On Semiparametric Exponential Family Graphical Models

Zhuoran Yang∗ Yang Ning† Han Liu‡

December 15th, 2014

Abstract

We study a general class of semiparametric exponential family graphical models for the analysis of high dimensional mixed data. Unlike existing methods that require the nodewise conditional distributions to be fully specified by known generalized linear models, we allow the nodewise conditional distributions to be generalized linear models with unspecified base measures. For graph inference, we propose a new procedure named NOSI which is invariant to arbitrary base measures and attains optimal rates of convergence for parameter estimation. We also provide theoretical guarantees for both local and global inference of the true graph. Thorough numerical simulations and a real data example are provided to back up our results.

1 Introduction

We consider the problem of inferring semiparametric exponential family graphical model. Let $G = (V, E)$ be an undirected graph with node set $V = \{1, 2, \ldots, d\}$ and edge set $E \subset \{(j, k): 1 \leq j < k \leq d\}$, a semiparametric exponential graphical model specifies the joint distribution for a $d$-dimensional random vector $X = (X_1, \ldots, X_d)^T$ such that for each $j \in V$, the conditional distribution of $X_j$ given $X_{\setminus j} := (X_k; k \neq j)^T$ is of the form

$$p(x_j \mid x_{\setminus j}) = \exp\left(\eta_j(x_{\setminus j}) \cdot x_j + f_j(x_j) - b_j(\eta_j; f_j)\right),$$

(1.1)

where $\eta_j(x_{\setminus j}) = \alpha_j + \sum_{k \neq j} \beta_{jk} x_k$ is the canonical parameter, $f_j(\cdot)$ is the base measure and $b_j(\cdot; \cdot)$ is the log-partition function. 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, it is easy to see that $\beta_{jk} \neq 0$ if and only if $X_j$ and $X_k$ are conditionally independent given $X_{\setminus \{j,k\}}$. Therefore, we set $(j, k) \in E$ if and only

∗Department of Mathematics, Tsinghua University and Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544. This work is carried out during an internship in the SMiLe (Statistical Machine Learning) lab at Princeton University. e-mail: zy6@princeton.edu.
†Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; e-mail: yning@princeton.edu.
‡Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; e-mail: hanliu@princeton.edu.
if \( \beta_{jk} \neq 0 \). The graph \( G \) thus characterizes the conditional independence relationship of the high dimensional distribution of \( X \).

We aim to infer the graph \( G \) based on independent observations \( X_1, \ldots, X_n \). For this, we propose the NOSI (Nuisance-Free Optimal Estimation and Simultaneous Inference) procedure which is invariant to arbitrary base measures \( \{ f_j \}_{j=1, \ldots, d} \) under model (1.1). More specifically, let \( \beta_j := (\beta_{jk}; k \neq j)^T \) be the parameters associated with node \( j \), NOSI estimates \( \beta_j \) by minimizing a loss function proposed by Ning and Liu (2014) with a nonconvex penalty function. To solve this nonconvex optimization, we use an iterative algorithm named Adaptive Multi-stage Convex Relaxation originated in Zhang (2010). Once a graph \( \hat{G} \) is estimated from the data, another important inferential task is to assess the uncertainty of \( \hat{G} \). More specifically, we are interested in two problems: local inference and global inference. The local inference aims to test whether an edge \((j,k)\) in \( \hat{G} \) belongs to the true graph \( G^* \). Based on the local inference, the global inference aims to test whether the point estimate \( \hat{G} \) is a super-graph of \( G^* \). We have developed a pairwise score test to conduct these two inferential tasks. For local inference, we test the null hypothesis \( H_0: \beta_{jk} = 0 \). For global inference, we propose test whether a given graph is a super-graph of the true graph with specified significance level. That is, for any given graph \( G \), we aim to test the null hypothesis \( H_0: G^* \subset G \). To derive the limiting distribution of the test statistic, we exploit the Gaussian approximation method in Chernozhukov et al. (2013).

1.1 Related Works

There is a huge literate on estimating undirected graphical models (Dempster, 1972; Lauritzen, 1996; Edwards, 2000; Whittaker, 2009; Koller and Friedman, 2009). In the following, we summarize different methods according to different data types: continuous data, discrete data, and mixed data.

For continuous data, the most commonly used method is Gaussian graphical models, which assume the random variables are jointly Gaussian with precision matrix \( \Theta \). Under this normality assumption, the Hammersley-Clifford theorem implies that the conditional independence graph is encoded by the sparsity pattern of \( \Theta \): \((j,k)\) is an edge if and only if \( \Theta_{jk} \neq 0 \). To estimate \( \Theta \) in high dimensions, Yuan and Lin (2007); Banerjee et al. (2008); Friedman et al. (2008) propose the penalized log-likelihood method named graphical LASSO. To reduce the estimation bias, Lam and Fan (2009); Johnson et al. (2012); Shen et al. (2012) propose either greedy algorithms or nonconvex penalizations for sparse precision matrix estimation. Bickel et al. (2008) propose a thresholding method and Ravikumar et al. (2011); Rothman et al. (2008) study the theoretical properties of nonconvex penalizations. Yuan (2010); Cai et al. (2011) and Sun and Zhang (2013) propose the graphical Dantzig selector, CLIME and scaled-LASSO respectively for sparse precision matrix estimation. Guo et al. (2011); Danaher et al. (2014) and Mohan et al. (2014) also consider the problem of estimating multiple Gaussian graphical models via precision matrix estimation. In addition, Meinshausen and Bühlmann (2006); Peng et al. (2009); Friedman et al. (2010) propose the penalized pseudo-likelihood idea, which treats each node as response and the rest of the graph as covariates, then conduct variable selection using penalized regression. Both the \( \ell_1 \)-penalty (Tibshirani, 1996; Bickel et al., 2009; Meinshausen and Yu, 2009; Zhao and Yu, 2006) and folded concave penalties (Fan and Li, 2001; Zhang et al., 2010; Zhang, 2010; Zhang et al., 2013; Breheny and Huang, 2011) can be applied to such models. In addition to Gaussian graphical models, Liu et al. (2009); Xue et al. (2012b); Liu et al. (2012) also study the semiparametric Gaussian copula models and Voorman et al. (2014a) study the joint additive models for graph estimation.

For discrete data, the most commonly used method is Ising graphical model, which is useful
for modeling binary data. In such models, calculating the log-partition function is in general computationally intractable. To overcome this challenge, Ravikumar et al. (2010) propose the nodewise regression approach for graph estimation, which can be viewed as a penalized pseudo likelihood method (Lee et al., 2006; Schmidt et al., 2007; Anandkumar et al., 2012; Guo et al., 2010). In particular, under Ising models, the nodewise regression approach reduces to solving a sparse logistic regression problem. Höfling and Tibshirani (2009); Ravikumar et al. (2010) and Xue et al. (2012a) have studied the theoretical properties of this approach using either \( \ell_1 \)-penalty or nonconvex penalties. Cheng et al. (2012) also propose a conditional Ising model with covariates to model binary data with additional covariates. In addition to binary data, Allen and Liu (2012) and Yang et al. (2013b) propose the Possion graphical models. Moreover, Yang et al. (2013a) propose univariate exponential family graphical models, which assume that the conditional distributions of all nodes follow a generalized linear model. In addition, Guo et al. (2014) study the graphical models for ordinal data and Tan et al. (2014) propose a general framework for graphical models with hubs.

For mixed data, Lee and Hastie (2014); Fellinghauer et al. (2013); Cheng et al. (2013); Chen et al. (2013) study mixed graphical models. To be more specific, Lee and Hastie (2014) consider the problem of structure learning of mixed graphical models; Fellinghauer et al. (2013) propose a random forest approach combined with stability selection for graph estimation; Chen et al. (2013) establish consistency of edge selection via conditional maximum likelihood estimation with \( \ell_1 \)-penalization; Cheng et al. (2013) propose a conditional Gaussian model for mixed data and propose graph estimation via maximizing conditional log-likelihood with weighted \( \ell_1 \)-penalty. Recently, Yang et al. (2014) propose a more general class of mixed graphical models where the nodes form different blocks. The edges within each block are undirected whereas the edges across different blocks are directed. Yang et al. (2014) propose an \( \ell_1 \)-penalized nodewise regression to fit this model and establish its graph estimation consistency.

