $\widetilde{\boldsymbol{\Lambda}} := \nabla^2 L_j(0, \widetilde{\boldsymbol{\beta}}_{j\backslash k})$. We let $\boldsymbol{\delta} := \widehat{\boldsymbol{\beta}}'_j - \boldsymbol{\beta}^*_j$ and denote $\nabla^2 L_j(\widehat{\boldsymbol{\beta}}'_j)$ and $\nabla^2(\boldsymbol{\beta}^*_j)$ as $\boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda}^*$ respectively. From the definition of Dantzig selector we obtain

$$\begin{aligned}
|I_{1j}| \le\ &\underbrace{\|\boldsymbol{\Lambda}_{jk,j\backslash k} - \widehat{\mathbf{w}}^T\boldsymbol{\Lambda}_{j\backslash k,j\backslash k}\|_\infty\|\boldsymbol{\delta}_{j\backslash k}\|_1}_{I_{11}} + \underbrace{\|\boldsymbol{\Lambda}_{jk,j\backslash k} - \widetilde{\boldsymbol{\Lambda}}_{jk,j\backslash k}\|_\infty\|\boldsymbol{\delta}_{j\backslash k}\|_1}_{I_{12}}\\
&+ \underbrace{\|\widehat{\mathbf{w}}^T(\boldsymbol{\Lambda}_{j\backslash k,j\backslash k} - \widetilde{\boldsymbol{\Lambda}}_{j\backslash k,j\backslash k})\boldsymbol{\delta}_{j\backslash k}\|_\infty}_{I_{13}}.
\end{aligned}$$

Theorem 5 implies that $\|\boldsymbol{\delta}\|_1 \leq Cs^*\lambda$ with probability tending to 1 for some constant $C > 0$. Then by the definition of Dantzig selector, $I_{11} \leq Cs^*\lambda\lambda_D$. with high probability. Moreover, the constant C is the same for all (j,k). By assumption 7, $I_{11} = o(n^{-1/2})$ with probability tending to one.

For term I_{12}, Hölder's inequality implies that

$$I_{12} \leq \|\boldsymbol{\Lambda}_{jk,j\backslash k} - \widetilde{\boldsymbol{\Lambda}}_{jk,j\backslash k}\|_\infty \|\boldsymbol{\delta}_{j\backslash k}\|_1. \tag{37}$$

By Lemma 26 we obtain

$$\|\boldsymbol{\Lambda} - \widetilde{\boldsymbol{\Lambda}}\|_\infty \leq \|\boldsymbol{\Lambda} - \boldsymbol{\Lambda}^*\|_\infty + \|\boldsymbol{\Lambda}^* - \widetilde{\boldsymbol{\Lambda}}\|_\infty \leq 2Cs^*\lambda \log^2 d. \tag{38}$$

Therefore combining (37) and (38) we have

$$I_{12} \leq 2C{s^*}^2\lambda^2 \log^2 d \lesssim s^*\lambda\lambda_D \quad \text{uniformly for } 1 \leq j < k \leq d.$$

Similarly by Hölder's inequality, we have

$$I_{13} \leq \|\widehat{\mathbf{w}}_{j,k}\|_1 \|\boldsymbol{\Lambda} - \widetilde{\boldsymbol{\Lambda}}\|_\infty \|\boldsymbol{\delta}\|_1. \tag{39}$$

Notice that by the optimality of $\widehat{\mathbf{w}}_{j,k}$, $\|\widehat{\mathbf{w}}_{j,k}\|_1 \leq \|\mathbf{w}^*_{j,k}\|_1 \leq w_0$. Combining (39) and (38) we have

$$I_{13} \leq Cw_0{s^*}^2\lambda^2 \log^2 d \lesssim s^*\lambda\lambda_D \quad \text{uniformly for } 1 \leq j < k \leq d.$$

where we use the fact that $\lambda_D \gtrsim \max\{1, w_0\}s^*\lambda \log^2 d$. Therefore we conclude that for all $j \in [d]$, $|I_{1j}| \lesssim s^*\lambda\lambda_D = o_{\mathbb{P}}(n^{-1/2})$. For I_{2j}, Hölder's inequality implies that $|I_{2j}| \leq \|\mathbf{w}^*_{j,k} - \widehat{\mathbf{w}}_{j,k}\|_1 \big\|\nabla L_j(\boldsymbol{\beta}^*_j)\big\|_\infty$. To control $\|\mathbf{w}^*_{j,k} - \widehat{\mathbf{w}}_{j,k}\|_1$, we need to the following lemma to obtain the estimation error of the Dantzig selector $\widehat{\mathbf{w}}_{j,k}$.

Lemma 13 *For $1 \leq j \neq k \leq d$, let $\widehat{\mathbf{w}}_{j,k}$ be the solution of the Dantzig-type optimization problem* (11) *and let $\mathbf{w}^*_{j,k} = \mathbf{H}^j_{jk,j\backslash k}(\mathbf{H}^j_{j\backslash k,j\backslash k})^{-1}$. Under Assumptions 2, 4, 6 and 7, with probability tending to one, we have*

$$\|\widehat{\mathbf{w}}_{j,k} - \mathbf{w}^*_{j,k}\|_1 \leq 37\nu_*^{-1}s_0^\star\lambda_D \;\; \textit{for all} \;\; 1 \leq j \neq k \leq d.$$

Proof See §D.2 for a detailed proof. ■

Now combining Lemma 13 and Theorem 10 we obtain that

$$|I_{2j}| \leq 37\nu_*^{-1}K_1 s_0^\star \lambda_D \sqrt{\log d/n} \asymp s_0^\star\lambda\lambda_D = o(n^{-1/2}).$$

Therefore we have shown that $I_{1j} + I_{2j} = o(n^{-1/2})$ with high probability. Similarly, we also have $I_{1k} + I_{2k} = o(n^{-1/2})$ with high probability. Moreover, since the bounds for $|I_{1j}|$ and $|I_{2j}|$ is independent of the choice of $(j,k) \in \{(j,k) \colon 1 \leq j \neq k \leq d\}$, we conclude that

$$\sqrt{n}\big[\widehat{S}_{jk} - S_{jk}\big(\boldsymbol{\beta}^*_{j\vee k}\big)\big] = o_{\mathbb{P}}(1) \quad \text{uniformly for } 1 \leq j < k \leq d.$$

Our next lemma characterizes the limiting distribution of $\nabla L_{jk}\big(\boldsymbol{\beta}^*_{j\vee k}\big)$ and is pivotal for establishing the validity of the composite pairwise score test.

Lemma 14 *For any* $\mathbf{b} \in \mathbb{R}^{2d-3}$ *with* $\|\mathbf{b}\|_2 = 1$ *and* $|\mathbf{b}\|_0 \le \widetilde{s}$, *if* $\lim_{n\to\infty} \widetilde{s}/n = 0$, *we have*

$$\sqrt{n}/2 \cdot \mathbf{b}^T \nabla L_{jk}\big(\boldsymbol{\beta}^*_{j\vee k}\big) \rightsquigarrow N\big(0, \mathbf{b}^T \boldsymbol{\Sigma}^{jk} \mathbf{b}\big). \tag{40}$$

By Lemma 14 we obtain

$$\sqrt{n}/2 \cdot S\big(\boldsymbol{\beta}^*_{j\vee k}\big) = \nabla_{jk} L_{jk}\big(\boldsymbol{\beta}^*_{j\vee k}\big) - \mathbf{w}^{*T}_{j,k} \nabla_{j\backslash k} L_{jk}\big(\boldsymbol{\beta}^*_{j\vee k}\big) - \mathbf{w}^{*T}_{k,j} \nabla_{j\backslash k} L_{jk}\big(\boldsymbol{\beta}^*_{j\vee k}\big) \rightsquigarrow N(0, \sigma^2_{jk}),$$

where the asymptotic variance σ^2_{jk} is given by

$$\sigma^2_{jk} = \boldsymbol{\Sigma}^{jk}_{jk,jk} - 2\boldsymbol{\Sigma}^{jk}_{jk,j\backslash k} \mathbf{w}^*_{j,k} - 2\boldsymbol{\Sigma}_{jk,k\backslash j} \mathbf{w}^*_{k,j} + \mathbf{w}^{*T}_{j,k} \boldsymbol{\Sigma}^{jk}_{j\backslash k,j\backslash k} \mathbf{w}^*_{j,k} + \mathbf{w}^{*T}_{k,j} \boldsymbol{\Sigma}^{jk}_{k\backslash j,k\backslash j} \mathbf{w}^*_{k,j}.$$

For a more accurate estimation of $\widehat{S}_{jk} - S_{jk}\big(\boldsymbol{\beta}^*_{j\vee k}\big)$, we have

$$\sqrt{n}\big|\widehat{S}_{jk} - S_{jk}\big(\boldsymbol{\beta}^*_{j\vee k}\big)\big| \le \sqrt{n}\big(|I_1| + |I_2|\big) \lesssim \sqrt{n}(s^* + s^\star_0)\lambda\lambda_D. \tag{41}$$

Finally, the following lemma, whose proof is deferred to the supplementary material, shows that $\widehat{\sigma}_{jk}$ is a consistent estimator of σ_{jk}.

Lemma 15 *For* $1 \le j \ne k \le d$, *we denote the asymptotic variance of* $\sqrt{n}/2 \cdot S_{jk}(\boldsymbol{\beta}^*_{j\vee k})$ *as* σ^2_{jk}. *Under Assumptions 2, 4, 6 and 7, the estimator* $\widehat{\sigma}_{jk}$ *satisfies* $\lim_{n\to\infty} \max_{j<k} |\widehat{\sigma}_{jk} - \sigma_{jk}| = 0$.

Proof See §D.3 for a proof. ■

Since $\widehat{\sigma}_{jk}$ is consistent for σ_{jk} by Lemma 15 and σ_{jk} is bounded away from zero by Assumption 6, Slutsky's theorem implies that $\sqrt{n}\widehat{S}_{jk}/(2\widehat{\sigma}_{jk}) \rightsquigarrow N(0,1)$. ■

Appendix B. Additional Estimation Results

We present the additional results of parameter estimation. In §B.1 we verify the sparse eigenvalue condition for Gaussian graphical models, which justifies Assumption 4 in our paper. In §B.2 we derive a more refined statistical rates of convergence for the iterates of Algorithm 1.

B.1. Verify the Sparse Eigenvalue Condition for Gaussian Graphical Models

In this subsection, we verify the sparse eigenvalue condition for Gaussian graphical models. Moreover, we show that such condition holds uniformly over a ℓ_1-ball centered at the true parameter $\boldsymbol{\beta}^*_j$.

Proposition 16 *Suppose* $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ *is a Gaussian graphical model and let* $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ *be the precision matrix. For all* $j \in [d]$, *the conditional distribution of* X_j *given* $\boldsymbol{X}_{\backslash j}$ *is a normal distribution with mean* $\boldsymbol{\beta}^{*T}_j \boldsymbol{X}_{\backslash j}$ *and variance* $\boldsymbol{\Theta}^{-1}_{jj}$, *where* $\boldsymbol{\beta}^*_j = \boldsymbol{\Theta}_{j\backslash j}$. *Let* $L_j(\cdot)$ *be the loss function defined in* (6). *We assume that there exist positive constants* D, c_λ *and* C_λ *such that* $\|\boldsymbol{\Sigma}\|_\infty \le D$ *and* $c_\lambda \le \lambda_{\min}(\boldsymbol{\Sigma}) \le \lambda_{\max}(\boldsymbol{\Sigma}) \le C_\lambda$. *We let* $s^* = \max_{j\in[d]} \|\boldsymbol{\beta}^*_j\|_0$ *and also assume that there exists a constant* $C_\beta > 0$ *such that* $\|\boldsymbol{\beta}^*_j\|_2 \le C_\beta$ *for all* $j \in [d]$.

Suppose $r > 0$ is a real number such that $r = \mathcal{O}(1/\sqrt{s^*})$. Then, there exist $\rho_*, \rho^* > 0$ such that for all $j \in [d]$, and $s = 1, \ldots, d-1$,

$$\rho_* \leq \rho_-\big(\mathbb{E}\big[\nabla^2 L_j\big], \boldsymbol{\beta}_j^*; s, r\big) \leq \rho_+\big(\mathbb{E}\big[\nabla^2 L_j\big], \boldsymbol{\beta}_j^*; s, r\big) \leq \rho^*.$$

Proof We prove this lemma in two steps. For any $\boldsymbol{\beta}_j \in \mathbb{R}^{d-1}$ such that $\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^*\|_1 \leq r$ and any $\mathbf{v} \in \mathbb{R}^{d-1}$ such that $\|\mathbf{v}\|_2 = 1$, we first give a lower bound for $\mathbf{v}^T \mathbb{E}\big[\nabla^2 L_j(\boldsymbol{\beta}_j)\big]\mathbf{v}$ by truncation. Then we give an upper bound in the second step.

Step (i): Lower Bound of $\mathbf{v}^T\mathbb{E}\big[\nabla^2 L_j(\boldsymbol{\beta}_j)\big]\mathbf{v}$. We denote $\mathbb{B}_j(r) := \big\{\boldsymbol{\beta} \in \mathbb{R}^{d-1} \colon \|\boldsymbol{\beta} - \boldsymbol{\beta}_j^*\|_1 \leq r\big\}$. For two truncation levels $\tau > 0$ and $R > 0$, we denote $\mathcal{A}_{ii'} := \big\{|X_{ij}| \leq \tau\big\} \cap \big\{|X_{i'j}| \leq \tau\big\}$, $\mathcal{B}_i := \big\{\big|\boldsymbol{X}_{i\backslash j}^T \boldsymbol{\beta}_j\big| \leq R, \forall \boldsymbol{\beta}_j \in \mathbb{B}_j(r)\big\}$ and $\mathcal{B}_{i'} := \big\{\big|\boldsymbol{X}_{i'\backslash j}^T \boldsymbol{\beta}_j^*\big| \leq R, \forall \boldsymbol{\beta}_j \in \mathbb{B}_j(r)\big\}$. The values of R and τ will be determined later. By the definition of $L_j(\cdot)$, for any $\boldsymbol{\beta}_j \in \mathbb{B}_j(r)$ and any $\mathbf{v} \in \mathbb{R}^{d-1}$ with $\|\mathbf{v}\|_2 = 1$, we have

$$\mathbf{v}^T \nabla^2 L_j(\boldsymbol{\beta}_j)\mathbf{v} \geq \frac{2C_1(R,\tau)}{n(n-1)} \sum_{i<i'} \big(X_{ij} - X_{i'j}\big)^2 \big[\big(\boldsymbol{X}_{i\backslash j} - \boldsymbol{X}_{i'\backslash j}\big)^T \mathbf{v}\big]^2 I(\mathcal{B}_i) T(\mathcal{B}_{i'}) I(\mathcal{A}_{ii'}), \quad (42)$$

where $C_1(R,\tau) := \exp(-4R\tau)\big[1 + \exp(-4R\tau)\big]^{-2}$. For notational simplicity, we denote the right-hand side of (42) as $C_1(R,\tau)\mathbf{v}^T \boldsymbol{\Delta} \mathbf{v}$. By the properties of Gaussian graphical models, the conditional density of X_{ij} given $\mathcal{I}_i := \big\{\boldsymbol{X}_{i\backslash j} = \boldsymbol{x}_{i\backslash j}\big\} \cap \mathcal{B}_i$ is

$$p\big(x_{ij}|\mathcal{I}_i\big) = p(\boldsymbol{x}_i|\mathcal{B}_i)\Big/\int_{\mathbb{R}} p(\boldsymbol{x}_i|\mathcal{B}_i)\mathrm{d}x_{ij} = p(x_{ij}|\boldsymbol{x}_{i\backslash j}),$$

where we use the fact that $p(\boldsymbol{x}_i|\mathbb{B}_i) = p(\boldsymbol{x}_i)/\mathbb{P}(\mathcal{B}_i)$ and that $\mathbb{P}(\mathcal{B}_i)$ is a constant. Recall that

$$p(x_{ij}|\boldsymbol{X}_{i\backslash j}) = \sqrt{\boldsymbol{\Theta}_{jj}/(2\pi)}\exp\big[-\boldsymbol{\Theta}_{jj}/2(x_{ij} - \boldsymbol{X}_{i\backslash j}^T\boldsymbol{\beta}_j^*)^2\big] \quad \text{where } \boldsymbol{\beta}_j^* = \boldsymbol{\Theta}_{j\backslash j}.$$

Thus the conditional expectation of $(X_{ij} - X_{i'j})^2 I(\mathcal{A}_{ij})$ given $\mathcal{I}_i$ and $\mathcal{I}_{i'}$ is

$$\begin{aligned}&\mathbb{E}\Big[(X_{ij} - X_{i'j})^2 I(\mathcal{A}_{ii'})\Big|\mathcal{I}_i \cap \mathcal{I}_{i'}\Big] \\ &\quad = \boldsymbol{\Theta}_{jj}/(2\pi)\int_{-\tau}^{\tau}\int_{-\tau}^{\tau}(x_{ij} - x_{ij})^2 \exp\Big\{-\boldsymbol{\Theta}_{jj}/2\big[(x_{ij} - \boldsymbol{\beta}_j^T\boldsymbol{x}_{i\backslash j})^2 + (x_{i'j} - \boldsymbol{\beta}_j^T\boldsymbol{x}_{i'\backslash j})^2\big]\Big\}\mathrm{d}x_{ij}\mathrm{d}x_{i'j}.\end{aligned}$$

Note that on event $\mathcal{I}_i$, $|\boldsymbol{\beta}_j^T\boldsymbol{X}_{i\backslash j}| \leq R$, hence the expression above can be lower-bounded by

$$\begin{aligned}&\mathbb{E}\Big[(X_{ij} - X_{i'j})^2 I(\mathcal{A}_{ii'})\Big|\mathcal{I}_i \cap \mathcal{I}_{i'}\Big] \\ &\quad \geq \boldsymbol{\Theta}_{jj}/(2\pi)\int_{-\tau}^{\tau}\int_{-\tau}^{\tau}(x_{ij} - x_{i'j})^2 \exp\Big\{-\boldsymbol{\Theta}_{jj}/2\big[x_{ij}^2 + x_{i'j}^2 + 2R^2 + 2R(|x_{ij}| + |x_{i'j}|)\big]\Big\}\mathrm{d}x_{ij}\mathrm{d}x_{i'j}.\end{aligned}$$

The last expression is positive and we denote it as $C_2(R,\tau)$ for simplicity. Thus by the law of total expectation we obtain

$$\mathbf{v}^T\mathbb{E}(\boldsymbol{\Delta})\mathbf{v} = \mathbf{v}^T\mathbb{E}\big[\mathbb{E}(\boldsymbol{\Delta}|\cap_{i=1}^n \mathcal{I}_i)\big]\mathbf{v} \geq C_2(R,\tau)\mathbb{E}\Big\{\big[(\boldsymbol{X}_{i\backslash j} - \boldsymbol{X}_{i'\backslash j})^T\mathbf{v}\big]^2 I(\mathcal{B}_i)I(\mathcal{B}_j)\Big\}.$$

By Cauchy-Schwarz inequality we have

$$\mathbb{E}\Big\{\big[(\boldsymbol{X}_{i\setminus j}-\boldsymbol{X}_{i'\setminus j})^T\mathbf{v}\big]^2\big[1-I(\mathcal{B}_i)I(\mathcal{B}_{i'})\big]\Big\}\le\sqrt{\mathbb{E}[(\boldsymbol{X}_{i\setminus j}-\boldsymbol{X}_{i'\setminus j})^T\mathbf{v}\big]^4}\sqrt{\mathbb{P}\big(\mathcal{B}_i^c\cup\mathcal{B}_{i'}^c\big)}. \quad (43)$$

Note that for Gaussian graphical model, the marginal distribution of $\boldsymbol{X}_{\setminus j}$ is $N(\mathbf{0},\boldsymbol{\Sigma}_{\setminus j\setminus j})$. If we denote $\boldsymbol{\Sigma}_{\setminus j\setminus j}$ as $\boldsymbol{\Sigma}_1$, we have $(\boldsymbol{X}_{i\setminus j}-\boldsymbol{X}_{i'\setminus j})^T\mathbf{v}\sim N(\mathbf{0},\sigma_v^2)$, $\boldsymbol{X}_{i\setminus j}^T\boldsymbol{\beta}_j^*\sim N(\mathbf{0},\sigma_1^2)$ and $\boldsymbol{X}_{i\setminus j}^T\boldsymbol{\beta}\sim N(\mathbf{0},\sigma_2^2)$ where $\sigma_v^2=2\mathbf{v}^T\boldsymbol{\Sigma}_1\mathbf{v}$, $\sigma_1^2=\boldsymbol{\beta}_j^{*T}\boldsymbol{\Sigma}_1\boldsymbol{\beta}_j^*$ and $\sigma_2^2=\boldsymbol{\beta}_j^T\boldsymbol{\Sigma}_1\boldsymbol{\beta}_j$. Hence we have $\mathbb{E}\big[(\boldsymbol{X}_{i\setminus j}-\boldsymbol{X}_{i'\setminus j})^T\mathbf{v}\big]^4=3\sigma_v^4$. Because the maximum eigenvalue of $\boldsymbol{\Sigma}_1$ is upper bounded by C_λ, we have $\sigma_1^2\le C_\lambda C_\beta^2$ and $\sigma_v^2\le 2C_\lambda$. Note that $\sigma_2^2-\sigma_1^2=\boldsymbol{\beta}_j^T\boldsymbol{\Sigma}_1\boldsymbol{\beta}_j-\boldsymbol{\beta}_j^{*T}\boldsymbol{\Sigma}_1\boldsymbol{\beta}_j^*$, the following lemma in linear algebra bounds this type of error.