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) establishes confidence subgraph of Gaussian graphical models. In high dimensions, Ren et al. (2013); Jankova and van de Geer (2014) study the local inference under Gaussian graphical models and Liu et al. (2013) studies global inference by controlling the false discovery rate. However, all these work rely on the Gaussian models. For non-Gaussian models, Wasserman et al. (2014) construct nonparametric confidence subgraphs using the multiplier bootstrap method. Compared with these results, this paper considers the inferential problems under a general family of semiparametric exponential models, which is much more challenging and requires new methods and theory.

1.2 Notation

We adopt the following notation throughout this paper. For any vector \( \mathbf{v} = (v_1, \ldots, v_d)^T \in \mathbb{R}^d \), define its support as \( \text{supp}(\mathbf{v}) = \{t: v_t \neq 0\} \). We define its \( \ell_0 \)-norm, \( \ell_p \)-norm and \( \ell_\infty \)-norm as \( \|\mathbf{v}\|_0 = |\text{supp}(\mathbf{v})| \), \( \|\mathbf{v}\|_p = \left(\sum_{i \in [d]} |v_i|^p\right)^{1/p} \) and \( \|\mathbf{v}\|_\infty = \max_{i \in [d]} |v_i| \). Let \( \mathbf{v} \otimes \mathbf{u} = \mathbf{v}^T \mathbf{u} \) be the Kronecker product of a vector \( \mathbf{v} \) and itself and let \( \mathbf{v} \circ \mathbf{u} = (v_1u_1, \ldots, v_du_d)^T \) be the Hadamard product of two vectors \( \mathbf{u}, \mathbf{v} \in \mathbb{R}^d \). In addition, we use \( |\mathbf{v}| = (|v_1|^p, \ldots, |v_d|^p)^{1/p} \) to denote the elementwise absolute value of vector \( \mathbf{v} \) and let \( \|\mathbf{v}\|_\infty = \max_{i \in [d]} |v_i| \). For any matrix \( \mathbf{A} = [a_{ij}] \in \mathbb{R}^{d_1 \times d_2} \), let \( \mathbf{A}_{S_1S_2} = [a_{ij}]_{i \in S_1, j \in S_2} \) be the submatrix of \( \mathbf{A} \) with indices in \( S_1 \times S_2 \); let \( \mathbf{A}_{j \setminus j} = [a_{jk}]_{k \neq j} \). Let \( \|\mathbf{A}\|_2 \),
∥A∥₁, ∥A∥∞, ∥A∥ℓ₁ are the spectral norm, elementwise supreme norm, elementwise ℓ₁-norm and operator ℓ₁-norm respectively. For two matrices A₁ and A₂, we denote A₁ < A₂ if A₂ − A₁ is positive semidefinite and denote A₁ ≤ A₂ if every entry of A₂ − A₁ is nonnegative. For a function f(x): Rᵈ → R, let ∇f(x), ∇²f(x), ∇²f(x) and ∂f(x) be the gradient of f(x), gradient of f(x) with respect to xₛ, the Hessian of f(x) and the subgradient of f(x). Let |d| = {1, 2, ..., d} be the first d positive integers. For a sequence of d-dimensional random vectors {Yᵢ}₁≤i≤d, we denote Yᵢ ∼ Y when Yᵢ converges to a random vector Y in distribution. Finally, for functions f(n) and g(n), we write f(n) ≲ g(n) to denote that f(n) ≤ cg(n) for a universal constant c ∈ (0, +∞) and we write f(n) ∼ g(n) when f(n) ≈ g(n) and g(n) ≈ 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 methodology 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 more discussions.

2 Semiparametric Exponential Family Graphical Models

The semiparametric exponential family graphical models are defined by specifying the conditional distribution of each variable Xⱼ on the rest of the variables {Xₖ; k ≠ j}. In particular, it requires such a conditional distribution to follow a generalized linear model with canonical link.

Definition 2.1 (Semiparametric exponential family graphical model). A d-dimensional random vector X = (X₁, ..., Xᵅ) ∈ Rᵈ follows a semiparametric exponential graphical model Markov to G = (V, E) if for any node j ∈ V, the conditional density of Xⱼ given Xⱼ̸j := (Xₖ; k ≠ j)ᵗ satisfies

\[ p(xⱼ | xⱼ̸j) = \exp \left\{ xⱼ(βⱼᵗ xⱼ̸j) + fⱼ(xⱼ) - bⱼ(βⱼ, fⱼ) \right\}, \tag{2.1} \]

where fⱼ(·) is a base measure that does not depend on Xⱼ̸j and bⱼ(·, ·) is the log-partition function. In particular, (j, k) ∈ E if and only if βⱼk ≠ 0.

This model is semiparametric since we treat both βⱼ and fⱼ(·) as parameters. If we treat Xⱼ as response and Xⱼ̸j as covariates, then Definition 2.1 is reduced to a semiparametric regression model. In this case, βⱼ is our parameter of interest and fⱼ(·) is an infinite-dimensional nuisance. Since the model in Definition 2.1 is only specified by the conditional distribution of each variable, it is important to understand the conditions that a valid joint distribution of X exists. This problem has been addressed by Chen et al. (2013). As shown in Proposition 1 of their paper, it is sufficient for the existence of joint distribution of X satisfying (2.1) for each Xⱼ if (i) βⱼk = βⱼ for 1 ≤ j, k ≤ d and (ii) g(x) := \exp \{ \sum_{j<k} βⱼk xⱼ xₖ + \sum_{j=1}^{d} fⱼ(xⱼ) \} is integrable.

This paper assumes the above conditions (i) and (ii), so that there exists a joint probability distribution for the model defined in (2.1), whose density has the form of

\[ p(x) = \exp \left\{ \sum_{k<ℓ} βⱼk xⱼ xₖ + \sum_{j=1}^{d} fⱼ(xⱼ) - A(\{βᵢ, fᵢ\}_{i=1}^{d}) \right\}, \tag{2.2} \]
where $\beta_{k\ell} \neq 0$ if and only if $(k, \ell) \in E$. The log-partition function $A(\cdot)$ is defined as

$$A\left(\{\beta_i, f_i\}_{i \in [d]}\right) := \log \left[ \int_{\mathbb{R}^d} \exp\left\{ \sum_{k<\ell} \beta_{k\ell} x_k x_\ell + \sum_{j=1}^d f_j(x_j) \right\} d\nu(x) \right],$$

(2.3)

where $\nu(\cdot)$ is the measure with respect to which the density (2.2) is taken. Note that we assume that $\beta_{k\ell} = \beta_{\ell k}$ for all pair of nodes $k, \ell$, we will use $\beta_{k\ell}$ and $\beta_{\ell k}$ interchangeably for notational simplicity.

### 2.1 Concrete Examples

We provide concrete examples to show that the family of semiparametric exponential family graphical models contains some of the most widely used graphical models as special cases.

**Gaussian Graphical Models:** The Gaussian graphical models assume that $X \in \mathbb{R}^d$ follows a multivariate Gaussian distribution $N(0, \Sigma)$. Let $\Theta = \Sigma^{-1}$ be the precision matrix, the conditional distribution of $X_j$ given $X_{\setminus j}$ satisfies

$$X_j \mid X_{\setminus j} = \alpha_j^T X_{\setminus j} + \epsilon_j \quad \text{with} \quad \epsilon_j \sim N(0, \sigma_j^2),$$

where $\alpha_j = \Theta_{j\setminus j}^{-1} \Theta_{j\setminus j}$ and $\sigma_j^2 = \Theta_{jj}^{-1}$. The conditional distribution is given by

$$p(x_j \mid x_{\setminus j}) = \sqrt{\Theta_{jj}/(2\pi)} \exp\left\{ -x_j (\Theta^{-1}_{j\setminus j} x_{\setminus j}) - \frac{1}{2} \Theta_{jj} x_j^2 - \frac{1}{2} (\Theta^{-1}_{j\setminus j} x_{\setminus j})^2 / \Theta_{jj} \right\}.$$

Therefore $\beta_j = -\Theta_{j\setminus j}, f_j(x) = -\Theta_{jj} x^2 / 2$ and $b_j(\beta_j, f_j) = \Theta_{jj}^{-1}(\Theta^{-1}_{j\setminus j} x_{\setminus j})^2 / 2 + \log(2\pi/\Theta_{jj}) / 2$.

**Ising Models:** In an Ising model with no external field, $X$ takes value in $\{0, 1\}^d$ and the joint probability mass function $p(x) \propto \exp\{ \sum_{j<k} \theta_{jk} x_j x_k \}$. Denoting $(\theta_{jk}; k \neq j)^T$ as $\theta_j$, the conditional distribution of $X_j$ given $X_{\setminus j}$ is of the form

$$p(x_j \mid x_{\setminus j}) = \frac{\exp\{ \sum_{k<\ell} \theta_{k\ell} x_k x_\ell \}}{\sum_{x_j \in \{0, 1\}} \exp\{ \sum_{k<\ell} \theta_{k\ell} x_k x_\ell \}} = \exp\left\{ x_j (\theta_j^T x_{\setminus j}) - \log(1 + \exp(\theta_j^T x_{\setminus j})) \right\}.$$ 

Therefore in this case we have $\beta_j = \theta_j, f_j(x) = 0$ and $b_j(\beta_j, f) = \log(1 + \exp(\theta_j^T x_{\setminus j})).$

**Exponential Graphical Models:** For exponential graphical models, $X$ takes values in $[0, +\infty)^d$ and the joint probability density satisfies

$$p(x) \propto \exp\left\{ -\sum_{i=1}^d \phi_i x_i - \sum_{k<\ell} \theta_{k\ell} x_k x_\ell \right\}.$$

In order to ensure 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 $X_{\setminus j}$:

$$p(x_j \mid x_{\setminus j}) = \exp\left\{ -\sum_{i=1}^d \phi_i x_i - \sum_{k<\ell} \theta_{k\ell} x_k x_\ell \right\} \int_{x_{\setminus j} \geq 0} \exp\left\{ -\sum_{i=1}^d \phi_i x_i - \sum_{k<\ell} \theta_{k\ell} x_k x_\ell \right\} dx_j$$

$$= \exp\left\{ -x_j (\phi_j + \theta_j^T x_{\setminus j}) - \log(\phi_j + \theta_j^T x_{\setminus j}) \right\},$$

(5)
where \( \theta_j = (\theta_{jk}; k \neq j)^T \). Thus we have \( \beta_j = -\theta_j \), \( f_j(x) = -\phi_j x \) and \( b_j(\beta_j, f_j) = \log(\beta_j^T x + \phi_j) \).

**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

\[
p(x) \propto \exp \left\{ \sum_{j=1}^d \phi_j x_j - \sum_{j=1}^d \log(x_j!) + \sum_{k \prec \ell} \theta_{k\ell} x_k x_\ell \right\}.
\]

Similar to 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 \( \theta_{jk} \leq 0 \) for all \( j, k \in [d] \).

Then we obtain the following conditional probability mass function defined on \( \mathbb{N}^d \)

\[
p(x_j | x_{\setminus j}) = \exp \left\{ x_j (\theta_j^T x_{\setminus j}) + \phi_j x_j - \log(x_j!) - b_j(\theta_j, f_j) \right\},
\]

where \( \beta_j = \theta_j^T \), \( f_j(x) = \phi_j x - \log(x!) \) and \( b_j(\beta_j, f_j) = \log \left\{ \sum_{y=0}^\infty \exp(y(\beta_j^T x_{\setminus j}) + f_j(y)) \right\} \).

## 3 Graph Estimation and Uncertainty Assessment

In this section, we first introduce a pseudo-likelihood loss for \( \{\beta_j\}_{j=1}^d \) that is invariant of the nuisance parameters \( f_1, \ldots, f_d \) in (2.1). Throughout our analysis, we use \( \{\beta^*_i, f^*_i\}_{i \in [d]} \) to denote the true parameters of the graphical model. We present an Adaptive Multi-stage Convex Relaxation algorithm to estimate each \( \beta^*_j \) by minimizing the loss function regularized by a nonconvex penalty function. We then proceed to introduce a pairwise score test for the null hypothesis \( H_0: \beta^*_{jk} = 0 \) and a super-graph test for testing whether a given graph contains the true graph.

### 3.1 A Nuisance-Free Loss Function

For graph estimation, we treat \( \beta_j \) as the parameter of interest and the base measures \( f_j(\cdot) \) as nuisance parameters. Let \( X_1, X_2, \ldots, X_n \) be \( n \) random samples of the graphical model. Due to the presence of \( f_j(\cdot) \), finding the conditional maximum likelihood estimator of \( \beta_j \) is intractable. To solve this problem, we exploit a pseudo-likelihood loss function from Ning and Liu (2014) that is invariant of the nuisance parameters. More specifically, this pseudo-likelihood loss is based on pairwise local order statistic. More details are presented as follows.

Let \( x_1, x_2, \ldots, x_n \) be \( n \) data points that are realizations of \( X_1, X_2, \ldots, X_n \). For any \( 1 \leq i < i' \leq n \), let \( A_{i'j}^i := \{ X_{ij}, X_{i'j} \} = \{ x_{ij}, x_{i'j} \} \), \( X_{i\setminus j} = x_{i\setminus j}, X_{i'\setminus j} = x_{i'\setminus j} \) be the event where we observe \( X_{i\setminus j} \) and \( X_{i'\setminus j} \) and the order statistics of \( X_{ij} \) and \( X_{i'j} \) (but not the relative ranks of \( X_{ij} \) and \( X_{i'j} \)). In another word, we know the values of a pair of observations \( x_i \) and \( x_{i'} \) except that we are not aware of the relative order of \( X_{ij} \) and \( X_{i'j} \). Ning and Liu (2014) show that

\[
\mathbb{P}(X_i = x_i, X_{i'} = x_{i'} | A_{i'j}^i) = \left\{ 1 + \exp\left( - (x_{ij} - x_{i'j}) \beta_j^T (x_{i\setminus j} - x_{i'\setminus j}) \right) \right\}^{-1}.
\]

If we denote \( R_{i'j}^i(\beta_j) := \exp\left\{ - (x_{ij} - x_{i'j}) \beta_j^T (x_{i\setminus j} - x_{i'\setminus j}) \right\} \), we construct a pseudo-likelihood loss function for \( \beta_j \) as

\[
L_j(\beta_j) := \frac{2}{n(n-1)} \sum_{1 \leq i < i' \leq n} \log(1 + R_{i'j}^i(\beta_j)).
\]
The gradient and Hessian of \( L_j(\beta_j) \) are given by
\[
\nabla L_j(\beta_j) = \left( \frac{2}{n(n-1)} \sum_{1 \leq i < i' \leq n} R_{ji}^j(\beta_j)(x_{ij} - x_{i'j})(x_{i\backslash j} - x_{i'\backslash j}) \right),
\]
\[

\nabla^2 L_j(\beta_j) = \left( \frac{2}{n(n-1)} \sum_{1 \leq i < i' \leq n} \frac{R_{ji}^j(\beta_j)(x_{ij} - x_{i'j})(x_{i\backslash j} - x_{i'\backslash j})^2}{(1 + R_{ji}^j(\beta_j))^2} \right). \tag{3.3}
\]

As shown in (3.1), \( L_j(\cdot) \) only involves \( \beta_j \) and its form resembles a logistic loss function. To justify that \( L_j(\beta_j) \) is a valid loss function, suppose \( \{\beta^*_j\} \) and \( \{f_j^*\} \) are the true parameters of the graphical model, then Lemma 3.1 in Ning and Liu (2014) shows that \( \mathbb{E}(\nabla L_j(\beta^*_j)) = 0 \) and \( \beta^*_j \) is the global minimizer of \( \mathbb{E}(L_j(\beta_j)) \), where \( \mathbb{E}(\cdot) \) is the expectation with respect to the density (2.2) under true parameters \( \{\beta^*_j\} \) and \( \{f_j^*\} \).