Lemma 17 *Let $\mathbf{M}\in\mathbb{R}^{d\times d}$ be a symmetric matrix and vectors $\mathbf{v}_1$ and $\mathbf{v}_2\in\mathbb{R}^d$, then*

$$\big|\mathbf{v}_1^T\mathbf{M}\mathbf{v}_1-\mathbf{v}_2^T\mathbf{M}\mathbf{v}_2\big|\le\|\mathbf{M}\|_\infty\|\mathbf{v}_1-\mathbf{v}_2\|_1^2+2\|\mathbf{M}\mathbf{v}_2\|_\infty\|\mathbf{v}_1-\mathbf{v}_2\|_1.$$

Proof Note that $\mathbf{v}_1^T\mathbf{M}\mathbf{v}_1-\mathbf{v}_2^T\mathbf{M}\mathbf{v}_2=(\mathbf{v}_1-\mathbf{v}_2)^T\mathbf{M}(\mathbf{v}_1-\mathbf{v}_2)+2\mathbf{v}_2^T\mathbf{M}(\mathbf{v}_1-\mathbf{v}_2)$, Hölder's inequality implies

$$\begin{aligned}\big|\mathbf{v}_1^T\mathbf{M}\mathbf{v}_1-\mathbf{v}_2^T\mathbf{M}\mathbf{v}_2\big|&\le\big|(\mathbf{v}_1-\mathbf{v}_2)^T\mathbf{M}(\mathbf{v}_1-\mathbf{v}_2)\big|+2\big|\mathbf{v}_2^T\mathbf{M}(\mathbf{v}_1-\mathbf{v}_2)\big|\\&\le\|\mathbf{M}\|_\infty\|\mathbf{v}_1-\mathbf{v}_2\|_1^2+2\|\mathbf{M}\mathbf{v}_2\|_\infty\|\mathbf{v}_1-\mathbf{v}_2\|_1.\end{aligned}$$

Hence, we conclude the proof of Lemma 17. ■

By Lemma 17, we have

$$\sigma_2^2-\sigma_1^2\le\|\boldsymbol{\Sigma}_1\|_\infty\|\boldsymbol{\beta}_j-\boldsymbol{\beta}_j^*\big\|_1^2+2\|\boldsymbol{\Sigma}_1\boldsymbol{\beta}_j^*\|_\infty\|\boldsymbol{\beta}_j-\boldsymbol{\beta}_j^*\|_1. \quad (44)$$

By Hölder's inequality and the relation between ℓ_1-norm and ℓ_2-norm of a vector, we have $\|\boldsymbol{\Sigma}_1\boldsymbol{\beta}_j^*\|_\infty\le\|\boldsymbol{\Sigma}_1\|_\infty\|\boldsymbol{\beta}_j^*\|_1\le\sqrt{s^*}C_\beta D$. Therefore the right-hand side of (44) can be bounded by

$$\sigma_2^2-\sigma_1^2\le r^2D+2\sqrt{s^*}rC_\beta D,$$

which shows that σ_2^2 is also bounded because $r=\mathcal{O}(1/\sqrt{s^*})$. In addition, by the bound $1-\Phi(x)\le\exp(-x^2/2)/(x\sqrt{2\pi})$ for the standard normal distribution function, we obtain that

$$\begin{aligned}\mathbb{P}\big(\mathcal{B}_i^c\big)&\le\mathbb{P}\big(\boldsymbol{X}_{i\setminus j}^T\boldsymbol{\beta}_j^*>R\big)+\mathbb{P}\big(\boldsymbol{X}_{i\setminus j}^T\boldsymbol{\beta}_j>R\big)\\&\le c\sigma_1\exp\big[-R^2/(2\sigma_1^2)\big]/R+c\sigma_2\exp\big[-R^2/(2\sigma_2^2)\big]/R,\end{aligned}$$

where the constant $c=1/\sqrt{2\pi}$. We denote the last expression as $C_3(R)$, then the right-hand side of (43) can be upper-bounded by $\sqrt{3\sigma_v^4}\sqrt{2C_3(R)}\le 2\sqrt{6C_3(R)}C_\lambda$. Hence we can choose a sufficiently large R such that $2\sqrt{6C_3(R)}C_\lambda=\lambda_{\min}(\boldsymbol{\Sigma})$ and we denote this particular choice of R as R_0.

Now we have

$$\mathbb{E}\Big\{\big[(\boldsymbol{X}_{i\setminus j}-\boldsymbol{X}_{i'\setminus j})^T\mathbf{v}\big]^2\big[1-I(\mathcal{B}_i)I(\mathcal{B}_{i'})\big]\Big\}\le\lambda_{\min}(\boldsymbol{\Sigma})$$

Note that $\mathbb{E}\big\{[(\boldsymbol{X}_{i\backslash j}-\boldsymbol{X}_{i'\backslash j})^T\mathbf{v}]^2\big\}=\sigma_v^2\geq 2\lambda_{\min}(\boldsymbol{\Sigma})$, we obtain that

$$\mathbf{v}^T\mathbb{E}\big[\nabla^2 L_j(\boldsymbol{\beta}_j)\big]\mathbf{v}\geq C_1(R_0,\tau)C_2(R_0,\tau)\lambda_{\min}(\boldsymbol{\Sigma})\ \text{ for all }\ \tau\in\mathbb{R}.$$

Therefore we conclude that for all $\boldsymbol{\beta}_j\in\mathbb{R}^{d-1}$ such that $\|\boldsymbol{\beta}_j-\boldsymbol{\beta}_j^*\|_1\leq r$,

$$\mathbf{v}^T\mathbb{E}\big[\nabla^2 L_j(\boldsymbol{\beta}_j)\big]\mathbf{v}\geq \max_{\tau\in\mathbb{R}}\big\{C_1(R_0,\tau)C_2(R_0,\tau)\big\}\lambda_{\min}(\boldsymbol{\Sigma}). \tag{45}$$

Step (ii): Upper Bound of $\mathbf{v}^T\mathbb{E}\big[\nabla^2 L_j(\boldsymbol{\beta}_j)\big]\mathbf{v}$. For any $\boldsymbol{\beta}_j\in\mathbb{R}^{d-1}$ such that $\|\boldsymbol{\beta}_j-\boldsymbol{\beta}_j^*\|_1\leq r$ and for any $\mathbf{v}\in\mathbb{R}^{d-1}$ with $\|\mathbf{v}\|_2=1$, by the definition of $\nabla^2 L_j(\boldsymbol{\beta}_j)$ we have

$$\mathbf{v}^T\nabla^2 L_j(\boldsymbol{\beta}_j)\mathbf{v}\leq (X_{ij}-X_{i'j})^2\big[(\boldsymbol{X}_{i\backslash j}-\boldsymbol{X}_{i'\backslash j})^T\mathbf{v}\big]^2. \tag{46}$$

Notice that conditioning on $\boldsymbol{X}_{i\backslash j}$, $X_{ij}\sim N\big(\boldsymbol{X}_{i\backslash j}^T\boldsymbol{\beta}_j^*,\boldsymbol{\Theta}_{jj}^{-1}\big)$, hence

$$\mathbb{E}\big[(X_{ij}-X_{i'j})^2\big|\boldsymbol{X}_{i\backslash j},\boldsymbol{X}_{i'\backslash j}\big]=\big[(\boldsymbol{X}_{i\backslash j}-\boldsymbol{X}_{i'\backslash j})^T\boldsymbol{\beta}_j^*\big]^2+2\boldsymbol{\Theta}_{jj}^{-1}. \tag{47}$$

Combining (46) and (47) we obtain

$$\begin{aligned}\mathbb{E}\big[\mathbf{v}^T\nabla^2 L_j(\boldsymbol{\beta}_j)\mathbf{v}\big]&\leq\mathbb{E}\Big\{\mathbb{E}\big[(X_{ij}-X_{i'j})^2\big|\boldsymbol{X}_{i\backslash j},\boldsymbol{X}_{i'\backslash j}\big]\cdot\big[(\boldsymbol{X}_{i\backslash j}-\boldsymbol{X}_{i'\backslash j})^T\mathbf{v}\big]^2\Big\}\\&\leq 2\boldsymbol{\Theta}_{jj}^{-1}\mathbb{E}\big((\boldsymbol{X}_{i\backslash j}-\boldsymbol{X}_{i'\backslash j})^T\mathbf{v}\big)^2+\mathbb{E}\Big\{\big[(\boldsymbol{X}_{i\backslash j}-\boldsymbol{X}_{i'\backslash j})^T\boldsymbol{\beta}_j^*\big]^2\big[(\boldsymbol{X}_{i\backslash j}-\boldsymbol{X}_{i'\backslash j})^T\mathbf{v}\big]^2\Big\}.\end{aligned} \tag{48}$$

Because $\boldsymbol{X}_{i\backslash j}\sim N(\mathbf{0},\boldsymbol{\Sigma}_1)$ where $\boldsymbol{\Sigma}_1:=\boldsymbol{\Sigma}_{\backslash j,\backslash j}$, and also note that the maximum eigenvalue of $\boldsymbol{\Sigma}_1$ is upper bounded by C_λ, we have

$$\mathbb{E}\big[(\boldsymbol{X}_{i\backslash j}-\boldsymbol{X}_{i'\backslash j})^T\mathbf{v}\big]^2=2\mathbf{v}^T\boldsymbol{\Sigma}_1\mathbf{v}\leq 2C_\lambda.$$

Moreover, by inequality $2ab\leq a^2+b^2$ we obtain

$$2\mathbb{E}\Big\{\big[(\boldsymbol{X}_{i\backslash j}-\boldsymbol{X}_{i'\backslash j})^T\boldsymbol{\beta}_j^*\big]^2\big[(\boldsymbol{X}_{i\backslash j}-\boldsymbol{X}_{i'\backslash j})^T\mathbf{v}\big]^2\Big\}\leq\mathbb{E}\big[(\boldsymbol{X}_{i\backslash j}-\boldsymbol{X}_{i'\backslash j})^T\boldsymbol{\beta}_j^*\big]^4+\mathbb{E}\big[(\boldsymbol{X}_{i\backslash j}-\boldsymbol{X}_{i'\backslash j})^T\mathbf{v}\big]^4.$$

Since $(\boldsymbol{X}_{i\backslash j}-\boldsymbol{X}_{i'\backslash j})^T\mathbf{v}\sim N(0,\sigma_v^2)$ and $(\boldsymbol{X}_{i\backslash j}-\boldsymbol{X}_{i'\backslash j})^T\boldsymbol{\beta}_j^*\sim N(0,2\sigma_1^2)$ where σ_v^2 and σ_1^2 are defined as $2\mathbf{v}^T\boldsymbol{\Sigma}_1\mathbf{v}$ and $\boldsymbol{\beta}_j^{*T}\boldsymbol{\Sigma}_1\boldsymbol{\beta}_j^*$ respectively, we obtain

$$\mathbb{E}\big[(\boldsymbol{X}_{i\backslash j}-\boldsymbol{X}_{i'\backslash j})^T\boldsymbol{\beta}_j^*\big]^4=3\sigma_v^4\leq 12C_\lambda^2\quad\text{and}\quad\mathbb{E}\big[(\boldsymbol{X}_{i\backslash j}-\boldsymbol{X}_{i'\backslash j})^T\mathbf{v}\big]^4=12\sigma_1^4\leq 12C_\lambda C_\beta^2.$$

Therefore we can bound the right-hand side of (48) by

$$\mathbb{E}\big[\mathbf{v}^T\nabla^2 L_j(\boldsymbol{\beta}_j)\mathbf{v}\big]\leq 4\boldsymbol{\Theta}_{jj}^{-1}C_\lambda+6C_\lambda^2+6C_\lambda C_\beta^2. \tag{49}$$

Combining (45) and (49) we conclude that Proposition 16 holds with

$$\rho_*=\max_{\tau\in\mathbb{R}}\big\{C_1(R_0,\tau)C_2(R_0,\tau)\big\}\lambda_{\min}(\boldsymbol{\Sigma})\ \text{ and }\ \rho^*=4\boldsymbol{\Theta}_{jj}^{-1}C_\lambda+6C_\lambda^2+6C_\lambda C_\beta^2.$$

Therefore, we conclude the proof of Proposition 16. ■

B.2. Refined Statistical Rates of Parameter Estimation

In this subsection we show more refined statistical rates of convergence for the proposed estimators. In specific, we consider the case where $\boldsymbol{\beta}_j^*$ contains nonzero elements with both strong and week magnitudes.

Theorem 18 (Refined statistical rates of convergence) *Under Assumptions 2 and 4, we let K_1 and K_2 be the constants defined in Theorem 10 and also let $\rho_* > 0$ and $r > 0$ be defined in Assumption 4. For all $j \in [d]$, we define the support of $\boldsymbol{\beta}_j^*$ as $S_j := \{(j,k)\colon \beta_{jk}^* \neq 0, k \in [d]\}$ and let $s^* = \max_{j\in[d]} \|\boldsymbol{\beta}_j^*\|_0$. The penalty function $p_\lambda(u) : [0,+\infty) \to [0,+\infty)$ in (7) satisfies regularity conditions (C.1), (C.2) and (C.3) listed in §3.2 with $c_1 = 0.91$ and $c_2 \geq 24/\rho_*$ for condition (C.3). We set the regularity parameter $\lambda = C\sqrt{\log d/n}$ such that $C \geq 25K_1$. Moreover, we assume that the penalty function $p_\lambda(u)$ satisfies an extra condition (C.4): there exists a constant $c_3 > 0$ such that $p_\lambda'(u) = 0$ for $u \in \big[c_3\lambda, +\infty\big)$. Suppose that the support of $\boldsymbol{\beta}_j^*$ can be partitioned into $S_j = S_{1j} \cup S_{2j}$ where $S_{1j} = \big\{(j,k)\colon |\beta_{jk}^*| \geq (c_2+c_3)\lambda\big\}$ and $S_{2j} = S_j - S_{1j}$. We denote constants $A_1 = 22\varrho$, $A_2 = 2.2c_2$, $B_1 = 32\varrho$, $B_2 = 3.2c_2$, $\varrho = c_2(c_2\rho_* - 11)^{-1}$, $\gamma = 11c_2^{-1}\rho_*^{-1} < 1$ and $a = 1.04$; we let $s_{1j}^* = |S_{1j}|$ and $s_{2j}^* = |S_{2j}|$. With probability at least $1 - d^{-1}$, we have the following more refined rates of convergence:*

$$\big\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}_j^*\big\|_2 \leq A_1\Big\{\big\|\nabla_{S_{1j}} L_j(\boldsymbol{\beta}_j^*)\big\|_2 + a\sqrt{s_{2j}^*}\lambda\Big\} + A_2\sqrt{s^*}\lambda\gamma^\ell \quad \text{and} \tag{50}$$

$$\big\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}_j^*\big\|_1 \leq B_1\Big\{\big\|\nabla_{S_{1j}} L_j(\boldsymbol{\beta}_j^*)\big\|_2 + a\sqrt{s_{2j}^*}\lambda\Big\} + B_2 s^*\lambda\gamma^\ell, \forall j \in [d]. \tag{51}$$

Proof Let $S_j = \{(j,k)\colon \beta_{jk}^* \neq 0, k \in [d]\}$ be the support of $\boldsymbol{\beta}_j^*$ and let index set G_j^ℓ, J_j^ℓ and I_j^ℓ be the same as defined in the proof of Theorem 5. For notational simplicity, we omit the subscript j in these index sets which stands for the j-th node of the graph; we simply write them as G^ℓ, J^ℓ and I^ℓ. Moreover, we let $\boldsymbol{\delta}^{(\ell)} = \widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}_j^*$, it is shown in Lemma 12 that

$$\big\|\boldsymbol{\delta}_{I^\ell}^{(\ell)}\big\|_2 \leq 10\rho_*^{-1}\Big(\big\|\nabla_{\widetilde{G}^\ell} L_j(\boldsymbol{\beta}_j^*)\big\|_2 + \big\|\boldsymbol{\lambda}_{S_j}^{(\ell-1)}\big\|_2\Big); \quad \widetilde{G}^\ell = (G^\ell)^c. \tag{52}$$

In the proof of Theorem 5, we show that $|\widetilde{G}^\ell| \leq 2s^*$ for all $j \in [d]$ and $\ell \geq 1$. Because $S_j = S_{1j} \cup S_{2j}$ where $S_{1j} = \big\{(j,k)\colon |\beta_{jk}^*| \geq (c_2 + c_3)\lambda\big\}$ and $S_{2j} = S_j - S_{1j}$, then by triangle inequality we have

$$\big\|\nabla_{S_j} L_j(\boldsymbol{\beta}_j^*)\big\|_2 \leq \big\|\nabla_{S_{1j}} L_j(\boldsymbol{\beta}_j^*)\big\|_2 + \sqrt{s_{2j}^*}\big\|\nabla_{S_{2j}} L_j(\boldsymbol{\beta}_j^*)\big\|_\infty.$$

Since $\lambda \geq 25\big\|\nabla L_j(\boldsymbol{\beta}_j^*)\big\|_\infty$ with high probability, by (28), we further have

$$\big\|\nabla_{\widetilde{G}^\ell} L_j(\boldsymbol{\beta}_j^*)\big\|_2 \leq \big\|\nabla_{S_{1j}} L_j(\boldsymbol{\beta}_j^*)\big\|_2 + \sqrt{s_{2j}^*}\lambda/25 + \big\|\boldsymbol{\delta}_{I^{\ell-1}}^{(\ell-1)}\big\|_2\Big/(25c_2). \tag{53}$$

Note that by the definition of S_{1j}, for any $(j,k) \in S_{1j}$, $p_\lambda'\big(|\beta_{jk}| - c_2\lambda\big) \leq p_\lambda'(c_3\lambda) = 0$, then we have

$$\Upsilon_j := \lambda\Big[\sum_{(j,k)\in S_j} p_\lambda'\big(|\beta_{jk}^*| - c_2\lambda\big)^2\Big]^{1/2} = \lambda\Big[\sum_{(j,k)\in S_{2j}} p_\lambda'(|\beta_{jk}^*| - c_2\lambda)^2\Big]^{1/2} \leq \sqrt{s_{2j}^*}\lambda.$$

Therefore (30) is reduced to

$$\big\|\boldsymbol{\lambda}_{S_j}^{(\ell-1)}\big\|_2 \le \Upsilon_j + \big\|\boldsymbol{\delta}_{I^{\ell-1}}^{(\ell-1)}\big\|_2/c_2 \le \sqrt{s_{2j}^*}\lambda + \big\|\boldsymbol{\delta}_{I^{\ell-1}}^{(\ell-1)}\big\|_2\Big/c_2. \tag{54}$$

Combining (52), (53) and (54) we obtain

$$\big\|\boldsymbol{\delta}_{I^\ell}^{(\ell)}\big\|_2 \le 10\rho_*^{-1}\Big\{\big\|\nabla_{S_{1j}}L_j(\boldsymbol{\beta}_j^*)\big\|_2 + 1.04\sqrt{s_{2j}^*}\lambda + 1.04\big\|\boldsymbol{\delta}_{I^{\ell-1}}^{(\ell-1)}\big\|_2/c_2\Big\}.$$

Then by recursion, we obtain the following estimation error:

$$\big\|\boldsymbol{\delta}_{I^\ell}^{(\ell)}\big\|_2 \le 10\varrho\Big\{\big\|\nabla_{S_{1j}}L_j(\boldsymbol{\beta}_j^*)\big\|_2 + 1.04\sqrt{s_{2j}^*}\lambda\Big\} + \gamma^{\ell-1}\big\|\boldsymbol{\delta}_{I^1}^{(1)}\big\|_2,$$

where $\gamma := 11c_2^{-1}\rho_*^{-1}$ and $\varrho := c_2(c_2\rho_* - 11)^{-1}$. Note that we assume $c_2 \ge 24\rho_*^{-1}$, for $k = 1$ by (32) we have

$$2.2\big\|\boldsymbol{\delta}_{I_1}^{(1)}\big\|_2 \le 2.2c_2\gamma\sqrt{s^*}\lambda \ \text{ and } \ 2.2\sqrt{2s^*}\big\|\boldsymbol{\delta}_{I_1}^{(1)}\big\|_2 \le 3.2c_2\gamma s^*\lambda.$$

Therefore, using the original notation, we obtain the refined rates of convergence by (23):

$$\begin{aligned}
&\big\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}_j^*\big\|_2 \le 22\varrho\Big\{\big\|\nabla_{S_{1j}}L_j(\boldsymbol{\beta}_j^*)\big\|_2 + 1.04\sqrt{s_{2j}^*}\lambda\Big\} + 2.2c_2\gamma^\ell\sqrt{s^*}\lambda \ \text{ and}\\
&\big\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}_j^*\big\|_1 \le 32\varrho\sqrt{s^*}\Big\{\big\|\nabla_{S_{1j}}L_j(\boldsymbol{\beta}_j^*)\big\|_2 + 1.04\sqrt{s_{2j}^*}\lambda\Big\} + 3.2c_2\gamma^\ell s^*\lambda,
\end{aligned}$$

where $s_{2j}^* = |S_{2j}|$. Moreover, it is easy to see that, with probability at least $1-d^{-1}$, these convergence rates hold for all $j \in [d]$. ■

Appendix C. Proof of the Auxiliary Results for Estimation

In this appendix, we prove the main results for estimation results presented in §4.1. In this appendix, we prove the auxiliary results for estimation. In specific, we give detailed proofs of Lemmas 10, 11, and 12, which are pivotal for the proof of Theorem5. We first prove Lemmas 10, which gives an upper bound for $\|\nabla L_j(\boldsymbol{\beta}_j^*)\|_\infty$.

C.1. Proof of Lemma 10

Proof By definition, $\nabla L_j(\boldsymbol{\beta}_j^*)$ is a centered second-order U-statistic with kernel function $\mathbf{h}_{ii'}^j(\boldsymbol{\beta}_j^*)\in\mathbb{R}^{d-1}$, whose entries are given by

$$\big[\mathbf{h}_{ii'}^j(\boldsymbol{\beta}_j^*)\big]_{jk} = \frac{R_{ii'}^j(\boldsymbol{\beta}_j^*)(X_{ij}-X_{i'j})(X_{ik}-X_{i'k})}{1+R_{ii'}^j(\boldsymbol{\beta}_j^*)}.$$

By the tail probability bound in (14), for any $i \in [n]$ and $j \in [d]$, we have

$$\begin{aligned}
\mathbb{P}\big(|X_{ij}| > x, \forall i \in [n], \forall j \in [d]\big) &\le \sum_{i\in[n],j\in[d]} \mathbb{P}(|X_{ij}| > x)\\
&\le 2\exp(\kappa_m + \kappa_h/2)\exp(-x + \log d + \log n).
\end{aligned} \tag{55}$$

By setting $x = C\log d$ for some constant C, we conclude that event $\mathcal{E} := \{|X_{ij}| \leq C\log d, \forall i \in [n], \forall j \in [d]\}$ holds with probability at least $1-(4d)^{-1}$. Following from the same argument as in Ning et al. (2017b), it is easy to show that, conditioning on $\mathcal{E}$, $\mathbf{h}^j_{ii'}(\boldsymbol{\beta}^*_j)$ is also centered. Note that conditioning on event $\mathcal{E}$, $\big\|\mathbf{h}^j_{ii'}(\boldsymbol{\beta}^*_j)\big\|_\infty \leq C\log^2 d$ for some generic constant C and for all $i, i' \in [d]$ and $j \in [d]$. The following Bernstein's inequality for U-statistics, presented in Arcones (1995), gives an upper bound for the tail probability of $\nabla L_j(\boldsymbol{\beta}^*_j)$.