3.2 Adaptive Multi-stage Convex Relaxation Algorithm

We present the algorithm for parameter estimation. For high dimensional sparse estimation, we minimize the loss functions \( L_j(\beta_j) \) with some penalty function. Two of the most prevalent methods are the LASSO (\( \ell_1 \)-penalization) (Tibshirani, 1996) and the folded concave penalization (Fan et al., 2014). Although the \( \ell_1 \)-penalization enjoys good computational properties as a convex optimization problem, they are known to incur significant estimation bias for parameters with large absolute values. Whereas nonconvex penalties such as smoothly clipped absolute deviation (SCAD) penalty, minimax concave penalty (MCP) and capped-\( \ell_1 \) penalty can eliminate such bias and attain improved rates of convergence. Therefore, we consider the nonconvex optimization problem
\[
\hat{\beta}_j = \arg \min_{\mathbb{R}^{d-1}} L_j(\beta_j) + \sum_{k \neq j} p_\lambda(|\beta_{jk}|),
\]
where and \( \lambda \) is a tuning parameter and \( p_\lambda(\cdot) : [0, +\infty) \to [0, +\infty) \) is a penalty function satisfying:

(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 exists \( c_1 \in [0, 1] \) 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 will denote \( p'_\lambda(u) \) as the right-hand derivative. Since \( p_\lambda(u) \) is continuously nondecreasing and concave, \( p'_\lambda(u) \) is nonincreasing and nonnegative in \( [0, \infty) \). Before we present the algorithm, we show that our formulation of nonconvex penalty function in (3.4) is quite general and cover many interesting examples.

Examples of nonconvex penalty functions

1. **SCAD penalty**: For a constant \( a > 2 \), the SCAD penalty function is defined as
\[
p_\lambda(u) = \lambda \int_0^u \left\{ 1(z \leq \lambda) + \frac{(a \lambda - z)_+}{(a-1)\lambda} 1(z > \lambda) \right\} dz.
\]

For the SCAD penalty function, condition (C.3) of the penalty function is satisfied with \( c_1 \in [0, 1] \) and \( c_2 = a(1-c_1) + c_1 \).
2. **MCP penalty**: For a constant $b > 0$, the MCP penalty function is defined as

$$ p_{\lambda}(u) = \lambda \int_0^u \left(1 - \frac{z}{\lambda b}\right)_+ dz. $$

The condition (C.3) is satisfied for MCP penalty function for $c_1 \in [0, 1]$ and $c_2 = (1 - c_1)b$. 

3. **Capped-$\ell_1$ penalty**: For $\theta > 0$, the capped-$\ell_1$ penalty is defined as $p_{\lambda}(u) = \lambda \min\{u, \theta\}$. Condition (C.3) is satisfied if there exists a constant $C > 0$ such that $\theta \geq c\lambda$; in this case we have $c_1 = 1$ and $c_2 = c$.

Optimization problem (3.4) is nonconvex and may have multiple local solutions. To overcome such difficulty, we exploit local linear approximation algorithm (Zou and Li, 2008; Fan et al., 2014) and multi-stage convex relaxation (Zhang, 2010; Zhang et al., 2013) to solve (3.4). Compared with previous works that mainly focus on the problem of sparse linear regression, our loss function $L_j(\beta_j)$ is a U-statistic based logistic loss, which requires nontrivial extension of the 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 an $\ell_1$-regularized optimization problem and we obtain the first-step solution $\hat{\beta}_{j1}$. Then in each subsequent iteration, we solve an adaptive $\ell_1$-regularized optimization problem where the weights of the penalty depend on the solution of the previous step. For example, in the $\ell$-th iteration, the regularization parameter $\lambda_{jk}^{(\ell-1)}$ in (3.5) is updated using the ($\ell-1$)-th step estimator $\hat{\beta}_{j}^{(\ell-1)}$. Note that $p_{\lambda}'(\beta_{jk}^{(\ell-1)})$ is the right-hand derivative of $p_{\lambda}(u)$ evaluated at $u = \beta_{jk}^{(\ell-1)}$.

Since the optimization problem in step 4 is convex, our method is computationally efficient. Moreover, the solution of (3.5) for $\ell = 1$ is the solution to the $\ell_1$-regularized problem. As we will show in §4.1, the estimator $\hat{\beta}_j$ of $\beta^*_j$ obtained by Algorithm 1 attains the optimal rates of convergence for parameter estimation.

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

1: Initialize $\lambda_{jk}^{(0)} = \lambda$ for $1 \leq j, k \leq d$.
2: for $j = 1, 2, \ldots, d$ do
3:   for $\ell = 1, 2, \ldots, $ until convergence do
4:     Solve the convex optimization problem

$$ \hat{\beta}_j^{(\ell)} = \arg \min_{\mathbb{R}^d} \left\{ L_j(\beta_j) + \sum_{k \neq j} \lambda_{jk}^{(\ell-1)}|\beta_{jk}| \right\}. $$

(3.5)

5:     Update $\lambda_{jk}^{(\ell)}$ by $\lambda_{jk}^{(\ell)} = p_{\lambda}'(\hat{\beta}_{jk}^{(\ell)})$ for $1 \leq k \leq d, k \neq j$.
6:   end for
7: Output $\hat{\beta}_j = \hat{\beta}_j^{(\ell_j)}$ where $\ell_j$ is the number of iterations until convergence appears.
8: end for
3.3 Local Graph Inference: Pairwise Score Test

We first consider the local hypothesis testing problem \( H_0 : \beta^*_{jk} = 0 \) versus \( H_1 : \beta^*_{jk} \neq 0 \) for any given \( 1 \leq j < k \leq d \). This problem is “local” since it only involves statistical inference for a single parameter \( \beta_{jk} \). This is in contrast to global inferential problems where we aim to infer \( \beta^* \) or a high dimensional subvector of \( \beta^* \). For local inference, we propose a new test procedure named pairwise score test, which involves the gradient of the pair of loss functions \( L_j(\beta_j) \) and \( L_k(\beta_k) \).

Letting \( \beta_{jk} = (\beta_{jk}; \ell \neq k)^T \), we denote the parameters associated with node \( j \) and node \( k \) as \( \beta_{jk} := (\beta_{jk}; \beta^T_{j\setminus k}, \beta^T_{k\setminus j})^T \in \mathbb{R}^{d-3} \). Let \( \mathbf{H}^j := \mathbb{E}\{\nabla^2 L_j(\beta^*_j)\} \) be the expectation of the Hessian of \( L_j(\beta_j) \) evaluated at the true parameter \( \beta^*_j \). We define two submatrices \( \mathbf{H}^j_{jk,j\setminus k} \) and \( \mathbf{H}^j_{k,j\setminus j} \) as

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

and define \( \mathbf{H}^k_{jk,k\setminus j} \) and \( \mathbf{H}^k_{k,j\setminus j} \) similarly. Let \( \mathbf{w}^*_k = \mathbf{H}^j_{jk,j\setminus k}^{-1} \mathbf{H}^j_{j\setminus k,j\setminus k}^{-1} \) and \( \mathbf{w}^*_k = \mathbf{H}^k_{jk,k\setminus j}^{-1} \mathbf{H}^k_{j\setminus k,j\setminus j}^{-1} \). The pairwise score function for parameter \( \beta_{jk} \) is defined as

\[
S_{jk}(\beta_{jk}) = \nabla_j L_j(\beta_j) + \nabla_k L_k(\beta_k) - \mathbf{w}^*_j \nabla_j L_j(\beta_k) - \mathbf{w}^*_k \nabla_k L_k(\beta_j). \tag{3.6}
\]

where \( \nabla_j L_j(\beta_j) = \partial L_j(\beta_j)/\partial \beta_{jk} \) and \( \nabla_k L_k(\beta_j) = \partial L_k(\beta_j)/\partial \beta_{jk} \).

Note that \( \mathbf{w}^*_k \) and \( \mathbf{w}^*_k \) involve unknown quantities, we estimate them using the Dantzig-type estimators. We define the empirical versions of \( \mathbf{H}^j_{jk,j\setminus k} \) and \( \mathbf{H}^j_{k,j\setminus j} \) as

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

We also define \( \tilde{\nabla}^2_{jk,k\setminus j} L_k(\beta_k) \) and \( \tilde{\nabla}^2_{k,j\setminus j} L_k(\beta_k) \) similarly. Then we estimate \( \mathbf{w}^*_k \) by solving the optimization problem:

\[
\hat{\mathbf{w}}_{j,k} = \arg\min \| \mathbf{w} \|_1 \quad \text{such that} \quad \| \tilde{\nabla}^2_{jk,j\setminus k} L_j(0, \hat{\beta}_j) - \hat{\mathbf{w}}^T \tilde{\nabla}^2_{k,j\setminus k} L_j(0, \hat{\beta}_{k\setminus j}) \|_\infty \leq \lambda_D, \tag{3.7}
\]

where \( \hat{\beta}_j \) is the estimator of \( \beta^*_j \) obtained from Algorithm 1 and \( \lambda_D \) is a tuning parameter. An estimator \( \hat{\mathbf{w}}_{k,j} \) of \( \mathbf{w}^*_k \) can be similarly obtained. With \( \hat{\mathbf{w}}_{j,k} \) and \( \hat{\mathbf{w}}_{k,j} \), we construct a pairwise score statistic for \( \beta_{jk} \) as

\[
\tilde{S}_{jk} = \nabla_j L_j(0, \hat{\beta}_j) + \nabla_k L_k(0, \hat{\beta}_k) - \hat{\mathbf{w}}^T_{j,k} \tilde{\nabla}^2_{jk,j\setminus k} L_j(0, \hat{\beta}_{k\setminus j}) - \hat{\mathbf{w}}^T_{k,j} \tilde{\nabla}^2_{k,j\setminus j} L_k(0, \hat{\beta}_{k\setminus j}). \tag{3.8}
\]

Comparing (3.6) and (3.8) we see that the pairwise score statistic is computed by replacing \( \beta_j \) and \( \beta_k \) with \( (0, \hat{\beta}_{j\setminus k}) \) and \( (0, \hat{\beta}_{k\setminus j}) \) respectively and replacing \( \mathbf{w}^*_j \) and \( \mathbf{w}^*_k \) with \( \hat{\mathbf{w}}_{j,k} \) and \( \hat{\mathbf{w}}_{k,j} \).

To obtain a valid test, we need to derive the limiting distribution of \( \tilde{S}_{jk} \) under the null hypothesis. \( \tilde{S}_{jk} \) is a linear combination of entries of \( \tilde{\nabla} L_j(\beta_j) \) and \( \tilde{\nabla} L_k(\beta_k) \), both of which are U-statistics. In the next section, we prove the asymptotic normality of \( \tilde{S}_{jk} \). More specifically, under the null hypothesis, \( \sqrt{n} \tilde{S}_{jk}/2 \sim N(0, \sigma^2_{jk}) \) where the limiting variance can be estimated consistently by \( \tilde{\sigma}^2_{jk} \).
(More details will be explained in the next section). Let $\alpha \in (0,1)$, our proposed pairwise score test function $\psi_{jk}(\alpha)$ is defined by

$$
\psi_{jk}(\alpha) = \begin{cases} 
1 & \text{if } |\sqrt{n}\tilde{S}_{jk}/(2\tilde{\sigma}_{jk})| > \Phi^{-1}(1-\alpha/2) \\
0 & \text{if } |\sqrt{n}\tilde{S}_{jk}/(2\tilde{\sigma}_{jk})| \leq \Phi^{-1}(1-\alpha/2) 
\end{cases},
$$

(3.9)

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

To summarize the pairwise score test for null hypothesis $H_0: \beta_{jk}^* = 0$: (i) Calculate $\hat{\beta}_j$ and $\hat{\beta}_k$ from Algorithm 1; (ii) Obtain $\hat{w}_{j,k}$ and $\hat{w}_{k,j}$ by solving two Dantzig-type problems (3.7); (iii) Compute the limiting variance $\tilde{\sigma}_{jk}^2$; (iv) Obtain the outcome of pairwise score test by (3.9).

We remark that there are only limited works on the uncertainty assessment for high dimensional models. Meinshausen et al. (2009); Wasserman and Roeder (2009); Meinshausen and Bühlmann (2010) and Shah and Samworth (2013) propose sample splitting or subsampling methods to obtain p-values. For the LASSO estimator, Lee et al. (2013); Lockhart et al. (2014); Taylor et al. (2014) consider statistical inference for high-dimensional linear regression after some variables are selected. Belloni et al. (2012, 2013) propose the method of instrumental variables together with a double selection procedure. In addition, Zhang and Zhang (2014); Javanmard and Montanari (2013); van de Geer et al. (2014) propose a bias correction method named the LDPE or the desparsity method to construct confidence intervals for linear models or generalized linear models with $l_1$-penalty. These works utilize the convexity and the Karush-Kuhn-Tucker 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. For statistical inference with nonconvex penalizations, Fan and Lv (2011); Bradic et al. (2011) establish the asymptotic normality for low dimensional nonzero parameters in high-dimensional models based on the oracle properties. However, their approach depends on the assumptions on the minimal signal strength, which is not applicable to situations where signals are weak. Recently, Voorman et al. (2014b) propose a penalized score test for linear model. Compared with their method, our pairwise score test does not hinge on an irrepresentable-type condition and is constructed for graphical models.

### 3.4 Global Graph Inference: Super-graph Test

We present a hypothesis testing procedure for testing whether a given graph $G = (V, E)$ is a super-graph of the true graph $G^* = (V, E^*)$ corresponding to the graphical model (2.1). Here we say $G^* \subset G$ if $(j,k) \in E^*$ implies that $(j,k) \in E$. Therefore the null hypothesis $H_0: G^* \subset G$ is equivalent to $H_0: \beta_{jk}^* = 0$ for all $(j,k) \in E^c$, where we denote $E^c$ as $\{(j,k): 1 \leq j < k \leq d, (j,k) \notin E\}$. Note that testing super-graph is a multiple testing problem and we consider it a global statistical inference on graph since it examines the structure of the entire graph instead of the existence of a single edge. It is clear that the null hypothesis is rejected if we reject $H_0: \beta_{jk}^* = 0$ for any $(j,k) \in E^c$, therefore it is possible to combine pairwise score tests for each $(j,k) \in E^c$ to obtain a valid test for super-graph. Recall that the pairwise score statistic $\tilde{S}_{jk}$ defined in (3.8) is a 2nd-order U-statistic with a Gaussian limiting distribution, we define

$$
T := \sqrt{n} \max_{(j,k) \in E^c} \tilde{S}_{jk}/2
$$

as the test statistic. We construct a valid asymptotically level-$\alpha$ test by rejecting the null hypothesis $H_0: G^* \subset G$ if $T$ is greater than the $(1-\alpha)$-th quantile of $T$, which is defined as $c_T(1-\alpha) := \inf \{ t \in \mathbb{R}: \mathbb{P}(T \leq t) \leq 1 - \alpha \}$. 