Lemma 19 (Bernstein's inequality for U-statistics) *Given n i.i.d. random variables $Z_1, \ldots Z_n$ taking values in a measurable space $(\mathbb{S}, \mathcal{B})$ and a symmetric and measurable kernel function $h\colon \mathbb{S}^m \to R$, we define the U-statistics with kernel h as*

$$U := \binom{n}{m}^{-1} \sum_{i_1<\ldots<i_m} h(Z_{i_1}, \ldots, Z_{i_m}).$$

Suppose that $\mathbb{E}[h(Z_{i_1}, \ldots, Z_{i_m})] = 0$, $\mathbb{E}\big\{\mathbb{E}[h(Z_{i_1}, \ldots, Z_{i_m}) \mid Z_{i_1}]\big\}^2 = \sigma^2$, and $\|h\|_\infty \leq b$ for some positive σ and b. There exists an absolute constant $K(m) > 0$ that only depends on m such that

$$\mathbb{P}(|U| > t) \leq 4\exp\big\{-nt^2/[2m^2\sigma^2 + K(m)bt]\big\},\ \forall t > 0. \tag{56}$$

Note that by (14), the fourth moment of $\boldsymbol{X}$ is bounded, which implies that $\mathbb{E}[\mathbf{h}^j_{ii'}(\boldsymbol{\beta}^*_j)]^2$ is uniformly bounded by an absolute constant for all $j \in [d]$. By Lemma 19, setting $b = C\log^2 d$ in (56) yields that

$$\mathbb{P}\big(\big|\nabla_{jk} L_j(\boldsymbol{\beta}^*_j)\big| > t\big|\mathcal{E}\big) \leq 4\exp\Big[-nt^2/(C_1 + C_2\log^2 d \cdot t)\Big] \tag{57}$$

for some generic constants C_1 and C_2. Taking a union bound over $\{(j,k)\colon j,k \in [d], k \neq j\}$ we obtain

$$\max_{j\in[d]}\Big\{\mathbb{P}\big(\big\|\nabla L_j(\boldsymbol{\beta}^*_j)\big\|_\infty > t\big|\mathcal{E}\big)\Big\} \lesssim d^2 \cdot \exp\big[-nt^2/(C_1 + C_2\log^2 d\cdot t)\big]. \tag{58}$$

Under Assumption 4 and conditioning on $\mathcal{E}$, by setting $t = K_1\sqrt{\log d/n}$ for a sufficiently large $K_1 > 0$, it holds probability greater than $1-(4d)^{-1}$ that

$$\big\|\nabla L_j(\boldsymbol{\beta}^*_j)\big\|_\infty \leq K_1\sqrt{\log d/n}\ \ \forall j \in [d].$$

Note that $\mathcal{E}$ holds with probability at least $1-(4d)^{-1}$, we conclude the proof of Lemma 10. ■

C.2. Proof of Lemma 11

Proof In what follows, for notational simplicity and readability, we omit j in the subscript and ℓ in the superscript by simply writing S_j, G^ℓ_j, J^ℓ_j and I^ℓ_j as S, G, J an I respectively. By the definition of G, $\big\|\boldsymbol{\lambda}^{(\ell-1)}_{G^\ell}\big\|_{\min} \geq p'_\lambda(\theta) \geq 0.91\lambda > 22.75\|\nabla L_j(\boldsymbol{\beta}^*_j)\|_\infty$. We prove this lemma in two steps. In the **first step** we show that $\big\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}_j^*\big\|_1 \leq 2.2\big\|\widehat{\boldsymbol{\beta}}_{G^c}^{\ell} - \boldsymbol{\beta}_{G^c}^*\big\|_1$. Suppose that $\widehat{\boldsymbol{\beta}}_j^{(\ell)}$ is the solution in the ℓ-th iteration and we denote $\nabla_{jk}L_j(\boldsymbol{\beta}_j) = \partial L_j(\boldsymbol{\beta}_j)/\partial\beta_{jk}$, the Karush-Kuhn-Tucker condition implies that

$$\begin{aligned}
&\nabla_{jk}L_j(\widehat{\boldsymbol{\beta}}_j^{(\ell)}) + \lambda_{jk}^{(\ell-1)}\mathrm{sign}(\widehat{\beta}_{jk}^{(\ell)}) = 0 \quad if \quad \widehat{\beta}_{jk}^{(\ell)} \neq 0;\\
&\nabla_{jk}L_j(\widehat{\boldsymbol{\beta}}_j^{(\ell)}) + \lambda_{jk}^{(\ell-1)}\xi_{jk}^{(\ell)} = 0, \quad \xi_{jk}^{(\ell)} \in [-1,1] \quad if \quad \widehat{\beta}_{jk}^{(\ell)} = 0.
\end{aligned}$$

The above Karush-Kuhn-Tuker condition can be written in a compact form as

$$\nabla L_j(\widehat{\boldsymbol{\beta}}_j^{(\ell)}) + \boldsymbol{\lambda}_j^{(\ell-1)} \circ \boldsymbol{\xi}_j^{(\ell)} = 0, \tag{59}$$

where $\boldsymbol{\xi}_j^{(\ell)} \in \partial\big\|\widehat{\boldsymbol{\beta}}_j^{(\ell)}\big\|_1$ and $\boldsymbol{\lambda}_j^{(\ell-1)} = \big(\lambda_{j1}^{(\ell-1)}, \ldots, \lambda_{jj-1}^{(\ell-1)}, \lambda_{jj+1}^{(\ell-1)}, \ldots, \lambda_{jd}^{(\ell-1)}\big)^T \in \mathbb{R}^{d-1}$.

For notational simplicity, we let $\boldsymbol{\delta} = \widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}_j^* \in \mathbb{R}^{d-1}$ and omit the superscript ℓ and subscript j in both $\boldsymbol{\lambda}_j^{(\ell-1)}$ and $\boldsymbol{\xi}_j^{(\ell)}$ by writing them as $\boldsymbol{\lambda}$ and $\boldsymbol{\xi}$. By definition, $I = G^c \cup J$. Note that we denote the support of $\boldsymbol{\beta}_j^*$ as S; we define $H := G^c - S$, then S, H and G is a partition of $\big\{(j,k)\colon k \in [d], k \neq j\big\}$.

By the Mean-Value theorem, there exists an $\alpha \in [0,1]$ such that $\widetilde{\boldsymbol{\beta}}_j := \alpha\boldsymbol{\beta}_j^* + (1-\alpha)\widehat{\boldsymbol{\beta}}_j^{(\ell)} \in \mathbb{R}^{d-1}$ satisfies

$$\nabla L_j(\widehat{\boldsymbol{\beta}}_j) - \nabla L_j(\boldsymbol{\beta}_j^*) = \nabla^2 L_j(\widetilde{\boldsymbol{\beta}}_j)\boldsymbol{\delta}.$$

Then (59) implies that

$$0 \leq \boldsymbol{\delta}^T\nabla^2 L_j(\widetilde{\boldsymbol{\beta}}_j)\boldsymbol{\delta} = -\underbrace{\big\langle \boldsymbol{\delta}, \boldsymbol{\lambda} \circ \boldsymbol{\xi} \big\rangle}_{\text{(i)}} - \underbrace{\big\langle \nabla L_j(\boldsymbol{\beta}_j^*), \boldsymbol{\delta} \big\rangle}_{\text{(ii)}}. \tag{60}$$

For term (ii) in (60), Hölder's inequality implies that

$$\text{(ii)} \geq -\big\|\nabla L_j(\boldsymbol{\beta}_j^*)\big\|_\infty \|\boldsymbol{\delta}\|_1. \tag{61}$$

For term (i) in (60), recall that we denote $|\mathbf{v}|$ as the vector that takes entrywise absolute value for $\mathbf{v}$. By the fact that $\xi_{jk}^{(\ell)}\widehat{\beta}_{jk}^{(\ell)} = |\widehat{\beta}_{jk}^{(\ell)}|$, we have $\boldsymbol{\xi}_G \circ \boldsymbol{\delta}_G = |\boldsymbol{\delta}_G|$ and $\boldsymbol{\xi}_H \circ \boldsymbol{\delta}_H = |\boldsymbol{\delta}_H|$. Since $\boldsymbol{\delta}_{S^c} = \widehat{\boldsymbol{\beta}}_{S^c}^{(\ell)}$. Hölder's inequality implies that

$$\begin{aligned}
\big\langle \boldsymbol{\delta}, \boldsymbol{\lambda} \circ \boldsymbol{\xi} \big\rangle &= \big\langle \boldsymbol{\delta}_S, (\boldsymbol{\lambda} \circ \boldsymbol{\xi})_S \big\rangle + \big\langle |\boldsymbol{\delta}_H|, \boldsymbol{\lambda}_H \big\rangle + \big\langle |\boldsymbol{\delta}_G|, \boldsymbol{\lambda}_G \big\rangle\\
&\geq -\|\boldsymbol{\delta}_S\|_1\|\boldsymbol{\lambda}_S\|_\infty + \|\boldsymbol{\delta}_G\|_1\|\boldsymbol{\lambda}_G\|_{\min} + \|\boldsymbol{\delta}_H\|_1\|\boldsymbol{\lambda}_H\|_{\min}.
\end{aligned} \tag{62}$$

Combining (60), (61) and (62) we have

$$-\|\boldsymbol{\delta}_S\|_1\|\boldsymbol{\lambda}_S\|_\infty + \|\boldsymbol{\delta}_G\|_1\|\boldsymbol{\lambda}_G\|_{\min} + \|\boldsymbol{\delta}_H\|_1\|\boldsymbol{\lambda}_H\|_{\min} - \big\|\nabla L_j(\boldsymbol{\beta}_j^*)\big\|_\infty\|\boldsymbol{\delta}\|_1 \leq 0. \tag{63}$$

By the definition of G, we have $\|\boldsymbol{\lambda}_G\|_{\min} \geq p_\lambda'(c_2\lambda) \geq 0.91\lambda$. Rearranging terms in (63) we have

$$p_\lambda'(c_2\lambda)\|\boldsymbol{\delta}_G\|_1 \leq \|\boldsymbol{\delta}_G\|_1\|\boldsymbol{\lambda}_G\|_{\min} \leq \big\|\nabla L_j(\boldsymbol{\beta}_j^*)\big\|_\infty\|\boldsymbol{\delta}\|_1 + \|\boldsymbol{\delta}_S\|_1\|\boldsymbol{\lambda}_S\|_\infty.$$

Using the decomposability of the ℓ_1-norm, we have

$$\Big[p'_\lambda(c_2\lambda) - \big\|\nabla L_j(\boldsymbol{\beta}^*_j)\big\|_\infty\Big]\|\boldsymbol{\delta}_G\|_1 \le \Big[\|\boldsymbol{\lambda}_S\|_\infty + \|\nabla L_j(\boldsymbol{\beta}^*_j)\|_\infty\Big]\|\boldsymbol{\delta}_{G^c}\|_1 \tag{64}$$

Recall that $\lambda > 25\big\|\nabla L_j(\boldsymbol{\beta}^*_j)\big\|_\infty$ and $p'_\lambda(\theta) \ge 0.91\lambda$, (64) implies

$$\big\|\boldsymbol{\delta}_G\big\|_1 \le \frac{\lambda + \big\|\nabla L_j(\boldsymbol{\beta}^*_j)\big\|_\infty}{p'_\lambda(c_2\lambda) - \big\|\nabla L_j(\boldsymbol{\beta}^*_j)\big\|_\infty}\|\boldsymbol{\delta}_{G^c}\|_1 \le 1.2\|\boldsymbol{\delta}_{G^c}\|_1, \tag{65}$$

where we use the fact that

$$\frac{\lambda + \big\|\nabla L_j(\boldsymbol{\beta}^*_j)\big\|_\infty}{p'_\lambda(c_2\lambda) - \big\|\nabla L_j(\boldsymbol{\beta}^*_j)\big\|_\infty} \le \frac{\lambda + 0.04\lambda}{0.91\lambda - 0.04\lambda} \le 1.2.$$

Going back to the original notation, (65) is equivalent to

$$\big\|\widehat{\boldsymbol{\beta}}^{(\ell)}_j - \boldsymbol{\beta}^*_j\big\|_1 \le 2.2\big\|\widehat{\boldsymbol{\beta}}^{(\ell)}_{\widetilde{G}^\ell_j} - \boldsymbol{\beta}^*_{\widetilde{G}^\ell_j}\big\|_1.$$

Now we show in the **second step** that $\big\|\widehat{\boldsymbol{\beta}}^{(\ell)}_j - \boldsymbol{\beta}^*_j\big\|_2 \le 2.2\big\|\widehat{\boldsymbol{\beta}}^{(\ell)}_{I^\ell_j} - \boldsymbol{\beta}^*_{I^\ell_j}\big\|_2$. Recall that J is the largest k^* components of $\widehat{\boldsymbol{\beta}}^{(\ell)}_G$ in absolute value where we omit the subscript j and superscript ℓ in the sets G^ℓ_j, J^ℓ_j and I^ℓ_j. By the definition of J we obtain that

$$\|\boldsymbol{\delta}_{I^c}\|_\infty \le \|\boldsymbol{\delta}_J\|_1/k^* \le \|\boldsymbol{\delta}_G\|_1/k^*, \quad \text{where} \quad \boldsymbol{\delta} = \widehat{\boldsymbol{\beta}}^{(\ell)}_j - \boldsymbol{\beta}^*_j.$$

By inequality (65) and the fact that $G^c \subset I$, we further have

$$\|\boldsymbol{\delta}_{I^c}\|_\infty \le 1.2/k^* \cdot \|\boldsymbol{\delta}_{G^c}\|_1 \le 1.2/k^* \cdot \|\boldsymbol{\delta}_I\|_1. \tag{66}$$

Then by Hölder' inequality and (66) we obtain that

$$\|\boldsymbol{\delta}_{I^c}\|_2 \le \big(\|\boldsymbol{\delta}_{I^c}\|_1\|\boldsymbol{\delta}_{I^c}\|_\infty\big)^{1/2} \le (1.2/k^*)^{1/2}\big(\|\boldsymbol{\delta}_I\|_1\|\boldsymbol{\delta}_{I^c}\|_1\big)^{1/2}. \tag{67}$$

By the definition of index sets G and I, we have $I^c \subset G$ and $G^c \subset I$. Then by (65) and (67) we obtain

$$\|\boldsymbol{\delta}_{I^c}\|_2 \le (1.2/k^*)^{1/2}\big(\|\boldsymbol{\delta}_{G^c}\|_1\|\boldsymbol{\delta}_G\|_1\big)^{1/2} \le 1.2\|\boldsymbol{\delta}_{G^c}\|_1/\sqrt{k^*}.$$

By the norm inequality between ℓ_1-norm and ℓ_2-norm, we have

$$\|\boldsymbol{\delta}_{I^c}\|_2 \le 1.2\|\boldsymbol{\delta}_{G^c}\|_1/\sqrt{k^*} \le 1.2\sqrt{2s^*/k^*}\|\boldsymbol{\delta}_{G^c}\|_2 \le 1.2\|\boldsymbol{\delta}_I\|_2,$$

where we use $k^* \ge 2s^*$ and the induction assumption that $|G| \le 2s^*$. Then triangle inequality for ℓ_2-norm yields that

$$\|\boldsymbol{\delta}\|_2 \le \|\boldsymbol{\delta}_{I^c}\|_2 + \|\boldsymbol{\delta}_I\|_2 \le 2.2\|\boldsymbol{\delta}_I\|_2. \tag{68}$$

Note that (65) and (68) are equivalent to

$$\big\|\widehat{\boldsymbol{\beta}}^{(\ell)}_j - \boldsymbol{\beta}^*_j\big\|_2 \le 2.2\big\|\widehat{\boldsymbol{\beta}}^{(\ell)}_{I^\ell_j} - \boldsymbol{\beta}^*_{I^\ell_j}\big\|_2 \quad \text{and} \quad \big\|\widehat{\boldsymbol{\beta}}^{(\ell)}_j - \boldsymbol{\beta}^*_j\big\|_1 \le 2.2\big\|\widehat{\boldsymbol{\beta}}^{(\ell)}_{\widetilde{G}^\ell_j} - \boldsymbol{\beta}^*_{\widetilde{G}^\ell_j}\big\|_1,$$

where $\widetilde{G}^\ell_j = \big(G^\ell_j\big)^c$, which concludes the proof. ■

C.3. Proof of Lemma 12

Proof We first show that $\widehat{\boldsymbol{\beta}}_j^{(\ell)}$ stays in the ℓ_1-ball centered at $\boldsymbol{\beta}_j^*$ with radius $r = C_\rho s^* \sqrt{\log d/n}$, where $C_\rho \ge 33\rho_*^{-1}$. For notational simplicity, we denote $\boldsymbol{\delta} = \widehat{\boldsymbol{\beta}}_j^{(\ell)} - \widehat{\boldsymbol{\beta}}_j$ and write S_j, G_j^ℓ, J_j^ℓ and I_j^ℓ as S, G, J an I respectively. We prove by contradiction. Suppose that $\|\boldsymbol{\delta}\|_1 > r$, then we define $\widetilde{\boldsymbol{\beta}}_j = \boldsymbol{\beta}_j^* + t(\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}_j^*) \in \mathbb{R}^{d-1}$ with $t \in [0,1]$ such that $\big\|\widetilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\big\|_1 \le r$. Letting $\widetilde{\boldsymbol{\delta}} := \widetilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*$, by (68) we obtain

$$\|\widetilde{\boldsymbol{\delta}}\|_2 = t\|\boldsymbol{\delta}\|_2 \le 2.2t\|\boldsymbol{\delta}_I\|_2 = 2.2\|\widetilde{\boldsymbol{\delta}}_I\|_2. \tag{69}$$

Moreover, by Lemma (11) and the relation between ℓ_1- and ℓ_2-norms we have

$$\|\widetilde{\boldsymbol{\delta}}\|_1 = t\|\boldsymbol{\delta}\|_1 \le 2.2t\|\boldsymbol{\delta}_{G^c}\|_1 \le 2.2\sqrt{2s^*}\|\widetilde{\boldsymbol{\delta}}_I\|_2, \tag{70}$$

where we use the fact that $G^c \subset I$ and the induction assumption that $|G^c| \le 2s^*$. By Mean-Value theorem, there exists a $\gamma \in [0,1]$ such that $\nabla L_j(\widetilde{\boldsymbol{\beta}}_j) - \nabla L_j(\boldsymbol{\beta}_j^*) = \nabla^2 L_j(\boldsymbol{\beta}_1)\widetilde{\boldsymbol{\delta}}$, where $\boldsymbol{\beta}_1 := \gamma\boldsymbol{\beta}_j^* + (1-\gamma)\widetilde{\boldsymbol{\beta}}_j \in \mathbb{R}^{d-1}$. In what follows we will derive an upper bound for $\|\widetilde{\boldsymbol{\delta}}_I\|_2$ from $\widetilde{\boldsymbol{\delta}}^T\nabla^2 L_j(\boldsymbol{\beta}_1)\widetilde{\boldsymbol{\delta}}$. Before doing that, we present two lemmas. The first one shows that the restricted correlation coefficients defined as follows are closely related to the sparse eigenvalues. This lemma also appear in Zhang (2010) and Zhang et al. (2013) for ℓ_2-loss.

Lemma 20 *(Local sparse eigenvalues and restricted correlation coefficients) Let m be a positive integer and $\mathbf{M}(\cdot)\colon \mathbb{R}^m \to \mathbb{S}^m$ be a mapping from $\mathbb{R}^m$ to the space of $m\times m$ symmetric matrices. We define the s-sparse eigenvalues of $\mathbf{M}(\cdot)$ over the ℓ_1-ball centered at $\mathbf{u}_0 \in \mathbb{R}^m$ with radius r as*

$$\rho_+\big(\mathbf{M}, \mathbf{u}_0; s, r\big) = \sup_{\mathbf{v},\mathbf{u}\in\mathbb{R}^m}\big\{\mathbf{v}^T\mathbf{M}(\mathbf{u})\mathbf{v}\colon \|\mathbf{v}\|_0 \le s, \|\mathbf{v}\|_2 = 1, \|\mathbf{u}-\mathbf{u}_0\|_1 \le r\big\};$$
$$\rho_-\big(\mathbf{M}, \mathbf{u}_0; s, r\big) = \inf_{\mathbf{v},\mathbf{u}\in\mathbb{R}^m}\big\{\mathbf{v}^T\mathbf{M}(\mathbf{u})\mathbf{v}\colon \|\mathbf{v}\|_0 \le s, \|\mathbf{v}\|_2 = 1, \|\mathbf{u}-\mathbf{u}_0\|_1 \le r\big\}.$$

In addition, we define the restricted correlation coefficients of $\mathbf{M}$ over the ℓ_1-ball centered at $\mathbf{u}_0$ with radius r as

$$\pi\big(\mathbf{M}, \mathbf{u}_0; s, k, r\big) := \sup_{\mathbf{v},\mathbf{w},\mathbf{u}\in\mathbb{R}^m}\left\{\frac{\mathbf{v}_I^T\mathbf{M}(\mathbf{u})\mathbf{w}_J\big\|\mathbf{v}_I\big\|_2}{\mathbf{v}_I^T\mathbf{M}(\mathbf{u})\mathbf{v}_I\big\|\mathbf{w}_J\big\|_\infty} : I\cap J = \emptyset, |I|\le s, |J|\le k, \big\|\mathbf{u}-\mathbf{u}_0\big\|_1 \le r\right\}.$$

Suppose that the local sparse eigenvalue $\rho_-\big(\mathbf{M}, \mathbf{u}_0; s+k, r\big) > 0$, then we have the following upper bound on the restricted correlation coefficient $\pi(\mathbf{M}, \mathbf{u}_0; s, k)$:

$$\pi\big(\mathbf{M}, \mathbf{u}_0; s, k, r\big) \le \frac{\sqrt{k}}{2}\sqrt{\rho_+\big(\mathbf{M}, \mathbf{u}_0; k, r\big)\Big/\rho_-\big(\mathbf{M}, \mathbf{u}_0; s+k, r\big) - 1}.$$

Proof See §E.1.1 for a detailed proof. ■

We denote the restricted correlation coefficients of $\nabla^2 L_j(\cdot)$ over the ℓ_1-ball centered at $\boldsymbol{\beta}_j^*$ with radius r as $\pi_j(s_1, s_2) := \pi\big(\nabla^2 L_j, \boldsymbol{\beta}_j^*; s_1, s_2, r\big)$ and denote the s-sparse eigenvalues $\rho_-\big(\nabla^2 L_j, \boldsymbol{\beta}_j^*; s, r\big)$ and $\rho_+\big(\nabla^2 L_j, \boldsymbol{\beta}_j^*; s, r\big)$ as $\rho_{j-}(s)$ and $\rho_{j+}(s)$ respectively. Applying Lemma 20 to $\pi_j(2s^*{+}k^*, k^*)$ we obtain

$$\pi_j(2s^*{+}k^*, k^*) \le {k^*}^{1/2}/2 \cdot \sqrt{\rho_{j+}(k^*)/\rho_{j-}(2s^*{+}2k^*) - 1}. \tag{71}$$

By the law of large numbers, if the sample size n is sufficiently large such that $\nabla^2 L_j$ is close to its expectation $\mathbb{E}\big[\nabla^2 L_j\big]$. When $\boldsymbol{\beta}_j$ is close to $\boldsymbol{\beta}_j^*$, by Assumption 4, we expect that the sparse eigenvalue condition also holds for $\nabla^2 L_j(\boldsymbol{\beta}_j)$ with high probability. The following lemma justifies this intuition.