10
Now the problem reduces to finding the quantiles of $T$, which is closely related to finding the quantiles of the maximum of a sum of high-dimensional random vector as considered in Chernozhukov et al. (2013). Recall that the quantity $T$ is the maximum of a high dimensional U-statistic, to obtain the limiting distribution of $T$, we propose a variant of the Gaussian multiplier bootstrap which is tailored towards the U-statistics structure. More specifically, recall that the gradient $\nabla L_j(\beta_j)$ is a 2nd-order U-statistic with kernel function

$$
\hat{h}_{ii'}^j(\beta_j) := \frac{R_{ii'}^j(\beta_j)(x_{ij} - x_{i'j})(x_{i'j} - x_{ij})}{1 + R_{ii'}^j(\beta_j)},
$$

we can represent $\hat{S}_{jk}$ as a summation

$$
\hat{S}_{jk} = \frac{2}{n(n-1)} \sum_{1 \leq i < i' \leq n} \left\{ (1, -\hat{w}_{j,k}^T)^T h_{ii'}^j(0, \hat{\beta}_j \setminus k) + (1, -\hat{w}_{k,j}^T)^T h_{ii'}^k(0, \hat{\beta}_k \setminus j) \right\}. \quad (3.10)
$$

Then we draw $n$ independent samples $\xi_1, \ldots, \xi_n$ from $N(0,1)$ and define a multiplier bootstrap statistic as

$$
\hat{S}_{jk}^B := \frac{2}{n(n-1)} \sum_{1 \leq i < i' \leq n} \left\{ (1, -\hat{w}_{j,k}^T)^T h_{ii'}^j(0, \hat{\beta}_j \setminus k) + (1, -\hat{w}_{k,j}^T)^T h_{ii'}^k(0, \hat{\beta}_k \setminus j) \right\} \cdot (\xi_i + \xi_{i'}). \quad (3.11)
$$

We denote a bootstrap estimator of $T$ as

$$
W := \sqrt{n} \max_{(j,k) \in E^c} \hat{S}_{jk}^B/2.
$$

Now we define the conditional $\alpha$-th quantile of $W$ by $c_W(\alpha) := \inf \{ t \in \mathbb{R} : \mathbb{P}_t(W \leq t) \leq \alpha \}$ where $\mathbb{P}_t$ is the probability measure induced by $\xi_1, \ldots, \xi_n$. Even though we cannot obtain $c_W(\alpha)$ directly, it can be easily obtained using Monte-Carlo methods. Theoretically, Chernozhukov et al. (2013) show that the quantiles of $W$ converges to that of $T$ uniformly in $\alpha$. Therefore a level-$\alpha$ super-graph test can be constructed as

$$
\psi_G(\alpha) = \begin{cases} 
1 & \text{if } T > c_W(1-\alpha) \\
0 & \text{if } T \leq c_W(1-\alpha)
\end{cases}. \quad (3.12)
$$

We conclude this section by summarizing the procedure of our super-graph test: (i) Obtain $\hat{\beta}_j$ for each $j = 1, \ldots, d$ using Algorithm 1; (ii) Compute $h_{ii'}^j(0, \hat{\beta}_j \setminus k)$ and $h_{ii'}^k(0, \hat{\beta}_k \setminus j)$ for $1 \leq i < i' \leq n$ and $(j,k) \in E^c$; (iii) Solve the Dantzig-type optimization problems to obtain $\hat{w}_{j,k}$ and $\hat{w}_{k,j}$ for $(j,k) \in E^c$; (iv) Draw $n \cdot B$ independent samples $\{\xi_1^{(i)}, \ldots, \xi_n^{(i)}\}_{i=1}^B$ from $N(0,1)$ for a large integer $B$; (v) Compute bootstrap statistics $W^{(i)}$ for $1 \leq i \leq B$ and obtain the $(1-\alpha)$-th sample quantiles $c_W(1-\alpha)$; (vi) Compute test statistic $T$ and obtain the outcome of the super-graph test by (3.12).

### 4 Theoretical Properties

We have three main results. We first prove that NOSI attains the optimal rate of convergence for parameter estimation. We provide theory of the pairwise score test and super-graph test.
4.1 Theoretical Results for Parameter Estimation

We first establish theoretical results on the rates of convergence of our adaptive multi-stage convex relaxation method for solving the nonconvex parameter estimation problem (3.4). We begin by listing several required assumptions. The first is about moment conditions of \( \{X_j\} \) and local smoothness of the log-partition function \( A(\cdot) \) defined in (2.3). This assumption also appears in Yang et al. (2013a) and Chen et al. (2013) as a pivotal technical condition for theoretical analysis.

**Assumption 4.1.** For all \( j \in [d] \), we assume that the first two moments of \( X_j \) are bounded, that is, there exist two constants \( \kappa_m \) and \( \kappa_v \) such that

\[
|E(X_j)| \leq \kappa_m \quad \text{and} \quad E(X_j^2) \leq \kappa_v.
\]

Moreover, we denote the true parameters of the graphical model as \( \{\beta^*_j, f^*_j\}_{j \in [d]} \) and define \( d \) univariate functions \( \bar{A}_j(\cdot) : \mathbb{R} \to \mathbb{R} \) as

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

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

Unlike in the Gaussian or Ising graphical models, we cannot directly assume the boundedness of \( \{X_j\}_{j \in [d]} \) for semiparametric exponential family graphical models. Instead, we impose mild conditions as in Assumption 4.1 to obtain a loose control of the tail behaviors of the graphical model so that the distribution of \( X \) is not too ill-conditioned. As shown in Yang et al. (2013a), Assumption 4.1 implies that for all \( j \in [d] \),

\[
\max \left\{ \log E(\exp(X_j)), \log E(\exp(-X_j)) \right\} \leq \kappa_m + \kappa_h/2.
\]

This further implies that for any \( x > 0 \) we have the tail probability:

\[
P(|X_j| \geq x) \leq 2 \exp(\kappa_m + \kappa_h/2) \exp(-x).
\]

Letting \( x = C \log d \), we immediately get \( X_j = \mathcal{O}_P(\log d) \).

Using (4.1), the following theorem bounds \( \|\nabla L_j(\beta_j^*)\|_\infty \) with high probability.

**Theorem 4.2.** Let \( s^* = \max_{j \in [d]} \|\beta_j^*\|_0 \) and suppose that \( s^* = o(n) \). Under Assumption 4.1, there exist two positive constants \( K_1 \) and \( K_2 \) such that, for any fixed \( j \in [d] \) and any index set \( S \subset \{(j,k) : k \in [d], k \neq j\} \) with \( |S| \leq s^* \),

\[
\|\nabla L_j(\beta_j^*)\|_\infty \leq K_1 \sqrt{\log^5 d/n} \quad \text{and} \quad \|\nabla S L_j(\beta_j^*)\|_2 \leq K_2 \sqrt{s^* \log^5 n/n}
\]

hold with probability at least \( 1 - (2d)^{-1} \).

**Proof of Theorem 4.2.** See §A.1 for a detailed proof.

12
Similar bounds for the gradient of loss function have been constructed for loss functions in other sparse estimation problems. Bickel et al. (2009) and Zhang (2010) establish similar bound for least-square loss function; Van de Geer (2008) bound the gradient of loss function for generalized linear models and Wang et al. (2014) analyze the semiparametric elliptical design loss.

In addition to Assumption 4.1, we also impose conditions to control the curvature of function $L_j(\cdot)$. In high-dimensional settings, although the loss function is generally not strongly convex, it is sometimes strongly convex in some directions. To characterize this phenomenon, we define the sparse eigenvalue condition for $L_j(\cdot)$ as follows.

**Definition 4.3** (Local sparse eigenvalues). Let $m$ be a positive integer and $M(\cdot): \mathbb{R}^m \to \mathbb{S}^m$ be a mapping from $\mathbb{R}^m$ to the space of $m \times m$ symmetric matrices. We define the $s$-sparse eigenvalues of $M(\cdot)$ over the $\ell_1$-ball centered at $u_0 \in \mathbb{R}^m$ with radius $r$ as

$$
\rho_+(M(\cdot), u_0; s, r) = \sup_{v,u} \left\{ v^T M(u) v : ||v||_0 \leq s, \|v\|_2 = 1, \|u - u_0\|_1 \leq r \right\};
$$

$$
\rho_-(M(\cdot), u_0; s, r) = \inf_{v,u} \left\{ v^T M(u) v : ||v||_0 \leq s, \|v\|_2 = 1, \|u - u_0\|_1 \leq r \right\}.
$$

**Assumption 4.4.** Let $s^* = \max_{j \in [d]} ||\beta_j^*||_0$. We assume that for any $j \in [d]$, there exist an integer $k^* \geq 2s^*$ and two positive numbers $\rho_*$ and $r$ satisfying $\lim_{n \to \infty} s^*(\log^2 d/n)^{1/2}/r = 0$ and $\lim_{n \to \infty} k^*(\log^9 d/n)^{1/2} = 0$ such that the sparse eigenvalues of $\mathbb{E}\{\nabla^2 L_j(\cdot)\}$ satisfy

$$
0 < \rho_* \leq \rho_{j-}(2s^* + k^*) < \rho_{j+}(k^*) < +\infty \quad \text{and} \quad \rho_{j+}(k^*)/\rho_{j-}(2s^* + k^*) \leq 1 + 0.2k^*/s^* \quad \text{for any} \ j \in [d],
$$

where we denote the $s$-sparse eigenvalues $\rho_-(\mathbb{E}\{\nabla^2 L_j(\cdot)\}, \beta_j^*; s, r)$ and $\rho_-(\mathbb{E}\{\nabla^2 L_j(\cdot)\}, \beta_j^*; s, r)$ as $\rho_{j-}^*(s)$ and $\rho_{j+}^*(s)$ respectively for notational simplicity.

The condition $\rho_{j+}^*(k^*)/\rho_{j-}^*(2s^* + k^*) \leq 1 + 0.2k^*/s^*$ requires the eigenvalue ratio $\rho_{j+}^*(s)/\rho_{j-}^*(s)$ to grow sub-linearly in $s$. Assumption 4.4 is commonly referred to as sparse eigenvalue condition. Conditions like this is standard for sparse estimation and have been studied by Bickel et al. (2009); Raskutti et al. (2010); Zhang (2010); Negahban et al. (2012); Xiao and Zhang (2013); Loh and Wainwright (2013) and Wang et al. (2014). In this paper, we only assume that the sparse eigenvalue condition holds within an $\ell_1$-ball centered at $\beta_j^*$ with radius $r$, which is weaker than the conditions in Wang et al. (2014). Our assumption is similar to that in Zhang (2010) and is weaker than the restricted isometry property (RIP) proposed by Candès and Tao (2005).

The following lemma justifies Assumption 4.4 for Gaussian graphical models.

**Proposition 4.5** (Sparse eigenvalue condition for Gaussian graphical models). Suppose $X \sim N(0, \Sigma)$ is a Gaussian graphical model and let $\Theta = \Sigma^{-1}$ be the precision matrix. For all $j \in [d]$, the conditional distribution of $X_j$ given $X_{\backslash j}$ is a normal distribution with mean $\beta_j^T X_{\backslash j}$ and variance $\Theta_{jj}^{-1}$, where $\beta_j^* = \Theta_{jj}^{-1}$. Let $L_j(\cdot)$ be the loss function defined in (3.1). We assume that there exist positive constants $D$, $c_2$ and $C_2$ such that $\|\Sigma\|_\infty \leq D$ and $c_2 \leq \lambda_{\max}(\Sigma) \leq \lambda_{\min}(\Sigma) \leq C_2$. We let $s^* = \max_{j \in [d]} ||\beta_j^*||_0$ and also assume that there exist a $C_3 > 0$ such that $\|\beta_j^*\|_2 \leq C_3$ for all $j \in [d]$. Suppose $r > 0$ is a real number such that $r = O(1/s^*)$, there exist $\rho_*, \rho^* > 0$ such that for all $j \in [d]$, and $s = 1, \ldots, d - 1$,

$$
\rho_* \leq \rho_-(\mathbb{E}\{\nabla L_j(\cdot)\}, \beta_j^*; s, r) \leq \rho_+(\mathbb{E}\{\nabla L_j(\cdot)\}, \beta_j^*; s, r) \leq \rho^*.
$$
Proof of Proposition 4.5. See §A.2 for a detailed proof.

Remark 4.6. Note that in Proposition 4.5, the sparse eigenvalue conditions hold for Gaussian graphical models within a ball centered at \( \beta_j^\ast \) with radius \( r = \mathcal{O}(1/\sqrt{s^\ast}) \). Because \( \rho_\ast \) and \( \rho^\ast \) in Proposition 4.5 are constants, we can find a \( k^\ast \approx s^\ast \) such that \( \rho^\ast/\rho_\ast \leq 1 + 0.2k^\ast/s^\ast \). If the scaling of \( s^\ast, n \) and \( d \) satisfies \( \lim_{n \to \infty} s^\ast((\sqrt{s^\ast}+\log^2 d)(\log^5 d/n)^{1/2} = 0 \), Assumption 4.4 holds with \( r \) satisfying \( \lim_{n \to \infty} s^\ast((\log^5 d/n)^{1/2}/r = 0 \) and \( r \leq 1/\sqrt{s^\ast} \).

By the law of large numbers, if the sample size \( n \) is sufficiently large such that \( \nabla^2 L_j(\cdot) \) is close to its expectation \( \mathbb{E}\{\nabla^2 L_j(\cdot)\} \), we expect that the sparse eigenvalue condition also holds for \( \nabla^2 L_j(\cdot) \). The following lemma justifies this intuition by showing that the sparse eigenvalue condition holds for \( \nabla^2 L_j(\cdot) \) with high probability.

Lemma 4.7. Under Assumption 4.1 and 4.4, if \( n \) is sufficiently large such that \( \rho_\ast \gtrsim s^\ast(\log^9 d/n)^{-1/2} \), with probability at least \( 1 - (2d)^{-1} \), for all \( j \in [d] \), the sparse eigenvalues of \( \nabla^2 L_j(\cdot) \) satisfy

\[
\rho_{j,-}(2s^\ast+k^\ast) - 0.05\rho_\ast \leq \rho_{j,-}(2s^\ast+k^\ast) \leq \rho_{j,+}(k^\ast) \leq \rho_{j,+}(k^\ast) + 0.05\rho_\ast, \quad \text{and} \quad \rho_{j,+}(k^\ast)/\rho_{j,-}(2s^\ast+k^\ast) \leq 1 + 0.27k^\ast/s^\ast,
\]

where \( k^\ast \) and \( r \) are defined in Assumption 4.4. and we denote \( s \)-sparse eigenvalues \( \rho_-(\nabla^2 L_j(\cdot), \beta_j^\ast ; s, r \), \( \rho_-(\mathbb{E}\{\nabla L_j(\cdot)\}, \beta_j^\ast ; s, r \), \( \rho_+(\nabla^2 L_j(\cdot), \beta_j^\ast ; s, r \) and \( \rho_+(\mathbb{E}\{\nabla L_j(\cdot)\}, \beta_j^\ast ; s, r \) as \( \rho_{j,-}(s), \rho_{j,-}(s), \rho_{j,+}(s) \) and \( \rho_{j,+}(s) \) respectively.

Proof of Lemma 4.7. See §A.3 for a detailed proof.

The significance of Lemma 4.7 is that given the sparse eigenvalue condition on \( \mathbb{E}\{\nabla^2 L_j(\cdot)\} \) (Assumption 4.4), similar condition will also hold for \( \nabla^2 L_j(\cdot) \) with high probability when \( n \) is sufficiently large. Because we have \( \lim_{n \to \infty} k^\ast(\log^9 d/n)^{1/2} = 0 \) under Assumption 4.4 , which implies \( \lim_{n \to \infty} s^\ast(\log^9 d/n)^{1/2} = 0 \), we can always select a sufficiently large \( n \) such that \( \rho_\ast \gtrsim s^\ast(\log^9 d/n)^{1/2} \), then Lemma 4.7 implies that we have sparse eigenvalue condition for \( \nabla^2 L_j(\cdot) \) in an \( \ell_1 \)-ball centered at \( \beta_j^\ast \) with high probability.