Lemma 21 *Recall that we define the sparse eigenvalues of $\mathbb{E}\big[\nabla^2 L_j(\boldsymbol{\beta}_j^*)\big]$ in Definition 3. Under Assumptions 2 and 4, if n is sufficiently large such that $\rho_* \gtrsim k^* \lambda \log^2 d$, with probability at least $1-(2d)^{-1}$, for all $j \in [d]$, there exists a constant $C_\rho \ge 33\rho_*^{-1}$ such that*

$$\rho_{j-}^*(2s^*{+}2k^*) - 0.05\rho_* \le \rho_{j-}(2s^*{+}2k^*) < \rho_{j+}(k^*) \le \rho_{j+}^*(k^*) + 0.05\rho_*, \quad \text{and}$$
$$\rho_{j+}(k^*)/\rho_{j-}(2s^*{+}2k^*) \le 1 + 0.27k^*/s^*,$$

where we denote the local sparse eigenvalues $\rho_-\big(\nabla^2 L_j, \boldsymbol{\beta}_j^*; s, r\big)$ and $\rho_+\big(\nabla^2 L_j, \boldsymbol{\beta}_j^*; s, r\big)$ with $r = C_\rho\sqrt{\log d/n}$ as $\rho_{j-}(s)$ and $\rho_{j+}(s)$, respectively.

Proof See §E.1.2 for a detailed proof. ■

Thus by Lemma 21 we have

$$\pi_j(2s^*{+}k^*, k^*) \le 0.5\sqrt{0.27{k^*}^2/s^*}. \tag{72}$$

By (65), (72) and $G^c \subset I$ we obtain

$$1 - 2\pi_j(2s^*{+}k^*, k^*){k^*}^{-1}\|\widetilde{\boldsymbol{\delta}}_G\|_1/\|\widetilde{\boldsymbol{\delta}}_I\|_2 \ge 1 - 1.2\sqrt{0.54} := \kappa_1, \tag{73}$$

where we denote $\kappa_1 := 1 - 1.2\sqrt{0.54} \ge 0.11$. Now we use the second lemma to get an lower bound of $\widetilde{\boldsymbol{\delta}}^T \nabla^2 L_j(\boldsymbol{\beta}_1)\widetilde{\boldsymbol{\delta}}$, which implies an upper bound for $\|\widetilde{\boldsymbol{\delta}}_I\|_2$.

Lemma 22 *Let $\mathbf{M}\colon \mathbb{R}^m \to \mathbb{S}^m$ be a mapping from $\mathbb{R}^m$ to the space of $m \times m$-symmetric matrices. Suppose that the sparse eigenvalue $\rho_-\big(\mathbf{M}, \mathbf{u}_0; s+k, r\big) > 0$, let the restricted correlation coefficients of $\mathbf{M}(\cdot)$ be defined in Lemma 20. We denote the restricted correlation coefficients $\pi\big(\mathbf{M}, \mathbf{u}_0; s, k, r\big)$ and s-sparse eigenvalue $\rho_-\big(\mathbf{M}, \mathbf{u}_0; s, r\big)$ as $\pi(s,k)$ and $\rho_-(s)$ respectively for notational simplicity. For any $\mathbf{v} \in \mathbb{R}^d$, let F be any index set such that $|F^c| \le s$, let J be the set of indices of the largest k entries of $\mathbf{v}_F$ in absolute value and let $I = F^c \cup J$. For any $\mathbf{u} \in \mathbb{R}^d$ such that $\|\mathbf{u} - \mathbf{u}_0\|_2 \le r$ and any $\mathbf{v} \in \mathbb{R}^d$ satisfying $1 - 2\pi(s{+}k, k)\|\mathbf{v}_F\|_1/\|\mathbf{v}_I\|_2 > 0$ we have*

$$\mathbf{v}^T\mathbf{M}(\mathbf{u})\mathbf{v} \ge \rho_-(s{+}k)\big[\|\mathbf{v}_I\|_2 - 2\pi(s{+}k, k)\|\mathbf{v}_F\|_1/k\big]\|\mathbf{v}_I\|_2.$$

Proof See §E.1.3 for a detailed proof. ■

Now applying Lemma 22 to $\nabla^2 L_j(\cdot)$ with $F = G$, $s = 2s^*$ and $k = k^*$ we obtain

$$\widetilde{\boldsymbol{\delta}}^T \nabla^2 L_j(\boldsymbol{\beta}_1)\widetilde{\boldsymbol{\delta}} \geq \rho_{j-}(2s^*+k^*)\|\widetilde{\boldsymbol{\delta}}_I\|_2\big[\|\widetilde{\boldsymbol{\delta}}_I\|_2 - 2\pi_j(2s^*+k^*, k^*)/k^*\|\widetilde{\boldsymbol{\delta}}_G\|_1\big]. \tag{74}$$

Then by (73), the right-hand side of (74) can be lower bounded by

$$\widetilde{\boldsymbol{\delta}}^T \nabla^2 \ell(\boldsymbol{\beta}_1)\widetilde{\boldsymbol{\delta}} \geq \kappa_1 \rho_{j-}(2s^*+k^*)\|\widetilde{\boldsymbol{\delta}}_I\|_2^2 \geq 0.95\kappa_1\rho_*\|\widetilde{\boldsymbol{\delta}}_I\|_2^2 = \kappa_2\rho_*\|\widetilde{\boldsymbol{\delta}}_I\|_2^2, \tag{75}$$

where we let $\kappa_2 := 0.95\kappa_1 \geq 0.1$. Now we derive an upper bound for $\widetilde{\boldsymbol{\delta}}^T \nabla^2 L_j(\boldsymbol{\beta}_1)\widetilde{\boldsymbol{\delta}}$. We define the symmetric Bregman divergence of $L_j(\boldsymbol{\beta}_j)$ as $D_j(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) := \big\langle \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2, \nabla L_j(\boldsymbol{\beta}_1) - \nabla L_j(\boldsymbol{\beta}_2)\big\rangle$, where $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^{d-1}$. Then by definition, $\widetilde{\boldsymbol{\delta}}^T \nabla^2 \ell(\boldsymbol{\beta}_1)\widetilde{\boldsymbol{\delta}} = D_j(\widetilde{\boldsymbol{\beta}}_j, \boldsymbol{\beta}_j^*)$. The following lemma relates $D_j(\widetilde{\boldsymbol{\beta}}_j, \boldsymbol{\beta}_j^*)$ with $D_j(\widehat{\boldsymbol{\beta}}_j, \boldsymbol{\beta}_j^*)$.

Lemma 23 *Let* $D_j(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) := \big\langle \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2, \nabla L_j(\boldsymbol{\beta}_1) - \nabla L(\boldsymbol{\beta}_2)\big\rangle$, $\boldsymbol{\beta}(t) = \boldsymbol{\beta}_1 + t(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)$, $t \in (0, 1)$ *be any point on the line segment between* $\boldsymbol{\beta}_1$ *and* $\boldsymbol{\beta}_2$*. Then we have*

$$D_j(\boldsymbol{\beta}(t), \boldsymbol{\beta}_1) \leq tD_j(\boldsymbol{\beta}_2, \boldsymbol{\beta}_1)$$

Proof See §E.1.4 for a detailed proof. ■

By Lemma 23 and (60),

$$D_j(\widetilde{\boldsymbol{\beta}}_j, \boldsymbol{\beta}_j^*) \leq tD_j(\widehat{\boldsymbol{\beta}}_j, \boldsymbol{\beta}_j^*) \leq \underbrace{-t\big\langle \nabla L_j(\boldsymbol{\beta}_j^*), \boldsymbol{\delta}\big\rangle}_{(i)} \underbrace{-t\big\langle \boldsymbol{\delta}, \boldsymbol{\lambda}_j \circ \boldsymbol{\xi}_j\big\rangle}_{(ii)}. \tag{76}$$

For term (i) in (76), by Hölder's inequality we have

$$\begin{aligned} -t\big\langle \nabla L_j(\boldsymbol{\beta}_j^*), \boldsymbol{\delta}\big\rangle &\leq t\big\|\nabla_{G^c} L_j(\boldsymbol{\beta}_j^*)\big\|_2\|\boldsymbol{\delta}_{G^c}\|_2 + t\big\|\nabla_G L_j(\boldsymbol{\beta}_j^*)\big\|_\infty\|\boldsymbol{\delta}_G\|_1 \\ &\leq \big\|\nabla_{G^c} L_j(\boldsymbol{\beta}_j^*)\big\|_2\|\widetilde{\boldsymbol{\delta}}_I\|_2 + \big\|\nabla_G L_j(\boldsymbol{\beta}_j^*)\big\|_\infty\|\widetilde{\boldsymbol{\delta}}_G\|_1, \end{aligned} \tag{77}$$

where the inequality follows from $G^c \subset I$. For term (ii) in (76), by (62) and Hölder's inequality we have

$$-t\big\langle \boldsymbol{\delta}, \boldsymbol{\lambda}_j \circ \boldsymbol{\xi}_j\big\rangle \leq -\big\langle \boldsymbol{\delta}_S, (\boldsymbol{\lambda}_j \circ \boldsymbol{\xi}_j)_S\big\rangle - \big\langle |\widetilde{\boldsymbol{\delta}}_G|, \boldsymbol{\lambda}_G\big\rangle \leq \|\boldsymbol{\lambda}_S\|_2\|\widetilde{\boldsymbol{\delta}}_I\|_2 - p_\lambda'(c_2\lambda)\|\widetilde{\boldsymbol{\delta}}_G\|_1, \tag{78}$$

where we use the Hölder's inequality and the definition of G. Combining (75),(77) and (78) we obtain that

$$\begin{aligned} \kappa_2\rho_*\big\|\widetilde{\boldsymbol{\delta}}_I\big\|_2^2 &\leq \big(\big\|\nabla_{G^c} L_j(\boldsymbol{\beta}_j^*)\big\|_2 + \|\boldsymbol{\lambda}_S\|_2\big)\|\widetilde{\boldsymbol{\delta}}_I\|_2 + \big[\big\|\nabla L_j(\boldsymbol{\beta}_j^*)\big\|_\infty - p_\lambda'(c_2\lambda)\big]\|\widetilde{\boldsymbol{\delta}}_G\|_1 \\ &\leq \big(\big\|\nabla_{G^c} L_j(\boldsymbol{\beta}_j^*)\big\|_2 + \|\boldsymbol{\lambda}_S\|_2\big)\big\|\widetilde{\boldsymbol{\delta}}_I\big\|_2, \end{aligned}$$

where the second inequality follows from $p_\lambda'(c_2\lambda) > \big\|\nabla L_j(\boldsymbol{\beta}_j^*)\big\|_\infty$. From the inequality above and the induction assumption $|G^c| \leq 2s^*$ we obtain that

$$\big\|\widetilde{\boldsymbol{\delta}}_I\big\|_2 \leq 10\rho_*^{-1}\big(\big\|\nabla_{G^c} L_j(\boldsymbol{\beta}_j^*)\big\|_2 + \|\boldsymbol{\lambda}_S\|_2\big) \leq 10\rho_*^{-1}\sqrt{s^*}\big(\sqrt{2}\big\|\nabla L_j(\boldsymbol{\beta}_j^*)\big\|_\infty + \lambda\big). \tag{79}$$

Thus (70), (79) and the the fact that $25\big\|\nabla L_j(\boldsymbol{\beta}_j^*)\big\|_\infty \leq \lambda$ imply that

$$\|\widetilde{\boldsymbol{\delta}}\|_1 \leq 22\sqrt{2}\rho_*^{-1}(1+\sqrt{2}/25)s^*\lambda < 33\rho_*^{-1}s^*\lambda \leq r, \tag{80}$$

where the last inequality follows from the definition of λ. Notice that (80) contradicts our assumption that $\|\widetilde{\boldsymbol{\delta}}\|_1 = r$, the reason for this contradiction is because we assume that $\big\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}_j^*\big\|_1 > r$, hence $\big\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}_j^*\big\|_1 \leq r$ and $\widetilde{\boldsymbol{\beta}}_j = \widehat{\boldsymbol{\beta}}_j^{(\ell)}$. This means that $\widehat{\boldsymbol{\beta}}_j^{(\ell)}$ stays in the ℓ_1-ball centered at $\boldsymbol{\beta}_j^*$ with radius r in each iteration.

Moreover, by (68) and (79), we obtain the following upper bound for $\|\boldsymbol{\delta}_I\|_2$:

$$\|\boldsymbol{\delta}\|_2 \leq 22\rho_*^{-1}\big(\big\|\nabla_{G^c}L_j(\boldsymbol{\beta}_j^*)\big\|_2 + \|\boldsymbol{\lambda}_S\|_2\big) \leq 24\rho_*^{-1}\sqrt{s^*}\lambda,$$

where we use the condition that $\lambda \geq 25\big\|\nabla L_j(\boldsymbol{\beta}_j^*)\big\|_\infty$. In addition, by(65) and (79) we obtain the following bound on $\|\boldsymbol{\delta}\|_1$

$$\|\boldsymbol{\delta}\|_1 \leq 2.2\|\boldsymbol{\delta}_{G^c}\|_1 \leq 22\sqrt{2s^*}\rho_*^{-1}\Big(\big\|\nabla_{G^c}L_j(\boldsymbol{\beta}_j^*)\big\|_2 + \|\boldsymbol{\lambda}_S\|_2\Big) \leq 33\rho_*^{-1}s^*\lambda, \tag{81}$$

Therefore going back to the original notations, note that $\kappa_2 \geq 0.1$, we establish the following crude rates of convergence for $\ell \geq 1$:

$$\big\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}_j^*\big\|_2 \leq 24\rho_*^{-1}\sqrt{s^*}\lambda \;\text{ and }\; \big\|\widehat{\boldsymbol{\beta}}_j^{(\ell)} - \boldsymbol{\beta}_j^*\big\|_1 \leq 33\rho_*^{-1}s^*\lambda. \tag{82}$$

And (79) is equivalent to

$$\big\|\widehat{\boldsymbol{\beta}}_{I_j^\ell}^{(\ell)} - \boldsymbol{\beta}_{I_j^\ell}^*\big\|_2 \leq 10\rho_*^{-1}\Big(\big\|\nabla_{\widetilde{G}_j^\ell}L_j(\boldsymbol{\beta}_j^*)\big\|_2 + \big\|\boldsymbol{\lambda}_{S_j}^{(\ell-1)}\big\|_2\Big), \;\; \widetilde{G}_j^\ell := (G_j^\ell)^c. \tag{83}$$

Note that we use Lemmas 10 and 21, hence (83) and (83) hold with probability at least $1 - d^{-1}$ for all $j \in [d]$. ■

Appendix D. Proof of Auxiliary Results for Asymptotic Inference

We prove the auxiliary results for asymptotic inference. More specifically, we first prove Lemma 14, which is pivotal for deriving the limiting distribution of the pairwise score statistic. Then we prove the lemmas presented in the proof of Theorem 8.

D.1. Proof of Lemma 14

Proof Before proving this lemma, we first let $\nabla^2 L_{jk}\big(\boldsymbol{\beta}_{j\vee k}\big)$ be the Hessian of $L_{jk}\big(\boldsymbol{\beta}_{j\vee k}\big)$ and define $\mathbf{H}^{jk} := \mathbb{E}\big[\nabla^2 L_{jk}\big(\boldsymbol{\beta}_{j\vee k}^*\big)\big]$. We also define

$$\boldsymbol{\Sigma}^{jk} := \mathbb{E}\big[\mathbf{g}_{jk}(\boldsymbol{X}_i)\mathbf{g}_{jk}(\boldsymbol{X}_i)^T\big] \;\text{ and }\; \boldsymbol{\Theta}^{jk} := \mathbb{E}\big[\mathbf{h}_{ii'}^{jk}(\boldsymbol{\beta}^*)\mathbf{h}_{ii'}^{jk}(\boldsymbol{\beta}^*)^T\big].$$

Under Assumption 2, we first show that there exists a positive constant D such that for any $j,k \in d$, $j \neq k$, $\max\big\{\big\|\boldsymbol{\Sigma}^{jk}\big\|_\infty, \big\|\mathbf{H}^{jk}\big\|_\infty, \big\|\boldsymbol{\Theta}^{jk}\big\|_\infty\big\} \leq D$. The reason is as follows.

Note that Hölder's inequality imply

$$\big\|\mathbf{H}^{jk}\big\|_\infty \lesssim \max_{j\in[d]}\mathbb{E}|X_{ij} - X_{i'j}|^4 \lesssim \max_{j\in[d]}\mathbb{E}|X_j|^4 \;\text{ for any } j,k \in [d], j \neq k.$$

Similarly, for $\mathbf{\Theta}^{jk}$, we also have $\left\|\mathbf{\Theta}^{jk}\right\|_\infty \lesssim \max_{j\in[d]} \mathbb{E}|X_j|^4$. By (14) we have

$$\mathbb{E}|X_j|^4 = \int_0^\infty \mathbb{P}(|X_4|^4 > t)dt \le \int_0^\infty c\exp(-t^{1/4})dt = 24c, \quad c = 2\exp(\kappa_m + \kappa_h/2).$$

Moreover, note that by the law of total variance, the diagonal elements of $\mathbf{\Sigma}^{jk}$ are no larger than the corresponding diagonal elements of $\mathbf{\Theta}^{jk}$; then by Cauchy-Schwarz inequality, $\|\mathbf{\Sigma}^{jk}\|_\infty \le \|\mathbf{\Theta}^{jk}\|_\infty$. Therefore there exists a constant D that does not depend on (s^*, n, d) such that

$$\max\left\{\|\mathbf{H}^{jk}\|_\infty, \|\mathbf{\Sigma}^{jk}\|_\infty, \|\mathbf{\Theta}^{jk}\|_\infty\right\} \le D, \quad 1\le j<k\le d. \tag{84}$$

Now we are ready to prove the lemma. Recall that $\nabla L_{jk}\big(\boldsymbol{\beta}_{j\vee k}\big)$ is a U-statistic with kernel function $\mathbf{h}_{ii'}^{jk}(\boldsymbol{\beta}_{j\vee k})$. Because $\mathbf{h}_{ii'}^{jk}\big(\boldsymbol{\beta}_{j\vee k}^*\big)$ is centered, the law of total expectation implies that $\mathbb{E}\big[\mathbf{g}_{jk}(\boldsymbol{X}_i)\big] = \mathbf{0}$. Note that the left-hand side of (40) can be written as

$$\begin{aligned}\frac{\sqrt{n}}{2}\mathbf{b}^T\nabla L_{jk}\big(\boldsymbol{\beta}_{j\vee k}^*\big) &= \frac{\sqrt{n}}{2}\mathbf{b}^T\mathbf{U}_{jk} + \frac{\sqrt{n}}{2}\mathbf{b}^T\left[\nabla L_{jk}\big(\boldsymbol{\beta}_{j\vee k}^*\big) - \mathbf{U}_{jk}\right] \\ &= \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{b}^T\mathbf{g}_{jk}(\boldsymbol{X}_i)}_{I_1} + \underbrace{\frac{\sqrt{n}}{2}\mathbf{b}^T\left[\nabla L_{jk}\big(\boldsymbol{\beta}_{j\vee k}^*\big) - \mathbf{U}_{jk}\right]}_{I_2}.\end{aligned}$$

Notice that I_1 is a weighted sum of i.i.d. random variables with the mean and variance given by

$$\mathbb{E}\big[\mathbf{b}^T\mathbf{g}_{jk}(\boldsymbol{X}_i)\big] = \mathbf{0} \quad \text{and} \quad \mathrm{Var}\big[\mathbf{b}^T\mathbf{g}_{jk}(\boldsymbol{X}_i)\big] = \mathbf{b}^T\mathbf{\Sigma}^{jk}\mathbf{b}.$$

Central limit theorem implies that $I_1 \rightsquigarrow N(0, \mathbf{b}^T\mathbf{\Sigma}^{jk}\mathbf{b})$. In what follows we use $\mathbf{h}_{ii'}$ and $\mathbf{h}_{ii'|i}$ to denote $\mathbf{h}_{ii'}^{jk}\big(\boldsymbol{\beta}_{j\vee k}^*\big)$ and $\mathbb{E}\big[\mathbf{h}_{ii'}^{jk}\big(\boldsymbol{\beta}_{j\vee k}^*\big)\big|\boldsymbol{X}_i\big] = \mathbf{g}_{jk}(\boldsymbol{X}_i)$. Thus we can write I_2 as

$$I_2 = \frac{1}{\sqrt{n}(n-1)}\sum_{i<i'}\mathbf{b}^T\boldsymbol{\chi}_{ii'}, \quad \text{where } \boldsymbol{\chi}_{ii'} = (\mathbf{h}_{ii'} - \mathbf{h}_{ii'|i} - \mathbf{h}_{ii'|i'}).$$

Then $\mathbb{E}(I_2^2)$ can be expanded as

$$\mathbb{E}(I_2^2) = \frac{1}{n(n-1)^2}\sum_{i<i',s<s'}\mathbf{b}^T\mathbb{E}(\boldsymbol{\chi}_{ii'}\boldsymbol{\chi}_{ss'}^T)\mathbf{b}. \tag{85}$$

By the definition of $\boldsymbol{\chi}_{ii'}$, we have

$$\begin{aligned}\mathbb{E}(\boldsymbol{\chi}_{ii'}\boldsymbol{\chi}_{ss'}^T) &= \mathbb{E}(\mathbf{h}_{ii'}\mathbf{h}_{ss'}^T) - \mathbb{E}(\mathbf{h}_{ii'}\mathbf{h}_{ss'|s}^T) - \mathbb{E}(\mathbf{h}_{ii'}\mathbf{h}_{ss'|s'}^T) - \mathbb{E}(\mathbf{h}_{ii'|i}\mathbf{h}_{ss'}^T) \\ &+ \mathbb{E}(\mathbf{h}_{ii'|i}\mathbf{h}_{ss'|s}^T) + \mathbb{E}(\mathbf{h}_{ii'|i}\mathbf{h}_{ss'|s'}^T) - \mathbb{E}(\mathbf{h}_{ii'|i'}\mathbf{h}_{ss'}^T) + \mathbb{E}(\mathbf{h}_{ii'|i'}\mathbf{h}_{ss'|s}^T) + \mathbb{E}(\mathbf{h}_{ii'|i'}\mathbf{h}_{ss'|s'}^T).\end{aligned} \tag{86}$$

Therefore, for $i \ne s, s'$ and $i' \ne s, s'$, law of total expectation implies that $\mathbb{E}(\boldsymbol{\chi}_{ii'}\boldsymbol{\chi}_{ss'}^T) = \mathbf{0}$. Similarly, if exactly one of i, i' is identical to one of s, s', say $i = s$, then (86) becomes

$$\mathbb{E}(\boldsymbol{\chi}_{ii'}\boldsymbol{\chi}_{ii''}^T) = \mathbb{E}(\mathbf{h}_{ii'}\mathbf{h}_{ii''}^T) - \mathbb{E}(\mathbf{h}_{ii'}\mathbf{h}_{ii''|i}^T) - \mathbb{E}(\mathbf{h}_{ii'|i}\mathbf{h}_{ii''}^T) + \mathbb{E}(\mathbf{h}_{ii'|i}\mathbf{h}_{ii''|i}^T), \ i\ne i'\ne i''.$$

Note that by the law of total expectation, for each term in (86) we have

$$\mathbb{E}(\mathbf{h}_{ii'}\mathbf{h}_{ii''}^T) = \mathbb{E}(\mathbf{h}_{ii'}\mathbf{h}_{ii''|i}^T) = \mathbb{E}(\mathbf{h}_{ii'|i}\mathbf{h}_{ii''}^T) = \mathbb{E}(\mathbf{h}_{ii'|i}\mathbf{h}_{ii''|i}^T).$$

Therefore, $\mathbb{E}(\boldsymbol{\chi}_{ii'}\boldsymbol{\chi}_{ii''}^T) = \mathbf{0}$. Finally, if $i = s$ and $i' = s'$, by the law of total expectation, (86) can be further reduced to $\mathbb{E}(\boldsymbol{\chi}_{ii'}\boldsymbol{\chi}_{ii'}^T) = \mathbb{E}(\mathbf{h}_{ii'}\mathbf{h}_{ii'}^T) - \mathbb{E}(\mathbf{h}_{ii'|i}\mathbf{h}_{ii'|i}^T) - \mathbb{E}(\mathbf{h}_{ii'|i'}\mathbf{h}_{ii'|i'}^T) = \boldsymbol{\Theta}^{jk} - 2\boldsymbol{\Sigma}^{jk}$. Thus by triangle inequality we have

$$\big\|\mathbb{E}(\boldsymbol{\chi}_{ii'}\boldsymbol{\chi}_{ii'}^T)\big\|_\infty \le \big\|\mathbb{E}(\mathbf{h}_{ii'}\mathbf{h}_{ii'}^T)\big\|_\infty + \big\|\mathbb{E}(\mathbf{h}_{ii'|i}\mathbf{h}_{ii'|i}^T)\big\|_\infty + \big\|\mathbb{E}(\mathbf{h}_{ii'|j}\mathbf{h}_{ii'|j}^T)\big\|_\infty \le 3D,$$

where the last inequality follows from Assumption 6. Then equation (85) can be reduced to

$$\mathbb{E}(I_2^2) = \frac{1}{n(n-1)^2}\sum_{i<i',s<s'} \mathbf{b}^T\mathbb{E}(\boldsymbol{\chi}_{ii'}\boldsymbol{\chi}_{ss}^T)\mathbf{b} = \frac{1}{n(n-1)^2}\sum_{i<i'} \mathbf{b}^T\mathbb{E}(\boldsymbol{\chi}_{ii'}\boldsymbol{\chi}_{ii'}^T)\mathbf{b}.$$

By Hölder's inequality we obtain

$$\begin{aligned}\mathbb{E}(I_2^2) &\le \frac{1}{2(n-1)}\|\mathbf{b}\|_1\big\|\mathbb{E}(\boldsymbol{\chi}_{ii'}\boldsymbol{\chi}_{ii'}^T)\mathbf{b}\big\|_\infty \\ &\le \frac{1}{2(n-1)}\|\mathbf{b}\|_1^2\big\|\mathbb{E}(\boldsymbol{\chi}_{ii'}\boldsymbol{\chi}_{ii'}^T)\big\|_\infty \le \frac{3D}{2(n-1)}\|\mathbf{b}\|_1^2. \end{aligned} \tag{87}$$

Since $\|\mathbf{b}\|_0 \le \widetilde{s}$, by the relationship between ℓ_1-norm and ℓ_2-norm, we can further bound the right-hand side of (87) by $\mathbb{E}(I_2^2) \le 1.5\widetilde{s}D/(n-1) \to 0$, where we use the condition that $\lim_{n\to\infty} \widetilde{s}/n = 0$. Therefore, we conclude the proof of Lemma 14. ■

D.2. Proof of Lemma 13

Proof By the definition of $\mathbf{w}_{j,k}^*$ we have $\mathbf{H}_{jk,j\backslash k}^j = \mathbf{w}_{j,k}^{*T}\mathbf{H}_{j\backslash k,j\backslash k}^j$. We let $\widehat{\boldsymbol{\beta}}_j' = (0, \widehat{\boldsymbol{\beta}}_{j\backslash k})$ and denote $\nabla^2 L_j(\widehat{\boldsymbol{\beta}}_j')$ and $\nabla^2 L_j(\boldsymbol{\beta}_j^*)$ as $\boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda}^*$ respectively. In addition, we write $\mathbf{H}^j, \mathbf{w}_{j,k}^*$ and $\widehat{\mathbf{w}}_{j,k}$ as $\mathbf{H}, \mathbf{w}^*$ and $\widehat{\mathbf{w}}$ respectively for notational simplicity. Triangle inequality implies that

$$\|\boldsymbol{\Lambda}_{jk,j\backslash k} - \mathbf{w}^{*T}\boldsymbol{\Lambda}_{j\backslash k,j\backslash k}\|_\infty \le \|\mathbf{H}_{jk,j\backslash k} - \boldsymbol{\Lambda}_{jk,j\backslash k}\|_\infty + \|\mathbf{w}^{*T}(\mathbf{H}_{j\backslash k,j\backslash k} - \boldsymbol{\Lambda}_{j\backslash k,j\backslash k})\|_\infty.$$

Hölder's inequality implies that

$$\|\boldsymbol{\Lambda}_{jk,j\backslash k} - \mathbf{w}^{*T}\boldsymbol{\Lambda}_{j\backslash k,j\backslash k}\|_\infty \le \|\boldsymbol{\Lambda} - \mathbf{H}\|_\infty(1 + \|\mathbf{w}^*\|_1). \tag{88}$$

Under null hypothesis, $\beta_{jk}^* = 0$. By Lemma 26, we have $\|\boldsymbol{\Lambda} - \mathbf{H}\|_\infty \lesssim s^*\lambda\log^2 d$. Then the right-hand side of (88) is bounded by

$$\|\boldsymbol{\Lambda}_{jk,j\backslash k} - \mathbf{w}^{*T}\boldsymbol{\Lambda}_{j\backslash k,j\backslash k}\|_\infty \lesssim (w_0 + 1)s^*\lambda\log^2 d.$$

Therefore, by the assumption that $\lambda_D \gtrsim \max\{1, w_0\}s^*\lambda\log^2 d$ we can ensure that $\mathbf{w}^*$ is in the feasible region of the Dantzig selector problem (11), hence we have $\|\widehat{\mathbf{w}}\|_1 \le \|\mathbf{w}^*\|_1 \le w_0$

by the optimality of $\widehat{\mathbf{w}}$. Let J be the support set of $\mathbf{w}^*$, that is, $J := \{(j,\ell)\colon [\mathbf{w}^*_{j,k}]_{j\ell} \neq 0, \ell \in [d], \ell \neq j\}$; the optimality of $\mathbf{w}^*$ is equivalent to $\|\widehat{\mathbf{w}}_{J^c}\|_1 + \|\widehat{\mathbf{w}}_J\|_1 \leq \|\mathbf{w}^*_J\|_1$. By triangle inequality, we have

$$\|\widehat{\mathbf{w}}_{J^c} - \mathbf{w}^*_{J^c}\|_1 = \|\widehat{\mathbf{w}}_{J^c}\|_1 \leq \|\mathbf{w}^*_J\|_1 - \|\widehat{\mathbf{w}}_J\|_1 \leq \|\widehat{\mathbf{w}}_J - \mathbf{w}^*_J\|_1, \tag{89}$$

where $J^c := \{(j,\ell)\colon (j,\ell) \notin J, j \text{ fixed}\}$. Letting $\widehat{\boldsymbol{\omega}} = \widehat{\mathbf{w}} - \mathbf{w}^*$, inequality (89) is equivalent to $\|\widehat{\boldsymbol{\omega}}_{J^c}\|_1 \leq \|\widehat{\boldsymbol{\omega}}_J\|_1$. Moreover, triangle inequality yields that

$$\|\mathbf{\Lambda}_{j\backslash k, j\backslash k}\widehat{\boldsymbol{\omega}}\|_\infty \leq \|\mathbf{\Lambda}_{jk, j\backslash k} - \mathbf{\Lambda}_{j\backslash k, j\backslash k}\widehat{\mathbf{w}}\|_\infty + \|\mathbf{\Lambda}_{jk, j\backslash k} - \mathbf{\Lambda}_{j\backslash k, j\backslash k}\mathbf{w}^*\|_\infty \leq 2\lambda_D,$$

where the last inequality follows from that both $\mathbf{w}^*$ and $\widehat{\mathbf{w}}$ are feasible for the Dantzig selector problem (11). Then triangle inequality implies that

$$|\widehat{\boldsymbol{\omega}}^T\mathbf{\Lambda}_{j\backslash k, j\backslash k}\widehat{\boldsymbol{\omega}}| \leq \underbrace{|\widehat{\boldsymbol{\omega}}_J^T\mathbf{\Lambda}_{J, j\backslash k}\widehat{\boldsymbol{\omega}}|}_{A_1} + \underbrace{|\boldsymbol{\omega}_{J^c}^T\mathbf{\Lambda}_{J^c, j\backslash k}\widehat{\boldsymbol{\omega}}|}_{A_2}.$$

By Hölder's inequality and inequality between ℓ_1-norm and ℓ_2-norms, we obtain that

$$A_1 \leq 2\lambda_D\|\widehat{\boldsymbol{\omega}}_J\|_1 \leq 2\sqrt{s_0^\star}\lambda_D\|\widehat{\boldsymbol{\omega}}_J\|_2 \quad \text{and} \quad A_2 \leq 2\lambda_D\|\widehat{\boldsymbol{\omega}}_{J^c}\|_1 \leq 2\lambda_D\|\widehat{\boldsymbol{\omega}}_J\|_1 \leq 2\sqrt{s_0^\star}\lambda_D\|\widehat{\boldsymbol{\omega}}_J\|_2.$$

Hence we conclude that $|\widehat{\boldsymbol{\omega}}^T\mathbf{\Lambda}_{j\backslash k, j\backslash k}\widehat{\boldsymbol{\omega}}| \leq 4\sqrt{s_0^\star}\lambda_D\|\widehat{\boldsymbol{\omega}}_J\|_2$.

We let J_1 be the set of indices of the largest $k_0^\star$ component of $\widehat{\boldsymbol{\omega}}_{J^c}$ in absolute value and let $I = J_1 \cup J$, then $|I| \leq s_0^\star + k_0^\star$. Under the null hypothesis, $\|\widehat{\boldsymbol{\beta}}'_j - \boldsymbol{\beta}^*_j\|_1 = \|\widehat{\boldsymbol{\beta}}_{j\backslash k} - \boldsymbol{\beta}^*_{j\backslash k}\|_1 \leq 33\rho_*^{-1}s^*\lambda$. We denote the s-sparse eigenvalue of $\nabla^2_{j\backslash k, j\backslash k}L_j(\boldsymbol{\beta}_j)$ over the ℓ_1-ball centered at $\boldsymbol{\beta}^*_j$ with radius r as $\rho'_{j+}(s)$ and $\rho'_{j-}(s)$ respectively and denote the corresponding restricted correlation coefficients as $\pi'_j(s_1, s_2)$. And we denote these quantities of $\nabla^2 L_j(\boldsymbol{\beta}^*_j)$ as $\rho_{j-}(s), \rho_{j+}(s)$ and $\pi_j(s_1, s_2)$. By definition, we immediately have $\rho_{j-}(s) \leq \rho'_{j-}(s) \leq \rho'_{j+}(s) \leq \rho_{j+}(s)$.

By Lemma 22 we have

$$|\widehat{\boldsymbol{\omega}}^T\mathbf{\Lambda}_{j\backslash k, j\backslash k}\widehat{\boldsymbol{\omega}}| \geq \rho'_{j-}\big(k^\star + s^\star\big)\big[\|\widehat{\boldsymbol{\omega}}_I\|_2 - 2\pi'_j(s^\star + k_0^\star, s_0^\star)\|\widehat{\boldsymbol{\omega}}_{J^c}\|_1/k^\star\big]\|\widehat{\boldsymbol{\omega}}_I\|_2. \tag{90}$$

The following lemma relates the sparse eigenvalues of $\nabla^2 L_j(\boldsymbol{\beta}_j)$ to those of $\mathbb{E}\nabla^2 L_j(\boldsymbol{\beta}^*_j)$.

Lemma 24 *Under Assumptions 2, 4 and 7, if n is sufficiently large such that $\rho_* \gtrsim s^*\lambda\log^2 d$, with probability at least $1-(2d)^{-1}$, for all $j \in [d]$, there exists a constant $C_\rho \geq 33\rho_*^{-1}$ such that*

$$\rho^*_{j-}(2s_0^\star + 2k_0^\star) - 0.05\nu_* \leq \rho_{j-}(2s_0^\star + 2k_0^\star) < \rho_{j+}(k_0^\star) \leq \rho^*_{j+}(k_0^\star) + 0.05\nu_*, \quad \textit{and}$$
$$\rho_{j+}(k_0^\star)/\rho_{j-}(2s_0^\star + 2k_0^\star) \leq 1 + 0.58k_0^\star/s_0^\star,$$

where we denote the local sparse eigenvalues $\rho_-\big(\nabla^2 L_j, \boldsymbol{\beta}^*_j; s, r\big)$ and $\rho_+\big(\nabla^2 L_j, \boldsymbol{\beta}^*_j; s, r\big)$ with $r = C_\rho\sqrt{\log d/n}$ as $\rho_{j-}(s)$ and $\rho_{j+}(s)$, respectively.

Proof The proof is similar to that of Lemma 3, hence is omitted here. ■

By $\|\widehat{\boldsymbol{\omega}}_{J^c}\|_1 \le \|\widehat{\boldsymbol{\omega}}_J\|_1 \le \sqrt{s_0^\star}\|\widehat{\boldsymbol{\omega}}_J\|_2$ and Lemma 24, the right-hand side of (90) can be reduced to

$$|\widehat{\boldsymbol{\omega}}^T \boldsymbol{\Lambda}_{j\backslash k, j\backslash k}\widehat{\boldsymbol{\omega}}| \ge 0.95\nu_*\big(\|\widehat{\boldsymbol{\omega}}_I\|_2 - 2\pi_j'(s_0^\star{+}k_0^\star, s^\star)\|\widehat{\boldsymbol{\omega}}_J\|_2\sqrt{s_0^\star}/k_0^\star)\|\widehat{\boldsymbol{\omega}}_I\|_2. \tag{91}$$

Using Lemma 20 we obtain

$$\begin{aligned}
&2\pi_j'(s_0^\star{+}k_0^\star, k_0^\star)\sqrt{s_0^\star}/k_0^\star \le \sqrt{s_0^\star/k_0^\star}\sqrt{\rho_{j+}'(k_0^\star)/\rho_{j-}'(s_0^\star{+}2k_0^\star) - 1} \\
&\qquad \le \sqrt{s_0^\star/k_0^\star}\sqrt{\rho_{j+}(k_0^\star)/\rho_{j-}(s_0^\star{+}2k_0^\star) - 1} \le \sqrt{s_0^\star/k_0^\star}\sqrt{0.58k_0^\star/s_0^\star} \le 0.76.
\end{aligned}$$

Thus the right-hand side of (91) can be reduced to

$$|\widehat{\boldsymbol{\omega}}^T \boldsymbol{\Lambda}_{j\backslash k, j\backslash k}\widehat{\boldsymbol{\omega}}| \ge 0.95\nu_*(1 - 0.76\|\widehat{\boldsymbol{\omega}}_J\|_2/\|\widehat{\boldsymbol{\omega}}_I\|_2)\|\widehat{\boldsymbol{\omega}}_I\|_2^2 \ge \nu_*\kappa\|\widehat{\boldsymbol{\omega}}_I\|_2^2, \tag{92}$$

where $\kappa = 0.22$. This inequality holds because $J \subset I$. By (92) we have

$$\nu_*\kappa\|\widehat{\boldsymbol{\omega}}_I\|_2^2 \le 4\sqrt{s_0^\star}\lambda_d\|\widehat{\boldsymbol{\omega}}_J\|_2 \le 4\sqrt{s_0^\star}\lambda_d\|\widehat{\boldsymbol{\omega}}_I\|_2, \quad \text{which implies } \|\widehat{\boldsymbol{\omega}}_I\|_2 \le 4\nu_*^{-1}\kappa^{-1}\sqrt{s_0^\star}\lambda_D.$$

Therefore the estimation error of $\widehat{\mathbf{w}}_{j,k}$ can be bounded by

$$\|\widehat{\boldsymbol{\omega}}\|_1 \le 2\|\widehat{\boldsymbol{\omega}}_J\|_1 \le 2\sqrt{s^\star}\|\widehat{\boldsymbol{\omega}}_J\|_2 \le 8\nu_*^{-1}\kappa^{-1}s_0^\star\lambda_D \le 37\nu_*^{-1}s_0^\star\lambda_D.$$

Returning to the original notations, we conclude that $\|\widehat{\mathbf{w}}_{j,k} - \mathbf{w}_{j,k}^*\|_1 \le 37\nu_*^{-1}s_0^\star\lambda_D$ for all (j,k) such that $j,k{\in}[d]$, $j \ne k$. ■

D.3. Proof of Lemma 15

Proof We only need to show that $\widehat{\sigma}_{jk}^2$ is a consistent estimator of σ_{jk}^2, which is equivalent to showing that $\lim\limits_{n\to\infty}|\widehat{\sigma}_{jk}^2 - \sigma_{jk}^2| = 0$. To begin with, triangle inequality implies that

$$\begin{aligned}
|\widehat{\sigma}_{jk}^2 - \sigma_{jk}^2| \le{}& \underbrace{\big|\widehat{\boldsymbol{\Sigma}}_{jk,jk}^{jk} - \boldsymbol{\Sigma}_{jk,jk}^{jk}\big|}_{I_1} + 2\underbrace{\big|\widehat{\mathbf{w}}_{j,k}^T\widehat{\boldsymbol{\Sigma}}_{j\backslash k,jk}^{jk} - \mathbf{w}_{j,k}^{*T}\boldsymbol{\Sigma}_{j\backslash k,jk}^{jk}\big|}_{I_{2j}} + \underbrace{\big|\widehat{\mathbf{w}}_{j,k}^T\widehat{\boldsymbol{\Sigma}}_{j\backslash k,j\backslash k}^{jk}\widehat{\mathbf{w}}_{j,k} - \mathbf{w}_{j,k}^{*T}\boldsymbol{\Sigma}_{j\backslash k,j\backslash k}^{jk}\mathbf{w}_{j,k}^*\big|}_{I_{3j}} \\
&+ 2\underbrace{\big|\widehat{\mathbf{w}}_{k,j}^T\widehat{\boldsymbol{\Sigma}}_{k\backslash j,jk}^{jk} - \mathbf{w}_{k,j}^{*T}\boldsymbol{\Sigma}_{k\backslash j,jk}^{jk}\big|}_{I_{2k}} + \underbrace{\big|\widehat{\mathbf{w}}_{k,j}{}^T\widehat{\boldsymbol{\Sigma}}_{k\backslash j,k\backslash j}^{jk}\widehat{\mathbf{w}}_{k,j} - \mathbf{w}_{k,j}^{*T}\boldsymbol{\Sigma}_{k\backslash j,k\backslash j}^{jk}\mathbf{w}_{k,j}^*\big|}_{I_{3k}},
\end{aligned}$$

where $\widehat{\boldsymbol{\Sigma}}^{jk} = \widehat{\boldsymbol{\Sigma}}^{jk}\big(\widehat{\boldsymbol{\beta}}_{j\vee k}'\big)$ and $\widehat{\boldsymbol{\Sigma}}^{jk}\big(\boldsymbol{\beta}_{j\vee k}\big)$ is defined as

$$\widehat{\boldsymbol{\Sigma}}^{jk}(\boldsymbol{\beta}_{j\vee k}) = \frac{1}{n}\sum_{i=1}^n\Big\{\frac{1}{n-1}\sum_{i'\ne i}\mathbf{h}_{ii'}^{jk}(\boldsymbol{\beta}_{j\vee k})\Big\}^{\otimes 2}. \tag{93}$$

To prove the consistency of $\widehat{\sigma}_{jk}^2$, we need the following theorem to show that $\widehat{\boldsymbol{\Sigma}}^{jk}$ is a consistent estimator of $\boldsymbol{\Sigma}^{jk}$ in the sense that $\big\|\widehat{\boldsymbol{\Sigma}}^{jk} - \boldsymbol{\Sigma}^{jk}\big\|_\infty$ is negligible.

Lemma 25 *For* $1\le j<k\le d$, *let* $\widehat{\boldsymbol{\Sigma}}^{jk}\big(\boldsymbol{\beta}_{j\vee k}\big)$ *be defined as* (93). *Suppose* $\widehat{\boldsymbol{\beta}}_j$ *and* $\widehat{\boldsymbol{\beta}}_k$ *are the estimators of* $\boldsymbol{\beta}^*_j$ *and* $\boldsymbol{\beta}^*_k$ *obtained from Algorithm 1 and we denote* $\widehat{\boldsymbol{\beta}}_{j\vee k}=(\widehat{\beta}_{jk},\widehat{\boldsymbol{\beta}}^T_{j\setminus k},\widehat{\boldsymbol{\beta}}^T_{k\setminus j})^T$. *Then* $\widehat{\boldsymbol{\Sigma}}^{jk}(\widehat{\boldsymbol{\beta}}_{j\vee k})$ *is a consistent estimator of* $\boldsymbol{\Sigma}^{jk}$. *There exists a constant* C_Σ *that does not depend on* (j,k) *such that, with probability tending to one,*

$$\big\|\widehat{\boldsymbol{\Sigma}}^{jk}(\widehat{\boldsymbol{\beta}}_{j\vee k})-\boldsymbol{\Sigma}^{jk}\big\|_\infty\le C_\Sigma s^*\lambda\log^2 d \quad \textit{for } 1\le j<k\le d.$$

Proof See §E.2.1 for a detailed proof. ■

In the rest of the proof, we will omit the superscripts in both $\widehat{\boldsymbol{\Sigma}}^{jk}$ and $\boldsymbol{\Sigma}^{jk}$ for notational simplicity. By Lemma 25,

$$I_1\le\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_\infty\le\mathcal{O}_{\mathbb{P}}\big(s^*\lambda\log^2 d\big). \tag{94}$$

By triangle inequality, we have the following inequality for I_2 :

$$I_{2j}\le\underbrace{\big|(\widehat{\mathbf{w}}_{j,k}-\mathbf{w}^*_{j,k})^T\big(\widehat{\boldsymbol{\Sigma}}_{j\setminus k,jk}-\boldsymbol{\Sigma}_{j\setminus k,jk}\big)\big|}_{I_{21}}+\underbrace{\big|(\widehat{\mathbf{w}}_{j,k}-\mathbf{w}^*_{j,k})^T\boldsymbol{\Sigma}_{j\setminus k,jk}\big|}_{I_{22}}+\underbrace{\big|\mathbf{w}^{*T}_{j,k}\big(\widehat{\boldsymbol{\Sigma}}_{j\setminus k,jk}-\boldsymbol{\Sigma}_{j\setminus k,jk}\big)\big|}_{I_{23}}.$$

By Hölder's inequality, Lemma 25 and the estimation error of $\widehat{\mathbf{w}}_{j,k}$, we obtain an upper-bound for I_{21} as follows:

$$I_{21}\le\|\widehat{\mathbf{w}}_{j,k}-\mathbf{w}^*_{j,k}\|_1\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_\infty=\mathcal{O}_{\mathbb{P}}\big(s^*s_0^\star\lambda_D\lambda\log^2 d\big). \tag{95}$$

Similarly, for I_{22}, Hölder's inequality implies that

$$I_{22}\le\|\widehat{\mathbf{w}}_{j,k}-\mathbf{w}^*_{j,k}\|_1\|\boldsymbol{\Sigma}\|_\infty=\mathcal{O}_{\mathbb{P}}\big(s_0^\star\lambda_D D\big), \tag{96}$$

where the constant D appears in (84). For I_{23}, by Hölder's inequality and 25 we obtain

$$I_{23}\le\|\mathbf{w}^*_{j,k}\|_1\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_\infty=\mathcal{O}_{\mathbb{P}}\big(w_0 s^*\lambda\log^2 d\big). \tag{97}$$

Combining (95), (96) and (97) we have

$$I_{2j}\lesssim(w_0+s_0^\star\lambda_D)s^*\lambda\log^2 d+s_0^\star\lambda_D. \tag{98}$$

For I_{3j}, by triangle inequality we have

$$I_{3j}\le\underbrace{\big|\widehat{\mathbf{w}}^T_{j,k}\big(\widehat{\boldsymbol{\Sigma}}_{j\setminus k,j\setminus k}-\boldsymbol{\Sigma}_{j\setminus k,j\setminus k}\big)\widehat{\mathbf{w}}_{j,k}\big|}_{I_{31}}+\underbrace{\big|\widehat{\mathbf{w}}^T_{j,k}\boldsymbol{\Sigma}_{j\setminus k,j\setminus k}\widehat{\mathbf{w}}_{j,k}-\mathbf{w}^{*T}_{j,k}\boldsymbol{\Sigma}_{j\setminus k,j\setminus k}\mathbf{w}^*_{j,k}\big|}_{I_{32}}.$$

For term I_{31}, Hölder's inequality and the optimality of $\widehat{\mathbf{w}}$ implies that

$$I_{31}\le\|\widehat{\mathbf{w}}_{j,k}\|_1^2\|\widehat{\boldsymbol{\Sigma}}_{j\setminus k,j\setminus k}-\boldsymbol{\Sigma}_{j\setminus k,j\setminus k}\|_\infty\le C_\Sigma w_0^2 s^*\lambda\log^2 d. \tag{99}$$

For term I_{32}, Lemma 17 implies that

$$\begin{aligned}I_{32}&\le\|\boldsymbol{\Sigma}_{j\setminus k,j\setminus k}\|_\infty\|\widehat{\mathbf{w}}_{j,k}-\mathbf{w}^*_{j,k}\|_1^2+\|\boldsymbol{\Sigma}_{j\setminus k,j\setminus k}\mathbf{w}^*_{j,k}\|_\infty\|\widehat{\mathbf{w}}_{j,k}-\mathbf{w}^*_{j,k}\|_1\\&\le\big(D\omega_0 s_0^\star\lambda_D+Ds_0^{\star 2}\lambda_D^2\big),\end{aligned} \tag{100}$$

where we use Hölder's inequality $\|\boldsymbol{\Sigma}_{j\backslash k,j\backslash k}\mathbf{w}^*_{j,k}\|_\infty \le \|\mathbf{w}^*_{j,k}\|_1\|\boldsymbol{\Sigma}\|_\infty \le Dw_0$. By (99), (100) and $\lambda_D \gtrsim w_0 s^* \lambda \log^2 d$, we obtain

$$I_{3j} \lesssim w_0^2 s^* \lambda \log^2 d + \bigl(D\omega_0 s_0^\star \lambda_D + D s_0^{\star 2}\lambda_D^2\bigr). \tag{101}$$

Therefore combining (94), (98) and (101) we obtain $I_1 + I_{2j} + I_{3j} = o_{\mathbb{P}}(1)$. We can show similarly that $I_{2k} + I_{3k} = o_{\mathbb{P}}(1)$. Thus $\lim\limits_{n\to\infty} \max\limits_{j<k}\bigl|\widehat{\sigma}^2_{jk} - \sigma^2_{jk}\bigr| = 0$ with probability converging to one. ■

Appendix E. Proof of Technical Lemmas

Finally, we prove the technical lemmas in this appendix. Specifically, we prove the lemmas introduced to derive the auxiliary results.

E.1. Proof of Technical Lemmas in §C

In this subsection we prove the technical lemmas we use to prove the auxiliary results of estimation. These lemmas are standard for high-dimensional linear regression, but proving them for our logistic-type loss function needs nontrivial extensions.

E.1.1. Proof of Lemma 20

Proof Let I and J be two index sets with $I \cap J = \emptyset, |I| \le s, |J| \le k$, for any $\mathbf{u} \in \mathbb{R}^d$ with $\|\mathbf{u} - \mathbf{u}_0\|_2 \le r$ and any $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$, let $\boldsymbol{\theta} = \mathbf{v}_I + \alpha\mathbf{w}_J$ with some $\alpha \in \mathbb{R}$, then by definition, $\|\boldsymbol{\theta}\|_0 \le s + k$. For notational simplicity, we denote s-sparse eigenvalues $\rho_+\bigl(\mathbf{M}, \mathbf{u}_0; s, r\bigr)$ and $\rho_-\bigl(\mathbf{M}, \mathbf{u}_0; s, r\bigr)$ as $\rho_-(s)$ and $\rho_+(s)$ respectively. By definition, we have

$$\rho_-(s{+}k)\|\boldsymbol{\theta}\|_2^2 \le \boldsymbol{\theta}^T\mathbf{M}(\mathbf{u})\boldsymbol{\theta} = \underbrace{\mathbf{v}_I^T\mathbf{M}(\mathbf{u})\mathbf{v}_I}_{A1} {+} 2\alpha \underbrace{\mathbf{v}_I^T\mathbf{M}(\mathbf{u})\mathbf{w}_J}_{A2} {+}\alpha^2 \underbrace{\mathbf{w}_J^T\mathbf{M}(\mathbf{u})\mathbf{w}_J}_{A3}. \tag{102}$$

Since $\|\boldsymbol{\theta}\|_2^2 = \|\mathbf{v}_I\|_2^2 + \alpha^2\|\mathbf{w}_J\|_2^2$. Rearranging the terms in (102) we have

$$\bigl[A_3 {-} \rho_-(s{+}k)\|\mathbf{w}_J\|_2^2\bigr]\alpha^2 + 2A_2\alpha + \bigl[A_1 {-} \rho_-(s{+}k)\|\mathbf{v}_I\|_2^2\bigr] \ge 0 \text{ for all } \alpha \in \mathbb{R}. \tag{103}$$

Note that the left-hand side (103) is a univariate quadratic function in α, thus (103) implies that

$$\bigl[A_1 {-} \rho_-(s{+}k)\|\mathbf{v}_I\|_2^2\bigr]\bigl[A_3 {-} \rho_-(s{+}k)\|\mathbf{w}_J\|_2^2\bigr] \ge A_2^2. \tag{104}$$

Therefore by multiplying $4\|\mathbf{v}_I\|_2^\infty/(A_1^2\|\mathbf{w}_J\|_2^2)$ to both sides of (104) we have

$$\frac{4A_2^2\|\mathbf{v}_I\|_2^2}{A_1^2\bigl\|\mathbf{w}_J\bigr\|_2^2} \le \frac{4\|\mathbf{v}_I\|_2^2}{A_1\|\mathbf{w}_J\|_2^2}\left[\frac{A_1 {-} \rho_-(s{+}k)\bigl\|\mathbf{v}_I\bigr\|_2^2}{A_1}\right]\bigl[A_3 {-} \rho_-(s{+}k)\|\mathbf{w}_J\|_2^2\bigr]. \tag{105}$$

By the inequality of arithmetic and geometric means, we have

$$\frac{\rho_-(s{+}k)\|\mathbf{v}_I\|_2^2}{A_1}\left[\frac{A_1 {-} \rho_-(s{+}k)\|\mathbf{v}_I\|_2^2}{A_1}\right] \le \frac{1}{4}.$$

Then the right-hand side of (104) can be bounded by

$$\frac{4A_2^2\|\mathbf{v}_I\|_2^2}{A_1^2\|\mathbf{w}_J\|_2^2} \le \frac{A_3 - \rho_-(s{+}k)\|\mathbf{w}_J\|_2^2}{\rho_-(s{+}k)\|\mathbf{w}_J\|_2^2} \le \frac{\rho_+(k)}{\rho_-(s{+}k)} - 1,$$

where the last inequality follows from $A_3 \le \rho_+(k)\|\mathbf{w}_J\|_2^2$. Note that by the relationship between ℓ_2- and ℓ_∞ norm, we have $\|\mathbf{w}_J\|_2 \le \sqrt{k}\|\mathbf{w}_J\|_\infty$, which further implies that

$$\frac{\mathbf{v}_I^T\mathbf{M}(\mathbf{u})\mathbf{w}_J\|\mathbf{v}_I\|_2}{\mathbf{v}_I^T\mathbf{M}(\mathbf{u})\mathbf{v}_I\|\mathbf{w}_J\|_\infty} \le \frac{\sqrt{k}\mathbf{v}_I^T\mathbf{M}(\mathbf{u})\mathbf{w}_J\|\mathbf{v}_I\|_2}{\mathbf{v}_I^T\mathbf{M}(\mathbf{u})\mathbf{v}_I\|\mathbf{w}_J\|_2} = \frac{\sqrt{k}A_2\|\mathbf{v}_I\|_2}{A_1\|\mathbf{w}_J\|_2} \le \frac{\sqrt{k}}{2}\sqrt{\rho_+(k)/\rho_-(s+k) - 1}.$$

Taking supremum over $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$ finally yields Lemma 20. ■

E.1.2. Proof of lemma 21

Proof Under Assumption 4, for any $\boldsymbol{\beta}_j \in \mathbb{R}^{d-1}$ such that $\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^*\|_2 \le r$ and any $\mathbf{v} \in \mathbb{R}^{d-1}$ with $\|\mathbf{v}\|_0 \le 2s^*{+}\, 2k^*$, we denote $\nabla^2 L_j(\boldsymbol{\beta}_j) - \nabla^2 L_j(\boldsymbol{\beta}_j^*)$ and $\nabla^2 L_j(\boldsymbol{\beta}_j) - \mathbb{E}\big[\nabla^2 L_j(\boldsymbol{\beta}_j^*)\big]$ as $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$ respectively. Our goal is to show that both $|\mathbf{v}^T\boldsymbol{\Lambda}_1\mathbf{v}|$ and $|\mathbf{v}^T\boldsymbol{\Lambda}_2\mathbf{v}|$ are negligible. Hölder's inequality implies that $\big|\mathbf{v}^T\boldsymbol{\Lambda}_2\mathbf{v}\big| \le \|\mathbf{v}\|_1\|\boldsymbol{\Lambda}_2\mathbf{v}\|_\infty \le \|\mathbf{v}\|_1^2\|\boldsymbol{\Lambda}_2\|_\infty$. We use the following lemma to control $|\mathbf{v}^\top\boldsymbol{\Lambda}_1\mathbf{v}|$ and $\|\boldsymbol{\Lambda}_2\|_\infty$.

Lemma 26 *We denote $s^* = \max_{j\in[d]}\|\boldsymbol{\beta}_j^*\|_0$. Let $r_1(s^*, n, d) > 0$ be a real number depending on s^*, n, and d that satisfy $\lim_{n\to\infty} r_1(s^*, n, d)\log^2 d = 0$. We define $\mathbb{B}_j(r_1) := \big\{\boldsymbol{\beta}_j \in \mathbb{R}^{d-1} \colon \big\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^*\big\|_1 \le r_1(s^*, n, d)\big\}$ as the ℓ_1-ball centered at $\boldsymbol{\beta}_j^*$ with radius $r_1(s^*, n, d)$. Under Assumptions 2 and 4, there exist absolute constants $C_h, C_r > 0$ such that, with probability at least $1 - (2d)^{-1}$, for all $j \in [d]$, $\boldsymbol{\beta}_j \in \mathbb{B}_j(r_1)$ and $\mathbf{v} \in \mathbb{R}^d$, it holds that,*

$$\big\|\nabla^2 L_j(\boldsymbol{\beta}_j^*) - \mathbb{E}\big[\nabla^2 L_j(\boldsymbol{\beta}_j^*)\big]\big\|_\infty \le C_h\sqrt{\log d/n}, \tag{106}$$

$$\big\|\nabla^2 L_j(\boldsymbol{\beta}_j) - \nabla^2 L_j(\boldsymbol{\beta}_j^*)\big\|_\infty \le C_r r_1(s^*, n, d)\cdot\log^2 d, \tag{107}$$

$$\big|\mathbf{v}^T\big[\nabla^2 L_j(\boldsymbol{\beta}_j) - \nabla^2 L_j(\boldsymbol{\beta}_j^*)\mathbf{v}\big| \le C_r r_1(s^*, n, d)\cdot\|\mathbf{v}\|_2^2. \tag{108}$$

Proof See §E.3 for a detailed proof. ■

Lemma 26 implies that $\|\boldsymbol{\Lambda}_2\|_\infty \le C_h\sqrt{\log d/n}$ with probability at least $1 - (2d)^{-1}$. By the relation between ℓ_1- and ℓ_2-norms, we have

$$\big|\mathbf{v}^T\boldsymbol{\Lambda}_2\mathbf{v}\big| \le (2s^*{+}\, 2k^*)\|\mathbf{v}\|_2^2\|\boldsymbol{\Lambda}\|_\infty \le (2s^*{+}\, 2k^*)C_h\sqrt{\log d/n}.$$

Moreover, setting $r = C_\rho s^*\sqrt{\log d/n}$ with $C_\rho \ge 33\rho_*^{-1}$, we have

$$|\mathbf{v}^T\boldsymbol{\Lambda}_1\mathbf{v}| \le C_r C_\rho\|\mathbf{v}\|_1^2 \le C_r C_\rho(2s^*{+}\, 2k^*)\sqrt{\log d/n}.$$

By Assumption 4, if n is large enough such that $(2s^*{+}\, 2k^*)(C_r C_\rho + C_h)\sqrt{\log d/n}] \le 0.05\rho_*$, then we have

$$0.95\rho_* \le \rho_{j-}^*(2s^*{+}\, 2k^*) - 0.05\rho_* \le \rho_{j-}(2s^*{+}\, 2k^*) < \rho_{j+}(k^*) \le \rho_{j+}^*(k^*) + 0.05\rho_*,$$

where we denote the s-sparse eigenvalues $\rho_-\big(\nabla^2 L_j, \boldsymbol{\beta}_j^*; s, r\big)$ and $\rho_+\big(\nabla^2 L_j, \boldsymbol{\beta}_j^*; s, r\big)$ as $\rho_{j-}(s)$ and $\rho_{j+}(s)$ respectively. Under Assumption 4, $\rho_{j+}^*(k^*)/\rho_{j-}^*(2s^*+2k^*) \le 1 + 0.2k^*/s^*$ and $k^* \ge 2s^*$, simple computation yields that

$$\frac{\rho_{j+}(k^*)}{\rho_{j-}(2s^*+2k^*)} \le \frac{\rho_{j+}^*(k^*) + 0.05\rho_*}{\rho_{j-}^*(2s^*+2k^*) - 0.05\rho_*} \le \frac{\rho_{j+}^*(k^*) + 0.05\rho_{j-}^*(2s^*+2k^*)}{0.95\rho_{j-}^*(2s^*+2k^*)} \le 1 + 0.27k^*/s^*.$$

Thus, we conclude the proof of Lemma 26. ■

E.1.3. Proof of Lemma 22

Proof For $\mathbf{v} = (v_1, \ldots, v_d)^T \in \mathbb{R}^d$, without loss of generality, we assume that $F^c = [s_1]$ where $s_1 = |F^c| \le s$. In addition, we assume that when $j > s_1$, v_j is arranged in descending order of $|v_j|$. That is, we rearrange the components of $\mathbf{v}$ such that $|v_j| \ge |v_{j+1}|$ for all $j \ge s_1$. Let $J_0 = [s_1]$ and $J_i = \{s_1 + (i-1)k + 1, \ldots, \min(s_1 + ik, d)\}$. By definition, we have $J = J_1$ and $I = J_0 \cup J_1$. Moreover, we have $\|\mathbf{v}_{J_i}\|_\infty \le \|\mathbf{v}_{J_{i-1}}\|_1/k$ when $i \ge 2$ because by the definition of J_i, we have $\sum_{i\ge 2} \|\mathbf{v}_{J_i}\|_\infty \le \|\mathbf{v}_F\|_1/k$. Note that by the definition of index sets I and J_i, $|J_i| \le k$ and $|I| = k + s_1 \le k + s$. We denote the restricted correlation coefficients $\pi(\mathbf{M}, \mathbf{u}_0; s, k, r)$ as $\pi(s, k)$, then by the definition of $\pi(s+k, k)$ we have

$$\big|\mathbf{v}_I^T \mathbf{M}(\mathbf{u}) \mathbf{v}_{J_i}\big| \le \pi(s+k, k)\big[\mathbf{v}_I^T \mathbf{M}(\mathbf{u}) \mathbf{v}_I\big] \|\mathbf{v}_{J_i}\|_\infty / \|\mathbf{v}_I\|_2.$$

Thus we have the following upper bound for $\big|\mathbf{v}_I^T \mathbf{M}(\mathbf{u}) \mathbf{v}_{I^c}\big|$:

$$\begin{aligned}\big|\mathbf{v}_I^T \mathbf{M}(\mathbf{u}) \mathbf{v}_{I^c}\big| &\le \sum_{i\ge 2} \big|\mathbf{v}_I^T \mathbf{M}(\mathbf{u}) \mathbf{v}_{J_i}\big| \le \pi(s+k, k)\|\mathbf{v}_I\|_2^{-1}\big[\mathbf{v}_I^T \mathbf{M}(\mathbf{u}) \mathbf{v}_I\big] \sum_{i\ge 2} \|\mathbf{v}_{J_i}\|_\infty \\ &\le \pi(s+k, k)\|\mathbf{v}_I\|_2^{-1}\big[\mathbf{v}_I^T \mathbf{M}(\mathbf{u}) \mathbf{v}_I\big] \|\mathbf{v}_F\|_1/k. \end{aligned} \tag{109}$$

Because $\mathbf{v}^T \mathbf{M}(\mathbf{u}) \mathbf{v} \ge \mathbf{v}_I^T \mathbf{M}(\mathbf{u}) \mathbf{v}_I + 2\mathbf{v}_I^T \mathbf{M}(\mathbf{u}) \mathbf{v}_{I^c}$, by (109) we have

$$\begin{aligned}\mathbf{v}^T \mathbf{M}(\mathbf{u}) \mathbf{v} &\ge \mathbf{v}_I^T \mathbf{M}(\mathbf{u}) \mathbf{v}_I - 2\pi(s+k, k)\|\mathbf{v}_I\|_2^{-1}\big[\mathbf{v}_I^T \mathbf{M}(\mathbf{u}) \mathbf{v}_I\big] \|\mathbf{v}_F\|_1/k \\ &= \big[\mathbf{v}_I^T \mathbf{M}(\mathbf{u}) \mathbf{v}_I\big]\big[1 - 2\pi(s+k, k)\|\mathbf{v}_I\|_2^{-1}\|\mathbf{v}_F\|_1/k\big].\end{aligned}$$

Thus we can bound the right-hand side of the last formula using the sparse eigenvalue condition

$$\mathbf{v}^T \mathbf{M}(\mathbf{u}) \mathbf{v} \ge \rho_-(s+k)\big[1 - 2\pi(s+k, k)k^{-1}\|\mathbf{v}_I\|_2^{-1}\|\mathbf{v}_F\|_1\big]\|\mathbf{v}_I\|_2^2, \tag{110}$$

where we denote s-sparse eigenvalue $\rho_-(\mathbf{M}, \mathbf{u}_0; s, r)$ as $\rho_-(s+k)$ for the simplicity of notations. Inequality (110) concludes the proof of Lemma 22. ■

E.1.4. Proof of Lemma 23

Proof Let $F(t) = L_j\big(\boldsymbol{\beta}(t)\big) - L_j(\boldsymbol{\beta}_1) - \big\langle \nabla L_j(\boldsymbol{\beta}_1), \boldsymbol{\beta}(t) - \boldsymbol{\beta}_1\big\rangle$. Since the derivative of $L_j\big(\boldsymbol{\beta}(t)\big)$ with respect to t is $\big\langle \nabla L_j\big(\boldsymbol{\beta}(t)\big), \boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\big\rangle$, the derivative of F is given by

$$F'(t) = \big\langle \nabla L_j\big(\boldsymbol{\beta}(t)\big) - \nabla L_j(\boldsymbol{\beta}_1), \boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\big\rangle.$$

Therefore the Bregman divergence $D_j\big(\boldsymbol{\beta}(t), \boldsymbol{\beta}_1\big)$ can be written as

$$D_j\big(\boldsymbol{\beta}(t), \boldsymbol{\beta}_1\big) = \big\langle \nabla L_j[\boldsymbol{\beta}(t)] - \nabla L_j(\boldsymbol{\beta}_1), t(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)\big\rangle = tF'(t).$$

By definition, it is easy to see that $F'(1) = D_j(\boldsymbol{\beta}_2, \boldsymbol{\beta}_1)$. To derive Lemma 23, it suffices to show that $F(t)$ is convex, which implies that $F'(t)$ is non-decreasing and $D_j\big(\boldsymbol{\beta}(t), \boldsymbol{\beta}_1\big) = tF'(t) \leq tF'(1) = tD_j(\boldsymbol{\beta}_2, \boldsymbol{\beta}_1)$.

For $\forall t_1, t_2 \in \mathbb{R}_+, t_1 + t_2 = 1, x, y \in (0, 1)$, by the linearity of $\boldsymbol{\beta}(t)$, $\boldsymbol{\beta}(t_1x + t_2y) = t_1\boldsymbol{\beta}(x) + t_2\boldsymbol{\beta}(y)$. Then we have

$$\big\langle \nabla L_j(\boldsymbol{\beta}_1), \boldsymbol{\beta}(t_1x + t_2y) - \boldsymbol{\beta}_1\big\rangle = t_1\big\langle \nabla L_j(\boldsymbol{\beta}_1), \boldsymbol{\beta}(x) - \boldsymbol{\beta}_1\big\rangle + t_2\big\langle \nabla L_j(\boldsymbol{\beta}_1), \boldsymbol{\beta}(y) - \boldsymbol{\beta}_1\big\rangle. \quad (111)$$

In addition, by convexity of function $L_j(\cdot)$, we obtain

$$L_j\big(\boldsymbol{\beta}(t_1x + t_2y)\big) \leq t_1L_j\big(\boldsymbol{\beta}(x)\big) + t_2L_j\big(\boldsymbol{\beta}(y)\big). \quad (112)$$

Adding (111) and (112) we obtain

$$F\big(t_1x + t_2y\big) \leq t_1F(x) + t_2F(y).$$

Therefore $F(t)$ is convex, thus we have $D_j(\boldsymbol{\beta}(t), \boldsymbol{\beta}_1) \leq tD_j(\boldsymbol{\beta}_2, \boldsymbol{\beta}_1)$. ■

E.2. Proof of Technical Lemmas in §D

Now we prove the lemmas that supports the auxiliary inferential results. We first prove Lemma 25, which implies that the $\widehat{\sigma}^2_{jk}$ is a consistent estimator of the asymptotic variance of σ_{jk}.

E.2.1. Proof of lemma 25

Proof Recall that we denote $\boldsymbol{\beta}_{j\vee k} = (\beta_{jk}, \boldsymbol{\beta}_{j\backslash k}, \boldsymbol{\beta}_{k\backslash j})$ and $L_{jk}\big(\boldsymbol{\beta}_{j\vee k}\big) = L_j(\boldsymbol{\beta}_j) + L_k(\boldsymbol{\beta}_k)$. We denote the kernel function of the second-order U-statistic $\nabla L_{jk}\big(\boldsymbol{\beta}_{j\vee k}\big)$ as $\mathbf{h}^{jk}_{ii'}\big(\boldsymbol{\beta}_{j\vee k}\big)$ where the subscripts i, i' indicate that $\mathbf{h}^{jk}_{ii'}(\cdot)$ depends on $\boldsymbol{X}_i$ and $\boldsymbol{X}_{i'}$. We define $\mathbf{V}^{jk}_{ii'i''}\big(\boldsymbol{\beta}_{j\vee k}\big) := \mathbf{h}^{jk}_{ii'}\big(\boldsymbol{\beta}_{j\vee k}\big)\mathbf{h}^{jk}_{ii'}\big(\boldsymbol{\beta}_{j\vee k}\big)^T$. Then by definition, $\widehat{\boldsymbol{\Sigma}}^{jk}\big(\boldsymbol{\beta}_{j\backslash k}\big)$can be written as

$$\widehat{\boldsymbol{\Sigma}}^{jk}\big(\boldsymbol{\beta}_{j\vee k}\big) = \frac{1}{n(n-1)^2}\sum_{i=1}^{n}\sum_{i'\neq i, i''\neq i}\mathbf{V}^{jk}_{ii'i''}\big(\boldsymbol{\beta}_{j\vee k}\big).$$

Note that $\widehat{\boldsymbol{\Sigma}}^{jk}\big(\boldsymbol{\beta}_{j\vee k}\big) - \boldsymbol{\Sigma}^{jk} = \underbrace{\widehat{\boldsymbol{\Sigma}}^{jk}\big(\boldsymbol{\beta}_{j\vee k}\big) - \widehat{\boldsymbol{\Sigma}}^{jk}\big(\boldsymbol{\beta}^*_{j\vee k}\big)}_{I_1} + \underbrace{\widehat{\boldsymbol{\Sigma}}^{jk}\big(\boldsymbol{\beta}^*_{j\vee k}\big) - \boldsymbol{\Sigma}^{jk}}_{I_2}$.

We first consider I_2. For notational simplicity, we use $\mathbf{h}_{ii'}$ and $\mathbf{h}_{ii'|i}$ to denote $\mathbf{h}^{jk}_{ij}\big(\boldsymbol{\beta}^*_{j\vee k}\big)$ and $\mathbf{h}^{jk}_{ii'|i}\big(\boldsymbol{\beta}^*_{j\vee k}\big) := \mathbb{E}\big[\mathbf{h}^{jk}_{ij}\big(\boldsymbol{\beta}^*_{j\vee k}\big)\big|\boldsymbol{X}_i\big]$ respectively. As shown in §D.1, for $i \neq i' \neq i''$,

$$\mathbb{E}\big(\mathbf{h}_{ii'}\mathbf{h}^T_{ii''}\big) = \mathbb{E}\big(\mathbf{h}_{ii'}\mathbf{h}^T_{ii''}\big|\boldsymbol{X}_i\big) = \mathbb{E}\big(\mathbf{h}_{ii'|i}\mathbf{h}^T_{ii''|i}\big) = \boldsymbol{\Sigma}^{jk} \text{ and } \mathbb{E}\big(\mathbf{h}_{ij}\mathbf{h}^T_{ij}\big) = \boldsymbol{\Theta}^{jk},$$

we can write I_2 as

$$I_2 = \frac{n-2}{n-1}\underbrace{\left\{\binom{n}{3}^{-1}\sum_{i<i'<i''}\big[\mathbf{V}_{ii'i''} - \mathbb{E}(\mathbf{V}_{ii'i''})\big]\right\}}_{I_{21}} + \frac{1}{n-1}\underbrace{\left\{\binom{n}{2}^{-1}\sum_{i<i'}\big[\mathbf{V}_{ii'i'} - \mathbb{E}(\mathbf{V}_{ii'i'})\big]\right\}}_{I_{22}} + \frac{1}{n-1}\big(\mathbf{\Theta}^{jk} - \mathbf{\Sigma}^{jk}\big),$$

where we use $\mathbf{V}_{ii'i''}$ to denote $\mathbf{V}_{ii'i''}^{jk}(\boldsymbol{\beta}_{j\vee k}^{*})$. Observing that I_{21} is a centered third order U-statistic, for x large enough such that $x^4 \geq \big\|\mathbb{E}\big[\mathbf{V}_{ijk}(\boldsymbol{\beta}_{j\vee k}^{*})\big]\big\|_{\infty}$ and for any $(a,b),(c,d) \in \big\{(p,q)\colon p,q \in \{j,k\}\big\}$ we have

$$\begin{aligned}\mathbb{P}\big(\big[\mathbf{V}_{ii'i''}^{jk}(\boldsymbol{\beta}_{j\vee k})\big]_{ab,cd} > 2x^4\big) &\leq \mathbb{P}\big[(X_{ia} - X_{i'a})(X_{ib} - X_{i'b})(X_{ic} - X_{i''c})(X_{id} - X_{i''d}) > x^4\big] \\ &\leq 8\exp(2\kappa_m + \kappa_h)\exp(-x).\end{aligned}$$

Thus there exist constants c_1 and C_1 that does not depend on n or d or (j,k) such that for any $x \in \mathbb{R}$, any $i,i',i'' \in [n]$ and any $j,k \in [d]$,

$$\mathbb{P}\big([\mathbf{V}_{ii'i''}^{jk}(\boldsymbol{\beta}_{j\vee k}^{*})]_{ab,cd} > x\big) \leq C_1\exp(c_1 x^{1/4}). \tag{113}$$

This implies that there exists some generic constant C such that $\|\mathbf{V}_{ii'i''}^{jk}(\boldsymbol{\beta}_{j\vee k}^{*})\|_{\infty} \leq C\log^4 d$ for all $j,k \in [d]$ and $i,i' \in [n]$ with probability tending to one. Similar to the method we use in §E.3, we define $\mathcal{E} := \big\{\|\mathbf{V}_{ii'i''}^{jk}(\boldsymbol{\beta}_{j\vee k}^{*})\|_{\infty} \leq C\log^4 d, \forall i,i',i'' \in [n], j,k \in [d]\big\}$. By Bernstein's inequality for U-statistics (Lemma 19) with $b = C\log^4 d$ in (56), for some generic constants C, it holds with high probability that

$$\binom{n}{2}^{-1}\sum_{i<i'}\big[\mathbf{V}_{ii'i'} - \mathbb{E}(\mathbf{V}_{ii'i'}|\mathcal{E})\big] \leq C\sqrt{\log d/n}, \quad \forall j,k \in [d], i,i',i'' \in [n]. \tag{114}$$

Moreover, by (113), we have

$$\begin{aligned}&\mathbb{E}\big\{[\mathbf{V}_{ii'i'}(\boldsymbol{\beta}_{j\vee k}^{*})]_{ab,cd}|\mathcal{E}\big\} - \mathbb{E}\big\{[V_{ii'i''}(\boldsymbol{\beta}_{j\vee k}^{*})]_{ab,cd}\big\} \\ &\quad\leq \int_{C\log^4 d}^{\infty}\mathbb{P}\big\{\big|[\mathbf{V}_{ii'i''}^{jk}(\boldsymbol{\beta}_{j\vee k}^{*})]_{ab,cd}\big| > x\big\} \leq c_1\log^3 d\cdot\exp(-c_2\log d)\end{aligned} \tag{115}$$

for some absolute constant c_1 and c_2. Since (115) holds uniformly, we have

$$\binom{n}{2}^{-1}\sum_{i<i'}\big[\mathbb{E}(\mathbf{V}_{ii'i'}|\mathcal{E}) - \mathbb{E}(\mathbf{V}_{ii'i''})\big] \leq \log^3 d\cdot\exp(-c_2\log d) \lesssim \sqrt{\log d/n}. \tag{116}$$

Combining (114) and (116) we obtain that

$$\|I_{21}\|_{\infty} = \mathcal{O}_{\mathbb{P}}\big(\sqrt{\log d/n}\big) \text{ uniformly for } 1 \leq j < k \leq n. \tag{117}$$

For the second part I_{22}, noting that it is a U-statistic of order 2, because (113) also holds for $\mathbf{V}_{ii'i''}(\boldsymbol{\beta}_{j\vee k}^{*})$, applying the same technique, we have $\|I_{21}\|_{\infty} = \mathcal{O}_{\mathbb{P}}\big(\sqrt{\log d/n}\big)$ uniformly for $1 \le j < k \le n$. Combining with (117), we conclude that, for some absolute constant C, we have

$$\big\|\widehat{\mathbf{\Sigma}}^{jk}\big(\boldsymbol{\beta}^*_{j\vee k}\big) - \mathbf{\Sigma}^{jk}\big\|_\infty \le C\sqrt{\log d/n}, \quad \forall 1 \le j < k \le n. \tag{118}$$

Now we turn to I_1. For any $\boldsymbol{\beta}_j, \boldsymbol{\beta}_k \in \mathbb{R}^{d-1}$ such that $\|\boldsymbol{\beta}_j - \boldsymbol{\beta}^*_j\|_1 \le r(s^*, n, d)$ and $\|\boldsymbol{\beta}_k - \boldsymbol{\beta}^*_k\|_1 \le r(s^*, n, d)$, we denote $\omega^j_{ii'} := \exp\big[-(X_{ij} - X_{i'j})(\boldsymbol{\beta}_j - \boldsymbol{\beta}^*_j)^T(\boldsymbol{X}_{i\setminus j} - \boldsymbol{X}_{i'\setminus j})\big]$ and denote $\omega^k_{ii'}$ similarly. Recall that we denote $R^j_{ii'}(\boldsymbol{\beta}_j) = \exp\big[-(x_{ij} - x_{i'j})\boldsymbol{\beta}_j{}^T(\boldsymbol{x}_{i\setminus j} - \boldsymbol{x}_{i'\setminus j})\big]$. Hence by definition we have $R^j_{ii'}(\boldsymbol{\beta}_j) = \omega^j_{ii'}R^j_{ii'}(\boldsymbol{\beta}^*_j)$. As shown in §E.3, we have

$$\min\{1, \omega^j_{ii'}, \omega^k_{ii'}\}\mathbf{h}^{jk}_{ii'}(\boldsymbol{\beta}^*_{j\vee k}) \le \mathbf{h}^{jk}_{ii'}(\boldsymbol{\beta}_{j\vee k}) \le \max\{1, \omega^j_{ii'}, \omega^k_{ii'}\}\mathbf{h}^{jk}_{ii'}(\boldsymbol{\beta}^*_{j\vee k}), \tag{119}$$

where the inequality is taken elementwisely. We denote $b := \max_{i,i'\in[n];j\in[d]} r(s^*, n, d)\big\|(X_{ij} - X_{i'j})(\boldsymbol{X}_{i\setminus j} - \boldsymbol{X}_{i'\setminus j})\big\|_\infty$. Note that when $\|\boldsymbol{\beta}_j - \boldsymbol{\beta}^*_j\|_1 \le r(s^*, n, d)$ and $\|\boldsymbol{\beta}_k - \boldsymbol{\beta}^*_k\|_1 \le r(s^*, n, d)$, we have $\omega^j_{ii'}, \omega^k_{ii'} \in [\exp(-b), \exp(b)]$. Therefore by (119) and the definition of $V^{jk}_{ii'i''}\big(\boldsymbol{\beta}_{j\setminus k}\big)$, we obtain the following elementwise inequality

$$\exp(-2b)\mathbf{V}^{jk}_{ii'i''}\big(\boldsymbol{\beta}^*_{j\setminus k}\big) \le \mathbf{V}^{jk}_{ii'i''}\big(\boldsymbol{\beta}_{j\setminus k}\big) \le \exp(2b)\mathbf{V}^{jk}_{ii'i''}\big(\boldsymbol{\beta}^*_{j\setminus k}\big),$$

which implies that

$$\big\|\widehat{\mathbf{\Sigma}}^{jk}\big(\boldsymbol{\beta}_{j\vee k}\big) - \widehat{\mathbf{\Sigma}}^{jk}\big(\boldsymbol{\beta}^*_{j\vee k}\big)\big\|_\infty \le \max\big\{1 - \exp(-2b), \exp(2b) - 1\big\}\big\|\widehat{\mathbf{\Sigma}}^{jk}\big(\boldsymbol{\beta}^*_{j\vee k}\big)\big\|_\infty. \tag{120}$$

As we show in §E.3, $b \le Cr(s^*, n, d)\log^2 d$ with high probability for some absolute constant $C > 0$. Since $\lim_{n\to\infty} r(s^*, n, d)\log^2 d = 0$, by (120) we have

$$\big\|\widehat{\mathbf{\Sigma}}^{jk}\big(\boldsymbol{\beta}_{j\vee k}\big) - \widehat{\mathbf{\Sigma}}^{jk}\big(\boldsymbol{\beta}^*_{j\vee k}\big)\big\|_\infty \lesssim b\big\|\widehat{\mathbf{\Sigma}}^{jk}\big(\boldsymbol{\beta}^*_{j\vee k}\big)\big\|_\infty\big\|_\infty \le b\big\|\widehat{\mathbf{\Sigma}}^{jk}\big(\boldsymbol{\beta}^*_{j\vee k}\big) - \mathbf{\Sigma}^{jk}\big\|_\infty + b\|\mathbf{\Sigma}^{jk}\|_\infty.$$

Note that we show $\|I_2\|_\infty = \big\|\widehat{\mathbf{\Sigma}}^{jk}\big(\boldsymbol{\beta}^*_{j\vee k}\big) - \mathbf{\Sigma}^{jk}\big\|_\infty = \mathcal{O}_{\mathbb{P}}\big(\sqrt{\log d/n}\big)$, which converges to zero asymptotically. Thus we conclude that

$$\big\|\widehat{\mathbf{\Sigma}}^{jk}\big(\boldsymbol{\beta}_{j\vee k}\big) - \widehat{\mathbf{\Sigma}}^{jk}\big(\boldsymbol{\beta}^*_{j\vee k}\big)\big\|_\infty = \mathcal{O}_{\mathbb{P}}\big(r(s^*, n, d)\log^2 d\big). \tag{121}$$

Combining (118) and (121), we have the following error bound for $\widehat{\mathbf{\Sigma}}^{jk}\big(\boldsymbol{\beta}_{j\vee k}\big)$:

$$\big\|\widehat{\mathbf{\Sigma}}^{jk}\big(\boldsymbol{\beta}_{j\vee k}\big) - \mathbf{\Sigma}^{jk}\big\|_\infty = \mathcal{O}_{\mathbb{P}}\Big(r(s^*, n, d)\log^2 d + \sqrt{\log d/n}\Big) \text{ for all } (j, k). \tag{122}$$

Finally, by the fact that $\max_{j\in[d]}\|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}^*_j\|_1 \lesssim s^*\lambda$, we conclude the proof of Lemma 25 by setting $r = Cs^*\lambda$. ■

E.3. Proof of Lemma 26

Now we turn to the last unproven result, namely Lemma 26, which characterizes the perturbation of $\nabla^2 L_j(\boldsymbol{\beta}_j)$.

Proof Note that $\nabla^2 L_j(\boldsymbol{\beta}_j)$ is a second-order U-statistic. Hence $\nabla^2 L_j(\boldsymbol{\beta}_j) - \mathbb{E}\bigl[\nabla^2 L_j(\boldsymbol{\beta}_j)\bigr]$ is a centered U-statistic. We denote its kernel as $\mathbf{T}_{ii'}(\boldsymbol{\beta}_j)$, then

$$\nabla^2 L_j(\boldsymbol{\beta}_j) - \mathbb{E}\bigl[\nabla^2 L_j(\boldsymbol{\beta}_j)\bigr] = \frac{2}{n(n-1)} \sum_{i<i'} \mathbf{T}_{ii'}(\boldsymbol{\beta}_j).$$

Note that $\bigl\|\mathbb{E}[\mathbf{T}_{ii'}(\boldsymbol{\beta}_j)]\bigr\|_\infty$ is bounded for all $\boldsymbol{\beta}_j \in \mathbb{R}^{d-1}$ because

$$\max_{\mathbf{u}\in\mathbb{R}^{d-1}} \bigl\|\mathbb{E}[T_{ii'}(\boldsymbol{\beta}_j)]\bigr\|_\infty \lesssim \max_{j\in[d]} \mathbb{E}|X_{ij} - X_{i'j}|^4 \lesssim \max_{j\in[d],i\in[n]} \mathbb{E}|X_{ij}|^4 \leq \int_0^\infty c\exp(-t^{1/4})dt = 24c,$$

where $c = 2\exp(\kappa_m + \kappa_h/2)$. Here the last inequality follows from (14). Let $\nabla^2_{jk,j\ell} L_j(\boldsymbol{\beta}_j) = \partial^2 L_j(\boldsymbol{\beta}_j)/\bigl(\partial\beta_{jk}\partial\beta_{j\ell}\bigr)$ and let $\bigl[\boldsymbol{T}_{ii'}(\boldsymbol{\beta}_j)\bigr]_{k\ell}$ be the corresponding kernel function. That is, $\nabla^2_{jk,j\ell} L_j(\boldsymbol{\beta}_j) = \binom{n}{2}^{-1}\sum_{i<i'}\bigl[\boldsymbol{T}_{ii'}(\boldsymbol{\beta}_j)\bigr]_{k\ell}$. For $x > 0$ such that $x^4 > 24c$ and $k, \ell \neq j$, we have

$$\begin{aligned}\mathbb{P}\bigl\{\bigl|[\mathbf{T}_{ii'}(\boldsymbol{\beta}_j^*)]_{k\ell}\bigr| > 2x^4\bigr\} &\leq \mathbb{P}\bigl[(X_{ij} - X_{i'j})^2(X_{ik} - X_{i'k})(X_{i\ell} - X_{i'\ell}) > x^4\bigr] \\ &\leq \mathbb{P}\bigl(|X_{ij} - X_{i'j}| > x\bigr) + \mathbb{P}\bigl(|X_{ik} - X_{i'k}| > x\bigr) + \mathbb{P}\bigl(|X_{i\ell} - X_{i'\ell}| > x\bigr). \qquad (123)\end{aligned}$$

As a direct implication of Assumption 2, we have $\mathbb{P}\bigl(|X_{ij} - X_{ij}| > x\bigr) \leq 2\exp(2\kappa_m + \kappa_k)\exp(-x)$ for all $j \in [d]$. Then we can bound the right-hand side of (123) by

$$\mathbb{P}\bigl\{\bigl|[\mathbf{T}_{ii'}(\boldsymbol{\beta}_j^*)]_{k\ell}\bigr| > 2x^4\bigr\} \leq 6\exp(2\kappa_m + \kappa_h)\exp(-x) \quad \text{when } x^4 > 48\exp(\kappa_m + \kappa_h/2).$$

Letting $C_T = \max\Bigl\{6\exp(2\kappa_m + \kappa_h), \exp\bigl\{[48\exp(\kappa_m + \kappa_h/2)]^{1/4}\bigr\}\Bigr\}$, it holds that

$$\mathbb{P}\bigl\{\bigl|[\mathbf{T}_{ii'}(\boldsymbol{\beta}_j^*)]_{k\ell}\bigr| > x\bigr\} \leq C_T\exp(-2^{-1/4}x^{1/4}) \quad \text{for all } x > 0. \qquad (124)$$

Thus by a union bound, we conclude that there exists some generic constant C such that $\|\mathbf{T}_{ii'}(\boldsymbol{\beta}_j^*)\|_\infty \leq C\log^4 d$ for all $j \in [d]$ and $i, i' \in [n]$ with probability at least $1 - (8d)^{-1}$. We define an event $\mathcal{E} := \bigl\{\|\mathbf{T}_{ii'}(\boldsymbol{\beta}_j^*)\|_\infty \leq C\log^4 d, \forall i, i' \in [n], j \in [d]\bigr\}$. By (124), it is easy to see that $\mathbf{T}_{ii'}(\boldsymbol{\beta}_j^*)$ is ℓ_2-integrable. By Bernstein's inequality for U-statistics (Lemma 19) with $b = C\log^4 d$ in (56), for some generic constants C_1 and C_2, we obtain that

$$\mathbb{P}\Bigl(\nabla^2 L_j(\boldsymbol{\beta}_j) - \mathbb{E}_1\bigl[\nabla^2 L_j(\boldsymbol{\beta}_j)\bigr] > t\bigl|\mathcal{E}\Bigr) \leq 4\exp\bigl[-nt^2/(C_1 + C_2\log^4 \cdot t)\bigr], \quad \forall j \in [d]. \qquad (125)$$

Here we use $\mathbb{E}_1\bigl[\nabla^2 L_j(\boldsymbol{\beta}_j)\bigr]$ to denote $\mathbb{E}\bigl[\nabla^2 L_j(\boldsymbol{\beta}_j)\bigl|\mathcal{E}\bigr]$. Thus under Assumption 4 we obtain that, conditioning on event $\mathcal{E}$,

$$\bigl\|\nabla^2 L_j(\boldsymbol{\beta}_j) - \mathbb{E}_1\bigl[\nabla^2 L_j(\boldsymbol{\beta}_j)\bigr]\bigr\|_\infty \leq C\sqrt{\log d/n}, \quad \forall j \in [d] \qquad (126)$$

with probability at least $1 - (8d)^{-1}$. Moreover, by (124) we obtain that

$$\mathbb{E}\bigl\{[\mathbf{T}_{ii'}(\boldsymbol{\beta}_j^*)]_{k\ell}\bigl|\mathcal{E}\bigr\} - \mathbb{E}\bigl\{[\mathbf{T}_{ii'}(\boldsymbol{\beta}_j^*)]_{k\ell}\bigr\} \leq \int_{C\log^4 d}^\infty \mathbb{P}\bigl\{\bigl|[\mathbf{T}_{ii'}(\boldsymbol{\beta}_j^*)]_{k\ell}\bigr| > x\bigr\} \leq c_1\log^3 d \cdot \exp(-c_2\log d)$$

for some absolute constant c_1 and c_2. Therefore we have

$$\bigl\|\mathbb{E}_1\bigl[\nabla^2 L_j(\boldsymbol{\beta}_j)\bigr] - \mathbb{E}\bigl[\nabla^2 L_j(\boldsymbol{\beta}_j)\bigr]\bigr\|_\infty \lesssim \log^3 d \cdot \exp(-c_2\log d) \lesssim \sqrt{\log d/n}. \qquad (127)$$

Combining (126) and (127) we show that, with probability at least $1-(4d)^{-1}$, $\big\|\nabla^2 L_j(\boldsymbol{\beta}_j^*) - \mathbb{E}[\nabla^2 L_j(\boldsymbol{\beta}_j^*)]\big\|_\infty \leq C_h\sqrt{\log d/n}$ for all $j \in [d]$.

For the second argument (107), let $\boldsymbol{\Delta} = \boldsymbol{\beta}_j - \boldsymbol{\beta}_j^*$ where $\boldsymbol{\beta}_j \in \mathbb{R}^{d-1}$ lies in the ℓ_1-ball centered at $\boldsymbol{\beta}_j^*$ with radius $r_1(s^*,n,d)$, that is, $\big\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^*\big\|_1 \leq r_1(s^*,n,d)$. By the independence between $\boldsymbol{X}_i$ and $\boldsymbol{X}_{i'}$, Assumption 2 implies that

$$\max\Big\{\log\mathbb{E}\big[\exp(X_{ij}-X_{i'j})\big], \log\mathbb{E}\big[\exp(X_{i'j}-X_{ij})\big]\Big\} \leq 2\kappa_m + \kappa_h,$$

which further implies that for any $x > 0$

$$\mathbb{P}\Big(\big|(X_{ij}-X_{i'j})\big| > x\Big) \leq 2\exp(2\kappa_m+\kappa_h)\exp(-x), \quad \forall j \in [d].$$

Hence for any $x > 0$ and $j,k \in [d]$, a union bound implies that

$$\begin{aligned}\mathbb{P}\big[\big|(X_{ij}-X_{i'j})(X_{ik}-X_{i'k})\big| > x^2\big] &\leq \mathbb{P}\big[\big|(X_{ij}-X_{i'j})\big| > x\big] + \mathbb{P}\big[\big|(X_{ik}-X_{i'k})\big| > x\big] \\ &\leq 4\exp(2\kappa_m+\kappa_h)\exp(-x). \end{aligned} \tag{128}$$

Taking a union bound over $1 \leq j < k \leq d$ and $1 \leq i < i' \leq n$ we obtain that

$$\mathbb{P}\Big[\max_{i,i'\in[n];j\in[d]} \big\|(X_{ij}-X_{i'j})(\boldsymbol{X}_{i\setminus j}-\boldsymbol{X}_{i'\setminus j})\big\|_\infty > x^2\Big] \lesssim n^2 d^2 \exp(-x).$$

If we denote $b := \max_{i,i'\in[n];j\in[d]} r_1(s^*,n,d)\big\|(X_{ij}-X_{i'j})(\boldsymbol{X}_{i\setminus j}-\boldsymbol{X}_{i'\setminus j})\big\|_\infty$, then we obtain that $b \leq Cr_1(s^*,n,d)\log^2 d$ with probability at least $1-(4d)^{-1}$ for some constant $C > 0$. Denoting $\omega_{ii'} := \exp\big\{-(X_{ij}-X_{i'j})\boldsymbol{\Delta}^T(\boldsymbol{X}_{i\setminus j}-\boldsymbol{X}_{i'\setminus j})\big\}$, by definition,

$$R_{ii'}^j(\boldsymbol{\beta}_j) = \exp\big\{-(X_{ij}-X_{i'j})(\boldsymbol{\Delta}+\boldsymbol{\beta}_j^*)^T(\boldsymbol{X}_{i\setminus j}-\boldsymbol{X}_{i'\setminus j})\big\} = \omega_{ii'}R_{ii'}^j(\boldsymbol{\beta}_j^*).$$

Thus we can write $\nabla^2 L_j(\boldsymbol{\beta}_j)$ as:

$$\nabla^2 L_j(\boldsymbol{\beta}_j) = \frac{2}{n(n-1)}\sum_{i<i'} \frac{R_{ii'}^j(\boldsymbol{\beta}^*)(X_{ij}-X_{i'j})^2(\boldsymbol{X}_{i\setminus j}-\boldsymbol{X}_{i'\setminus j})^{\otimes 2}}{\big(1+R_{ii'}^j(\boldsymbol{\beta}^*)\big)^2} \frac{\omega_{ii'}\big(1+R_{ii'}^j(\boldsymbol{\beta}^*)\big)^2}{\big(1+\omega_{ii'}R_{ii'}^j(\boldsymbol{\beta}^*)\big)^2}. \tag{129}$$

If $\omega_{ii'} \geq 1$, then $(\omega_{ii'})^{-2} \leq \big(1+R_{ii'}^j(\boldsymbol{\beta}^*)\big)^2/\big(1+\omega_{ii'}R_{ii'}^j(\boldsymbol{\beta}^*)\big)^2 \leq 1$; otherwise we have $1 \leq \big(1+R_{ii'}^j(\boldsymbol{\beta})\big)^2/\big(1+\omega_{ii'}R_{ii'}^j(\boldsymbol{\beta}^*)\big)^2 \leq (\omega_{ii'})^{-2}$. This observation implies

$$\min\big\{\omega_{ii'}, 1/\omega_{ii'}\big\} \leq \frac{\omega_{ii'}\big(1+R_{ii'}^j(\boldsymbol{\beta})\big)^2}{\big(1+\omega_{ii'}R_{ii'}^j(\boldsymbol{\beta}^*)\big)^2} \leq \max\big\{\omega_{ii'}, 1/\omega_{ii'}\big\}. \tag{130}$$

By the definition of $\omega_{ii'}$, Hölder's inequality implies that $\big|(X_{ij}-X_{i'j})\boldsymbol{\Delta}^T(\boldsymbol{X}_{i\setminus j}-\boldsymbol{X}_{i'\setminus j})\big| \leq b$, thus we have

$$\exp(-b) \leq \min\big\{\omega_{ii'}1/\omega_{ii'}\big\} \leq \max\big\{\omega_{ii'}, 1/\omega_{ii'}\big\} \leq \exp(b). \tag{131}$$

Combining (129),(130) and (131) we obtain

$$\exp(-b)\nabla^2 L_j(\boldsymbol{\beta}_j^*) \le \nabla^2 L_j(\boldsymbol{\beta}_j) \le \exp(b)\nabla^2 L_j(\boldsymbol{\beta}_j^*). \tag{132}$$

Then by (132), since $\lim_{n\to\infty} r_1(s^*, n, d)\log^2 d = 0$, we have

$$\big\|\nabla^2 L_j(\boldsymbol{\beta}_j) - \nabla^2 L_j(\boldsymbol{\beta}_j^*)\big\|_\infty \le \max\big\{1-\exp(-b), \exp(b)-1\big\}\big\|\nabla^2 L_j(\boldsymbol{\beta}_j^*)\big\|_\infty \lesssim b\big\|\nabla^2 L_j(\boldsymbol{\beta}_j^*)\big\|_\infty.$$

Notice that under Assumption 2, as shown in §D.1, we can assume that $\big\|\mathbb{E}\big[\nabla^2 L_j(\boldsymbol{\beta}_j^*)\big]\big\|_\infty \le D$ where D appears in (84). By triangle inequality,

$$\big\|\nabla^2 L_j(\boldsymbol{\beta}_j^*)\big\|_\infty \le \big\|\nabla^2 L_j(\boldsymbol{\beta}_j^*) - \mathbb{E}\big[\nabla^2 L_j(\boldsymbol{\beta}_j^*)\big]\big\|_\infty + \big\|\mathbb{E}\big[\nabla^2 L_j(\boldsymbol{\beta}_j^*)\big]\big\|_\infty \le D + C_h\sqrt{\log d/n} \le 2D$$

with probability at least $1-(4d)^{-1}$, where the last inequality follows from the fact that $\big(\log^9 d/n\big)^{1/2}$ tends to zero as n goes to infinity. Then we obtain that

$$\big\|\nabla^2 L_j(\boldsymbol{\beta}_j) - \nabla^2 L_j(\boldsymbol{\beta}_j^*)\big\|_\infty \le C_r r_1(s^*, n, d)\log^2 d$$

holds for some absolute constant $C_r > 0$ and uniformly for all $j \in [d]$ and $\boldsymbol{\beta}_j \in \mathbb{B}_j(r_1)$ with probability at least $1-(2d)^{-1}$.

Finally, for the last argument (108), for any $\mathbf{v} \in \mathbb{R}^{d-1}$, by (132) we have

$$\exp(-b)\mathbf{v}^T\nabla^2 L_j(\boldsymbol{\beta}_j^*)\mathbf{v} \le \mathbf{v}^T\nabla^2 L_j(\boldsymbol{\beta}_j)\mathbf{v} \le \exp(b)\mathbf{v}^T\nabla^2 L_j(\boldsymbol{\beta}_j^*)\mathbf{v}.$$

Thus we have

$$\big|\mathbf{v}^T\big[\nabla^2 L_j(\boldsymbol{\beta}_j) - \nabla^2 L_j(\boldsymbol{\beta}_j^*)\big]\mathbf{v}\big| \lesssim b\big|\mathbf{v}^T\nabla^2 L_j(\boldsymbol{\beta}_j^*)\mathbf{v}\big| \le b\|\mathbf{v}\|_1^2\big\|\nabla^2 L_j(\boldsymbol{\beta}_j^*)\big\|_\infty,$$

which implies (108). ■

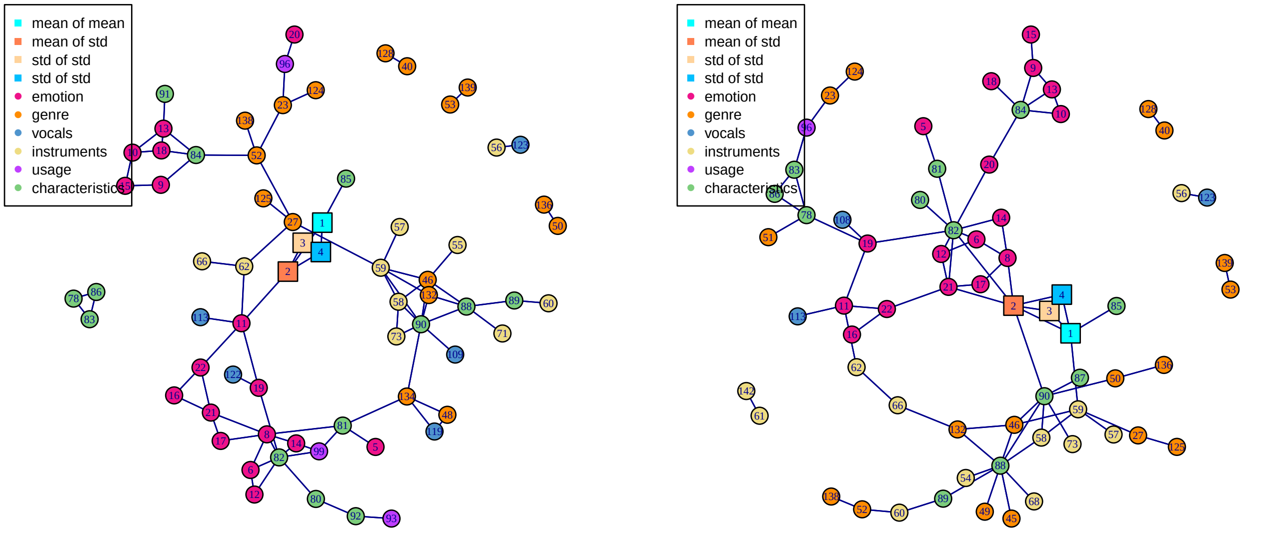
References

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*, 2012.

Miguel A Arcones. A Bernstein-type inequality for U-statistics and U-processes. *Statistics & probability letters*, 22(3):239–247, 1995.

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

Alexandre Belloni, Daniel Chen, Victor Chernozhukov, and Christian Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, 2012.



(a). The asymmetric score test based on $L_k(\boldsymbol{\beta}_k)$. (b). Inconsistent edges of the asymmetric score test

Figure 4: In (a) we plot estimated graph in the CAL500 dataset inferred by the asymmetric score test based on the loss function $L_k(\boldsymbol{\beta}_k)$ for testing $H_0\colon \beta^*_{jk} = 0$ for any $1 \leq j < k \leq d$. We plot the connected components of the estimated graph for illustration. Compared with Figure 3, we observe that the two asymmetric score tests yields different graphs. In (b) we plot the edges that appear in (a) and Figure 3-(a) but not in Figure 3-(b). In other words, we plot the inconsistent edges of these two asymmetric score tests that are discovered by the pairwise score test. Thus, by taking symmetry into consideration, the pairwise score test is able to correct such inconsistency.

Alexandre Belloni, Victor Chernozhukov, and Ying Wei. Honest confidence regions for a regression parameter in logistic regression with a large number of controls. *arXiv:1304.3969*, 2013.

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.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of LASSO and Dantzig selector. *The Annals of Statistics*, 37:1705–1732, 2009.

Jelena Bradic, Jianqing Fan, and Weiwei Wang. Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):325–349, 2011.

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 Candés, Terence Tao, et al. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

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

Kwun Chuen Gary Chan. Nuisance parameter elimination for proportional likelihood ratio models with nonignorable missingness and random truncation. *Biometrika*, 100(1):269–276, 2012.

Shizhe Chen, Daniela M Witten, and Ali Shojaie. Selection and estimation for mixed graphical models. *Biometrika*, 102(1):47–64, 2015.

Jie Cheng, Elizaveta Levina, Pei Wang, and Ji Zhu. A sparse Ising model with covariates. *Biometrics*, 70(4):943–953, 2014.

Jie Cheng, Tianxi Li, Elizaveta Levina, and Ji Zhu. High-dimensional mixed graphical models. *Journal of Computational and Graphical Statistics*, 26(2):367–378, 2017.

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

Guoqing Diao, Jing Ning, et al. Maximum likelihood estimation for semiparametric density ratio model. *The International Journal of Biostatistics*, 8(1):1–29, 2012.

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

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

David Edwards. *Introduction to Graphical Modelling.* Springer, 2000.

Sasha Epskamp. *Sampling Methods and Distribution Functions for the Ising Model*, 2015. R package.

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

Jianqing Fan, Lingzhou Xue, Hui Zou, et al. Strong oracle optimality of folded concave penalized estimation. *The Annals of Statistics*, 42(3):819–849, 2014.

Jianqing Fan, Han Liu, Yang Ning, and Hui Zou. High dimensional semiparametric latent graphical model for mixed data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):405–421, 2017.

Jianqing Fan, Han Liu, Qiang Sun, and Tong Zhang. I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *Annals of statistics*, 46 (2):814–841, 2018.

Bernd Fellinghauer, Peter Bühlmann, Martin Ryffel, Michael Von Rhein, and Jan D Reinhardt. Stable graphical model estimation with random forests for discrete, continuous, and mixed variables. *Computational Statistics & Data Analysis*, 64:132–152, 2013.

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 Rob Tibshirani. Applications of the LASSO and grouped LASSO to the estimation of sparse graphical models. Technical report, 2010.

J Ge, X Li, H Jiang, H Liu, T Zhang, M Wang, and T Zhao. *Picasso: A Sparse Learning Library for High Dimensional Data Analysis in R and Python*, 2017. R package.

Quanquan Gu, Yuan Cao, Yang Ning, and Han Liu. Local and global inference for high dimensional gaussian copula graphical models. *arXiv preprint arXiv:1502.02347*, 2015.

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

Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu. Graphical models for ordinal data. *Journal of Computational and Graphical Statistics*, 24(1):183–204, 2015.

Holger Höfling and Robert Tibshirani. Estimation of sparse binary pairwise markov networks using pseudo-likelihoods. *The Journal of Machine Learning Research*, 10:883–906, 2009.

Jana Janková and Sara van de Geer. Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics*, 9(1):1205–1229, 2015. doi: 10.1214/15-EJS1031.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

Clifford Lam and Jianqing Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of statistics*, 37(6B):4254–4278, 2009.

Steffen L Lauritzen. *Graphical Models*. Oxford University Press, 1996.

Jason D Lee and Trevor J Hastie. Learning the structure of mixed graphical models. *Journal of Computational and Graphical Statistics*, 24(1):230–253, 2015.

Jason D Lee, Dennis L Sun, Yuekai Sun, Jonathan E Taylor, et al. Exact post-selection inference, with application to the LASSO. *The Annals of Statistics*, 44(3):907–927, 2016.

Su-In Lee, Varun Ganapathi, and Daphne Koller. Efficient structure learning of Markov networks using ℓ_1-regularization. In *Advances in neural Information processing systems*, 2006.

Kung-Yee Liang and Jing Qin. Regression analysis under non-standard situations: a pairwise pseudolikelihood approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):773–786, 2000.

Han Liu, John Lafferty, and Larry Wasserman. The Nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *The Journal of Machine Learning Research*, 10:2295–2328, 2009.

Han Liu, Fang Han, Ming Yuan, John Lafferty, Larry Wasserman, et al. High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, 2012.

Weidong Liu et al. Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics*, 41(6):2948–2978, 2013.

Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, Robert Tibshirani, et al. A significance test for the LASSO. *The Annals of Statistics*, 42(2):413–468, 2014.

Po-Ling Loh and Martin J Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16:559–616, 2015.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the LASSO. *The Annals of Statistics*, 34(3):1436–1462, 2006.

Karthik Mohan, Palma London, Maryam Fazel, Daniela Witten, and Su-In Lee. Node-based learning of multiple Gaussian graphical models. *The Journal of Machine Learning Research*, 15(1):445–488, 2014.

Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 11 2012.

Matey Neykov, Yang Ning, Jun S Liu, Han Liu, et al. A unified theory of confidence regions and testing for high-dimensional estimating equations. *Statistical Science*, 33(3):427–443, 2018.

Yang Ning and Han Liu. High-dimensional semiparametric bigraphical models. *Biometrika*, 100(3):655–670, 2013.

Yang Ning, Han Liu, et al. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1):158–195, 2017a.

Yang Ning, Tianqi Zhao, Han Liu, et al. A likelihood ratio framework for high-dimensional semiparametric regression. *The Annals of Statistics*, 45(6):2299–2327, 2017b.

Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated Gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

Pradeep Ravikumar, Martin J Wainwright, John D Lafferty, et al. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3): 1287–1319, 2010.

Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, Bin Yu, et al. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Zhao Ren, Tingni Sun, Cun-Hui Zhang, Harrison H Zhou, et al. Asymptotic normality and optimalities in estimation of large gaussian graphical models. *The Annals of Statistics*, 43(3):991–1026, 2015.

Adam J Rothman, Peter J Bickel, Elizaveta Levina, and Ji Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

Xiaotong Shen, Wei Pan, and Yunzhang Zhu. Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107(497):223–232, 2012.

Tingni Sun and Cun-Hui Zhang. Sparse matrix inversion with scaled LASSO. *The Journal of Machine Learning Research*, 14(1):3385–3418, 2013.

Kean Ming Tan, Palma London, Karthik Mohan, Su-In Lee, Maryam Fazel, and Daniela Witten. Learning graphical models with hubs. *The Journal of Machine Learning Research*, 15(1):3297–3331, 2014.

Kean Ming Tan, Yang Ning, Daniela M Witten, and Han Liu. Replicates in high dimensions, with applications to latent variable graphical models. *Biometrika*, 103(4):761–777, 2016.

Robert Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

Grigorios Tsoumakas, Eleftherios Spyromitros-Xioufis, Jozef Vilcek, and Ioannis Vlahavas. Mulan: A java library for multi-label learning. *The Journal of Machine Learning Research*, 12:2411–2414, 2011.

Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet. Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):467–476, 2008.

George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *Speech and Audio Processing, IEEE transactions on*, 10(5):293–302, 2002.

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42 (3):1166–1202, 06 2014. doi: 10.1214/14-AOS1221.

Aad W Van der Vaart. *Asymptotic Statistics*. Cambridge university press, 2000.

Arend Voorman, Ali Shojaie, and Daniela Witten. Graph estimation with joint additive models. *Biometrika*, 101(1):85–101, 2014.

Zhaoran Wang, Han Liu, and Tong Zhang. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *The Annals of Statistics*, 42(6): 2164–2201, 12 2014. doi: 10.1214/14-AOS1238.

Joe Whittaker. *Graphical Models in Applied Multivariate Statistics*. Wiley Publishing, 2009.

Lin Xiao and Tong Zhang. A proximal-gradient homotopy method for the sparse least-squares problem. *SIAM Journal on Optimization*, 23(2):1062–1091, 2013.

Lingzhou Xue, Hui Zou, Tianxi Cai, et al. Nonconcave penalized composite conditional likelihood estimation of sparse Ising models. *The Annals of Statistics*, 40(3):1403–1429, 2012a.

Lingzhou Xue, Hui Zou, et al. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics*, 40(5):2541–2571, 2012b.

Eunho Yang, Pradeep Ravikumar, Genevera I Allen, and Zhandong Liu. On graphical models via univariate exponential family distributions. *arXiv preprint arXiv:1301.4183*, 2013a.

Eunho Yang, Pradeep Ravikumar, Genevera I Allen, and Zhandong Liu. On Poisson graphical models. *Advances in Neural Information Processing Systems*, pages 1718–1726, 2013b.

Eunho Yang, Pradeep Ravikumar, Genevera I Allen, Yulia Baker, Ying-Wooi Wan, and Zhandong Liu. A general framework for mixed graphical models. *arXiv preprint arXiv:1411.0288*, 2014.

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 the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

Cun-Hui Zhang and Jian Huang. The sparsity and bias of the LASSO selection in high-dimensional linear regression. *The Annals of Statistics*, pages 1567–1594, 2008.

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

Tong Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *The Journal of Machine Learning Research*, 11:1081–1107, 2010.

Tong Zhang et al. Multi-stage convex relaxation for feature selection. *Bernoulli*, 19(5B): 2277–2293, 2013.

Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics*, 36(4):1509, 2008.