Now we are ready to present the main theorem of this section. Recall that the penalty function \( p_\lambda(u) \) satisfy conditions (C.1), (C.2) and (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 4.8 (\( \ell_2 \) and \( \ell_1 \)-rates of convergence). Under Assumption 4.1 and 4.4, we let \( K_1 \) and \( K_2 \) be the constants defined in Theorem 4.2 and also let \( \rho_\ast > 0 \) and \( r > 0 \) be defined in Assumption 4.4. For all \( j \in [d] \), we define the support of \( \beta_j^\ast \) as \( S_j := \{(j,k) : \beta_{jk}^\ast \neq 0, k \in [d]\} \) and let \( s^\ast = \max_{j \in [d]} \|\beta_j^\ast\|_0 \). The penalty function \( p_\lambda(u) : [0, +\infty) \to [0, +\infty) \) in (3.4) satisfies conditions (C.1), (C.2) and (C.3) listed in §3.2 with \( c_1 = 0.91 \) and \( c_2 \geq 24/\rho_\ast \) for condition (C.3). We set the regularization parameter \( \lambda = C(\log^5 d/n)^{1/2} \) with \( C \geq 25K_1 \) and \( \lambda \leq r\rho_\ast/(33s^\ast) \). We denote constants \( A_1 = 22/\rho_\ast, A_2 = 2.2c_2, B_1 = 32/\rho_\ast, B_2 = 3.2c_2, \gamma = 11c_2^{-1}\rho_\ast^{-1} < 1 \) and define \( Y_j := \left(\sum_{(j,k) \in S_j} p_\lambda(\|\beta_{jk}^\ast\| - c_2\lambda)^{1/2}\right. :) \). Then the
following results hold with probability at least 1 \( -d^{-1} \):
\[
\|\hat{\beta}^{(\ell)}_j - \beta^*_j\|_2 \leq A_1 \left( \sqrt{s^* \log^5 n/n + Y_j} \right) + A_2 \sqrt{s^* \lambda^\ell} \quad \text{and} \quad (4.3)
\]
\[
\|\hat{\beta}^{(\ell)}_j - \beta^*_j\|_1 \leq B_1 \sqrt{s^*} \left( \sqrt{s^* \log^5 n/n + Y_j} \right) + B_2 s^* \lambda^\ell. \quad (4.4)
\]

**Proof of Theorem 4.8.** See §A.4 for a detailed proof.

**Remark 4.9.** In Theorem 4.8 we assume that \( \lambda \leq r \rho_*/(33s^*) \). To justify this assumption, by Assumption 4.4, \( \lim_{n \to \infty} s^*(\log^5 d/n)^{1/2}/r = 0 \), which indicates that \( \lim_{n \to \infty} s^* \lambda/r = 0 \); hence \( \lambda \leq r \rho_*/(33s^*) \) is satisfied when \( n \) is sufficiently large. In addition, Theorem 4.2 suggests that the regularization parameter \( \lambda \asymp (\log^5 d/n)^{1/2} \gtrsim \max_{j \in [d]} \|\nabla L_j(\beta^*_j)\|_\infty \) with high probability.

From Theorem 4.8 we see that the estimation error is dominated by the second term if \( p'_\lambda(|\beta^*_j| - c_2 \lambda) \) is not negligible. If the signal strength is large enough such that \( p'_\lambda(\beta - c_2 \lambda) = 0 \) where \( \beta = \min_{(j,k) \in S_j} |\beta^*_j| \), after sufficient number of iterations, the \( \ell_2 \) and \( \ell_1 \)-rates will be of order
\[
\|\hat{\beta}^{(\ell)}_j - \beta^*_j\|_2 = O\left( \sqrt{s^* \log^5 n/n} \right) \quad \text{and} \quad \|\hat{\beta}^{(\ell)}_j - \beta^*_j\|_1 = O\left( s^* \sqrt{\log^5 n/n} \right).
\]
However, if the signals are uniformly small such that \( p'_\lambda(|\beta^*_j| - c_2 \lambda) > 0 \) for all \((j, k) \in S_j\), the rates of convergence will be of order
\[
\|\hat{\beta}^{(\ell)}_j - \beta^*_j\|_2 = O\left( s^* \right) \quad \text{and} \quad \|\hat{\beta}^{(\ell)}_j - \beta^*_j\|_1 = O\left( s^* \right),
\]
which are identical to the \( \ell_2 \) and \( \ell_1 \)-rates of convergence of LASSO estimator respectively. Therefore, \( c_2 \lambda \) can be viewed as the threshold of signal strength. Combining with Theorem 4.2, for the \( \ell_2 \) and \( \ell_1 \)-rates of convergence of \( \hat{\beta}^{(\ell)}_j \), it holds with high probability that
\[
\|\hat{\beta}^{(\ell)}_j - \beta^*_j\|_2 \leq A_1 \left\{ \|\nabla S_j L_j(\beta^*_j)\|_2 + Y_j \right\} + A_2 \sqrt{s^*} \gamma^\ell \quad \text{and} \quad (4.5)
\]
\[
\|\hat{\beta}^{(\ell)}_j - \beta^*_j\|_1 \leq B_1 \sqrt{s^*} \left\{ \|\nabla S_j L_j(\beta^*_j)\|_2 + Y_j \right\} + B_2 s^* \gamma^\ell, \quad (4.6)
\]
where \( S_j = \{ (j, k) : \beta^*_j \neq 0, k \in [d] \} \) is the support of \( \beta^*_j \). Therefore the estimators \( \hat{\beta}_j \) obtained from the multi-stage convex relaxation algorithms attains the following more refined rates of convergence:
\[
\|\hat{\beta}_j - \beta^*_j\|_2 = O\left( \|\nabla S_j L_j(\beta^*_j)\|_2 + Y_j \right) \quad \text{and} \quad \|\hat{\beta}_j - \beta^*_j\|_1 = O\left( \sqrt{s^*} \right).\]
We claim that the sparsity level \( s^* \) in (4.5) and (4.6) can be changed to the sparsity level of each \( \beta^*_j \). Let \( s^*_j = \|\beta^*_j\|_0 \) be the sparsity level of \( \beta^*_j \) and \( \lambda_j \) be the regularization parameter for optimization problem (3.4) such that \( \lambda_j \gtrsim \|\nabla L_j(\beta^*_j)\|_\infty \). Let \( \bar{\beta}^{(\ell)}_j \) be the solution of the \( \ell \)-th convex relaxation problem (3.5) of Algorithm 1, the \( \ell_2 \) and \( \ell_1 \)-rates of convergence for each \( \bar{\beta}^{(\ell)}_j \) can be improved to
\[
\|\bar{\beta}^{(\ell)}_j - \beta^*_j\|_2 = O(\sqrt{s^*_j \lambda_j}) \quad \text{and} \quad \|\bar{\beta}^{(\ell)}_j - \beta^*_j\|_1 = O(s^*_j \lambda_j).\]
We use the uniform sparsity level \( s^* = \max_{j \in [d]} s^*_j \) and the same regularization parameter \( \lambda \) for simplicity, but the proof can be easily adapted to individual \( s^*_j \) and \( \lambda_j \) for each \( j \in [d] \).

The next theorem shows that if there are both strong and weak signals in \( \beta^*_j \), we are able to derive a more refined rates of convergence for \( \hat{\beta}^{(\ell)}_j \), which shows that NOSI method reaches the optimal rates of convergence for parameter estimation.

**Theorem 4.10** (Refined Rates of Convergence). Under Assumption 4.1 and 4.4, we let \( K_1 \) and \( K_2 \) be the constants defined in Theorem 4.2 and also let \( \rho_s > 0 \) and \( r > 0 \) be defined in Assumption 4.4. For all \( j \in [d] \), we define the support of \( \beta^*_j \) as \( S_j := \{ (j, k) : \beta^*_j \neq 0, k \in [d] \} \) and let \( s^* = \max_{j \in [d]} \|\beta^*_j\|_0 \). The penalty function \( p_\lambda(u) : [0, +\infty) \rightarrow [0, +\infty) \) in (3.4) satisfies regularity conditions (C.1), (C.2) and (C.3) listed in §3.2 with \( c_1 = 0.91 \) and \( c_2 \geq 24/\rho_s \) for condition (C.3). We set the regularity parameter \( \lambda = C(\log^5 d/n)^{1/2} \) such that \( C \geq 25K_1 \) and \( \lambda \leq r\rho_s/(33s^*) \). Moreover, we assume that the penalty function \( p_\lambda(u) \) satisfies an extra condition (C.4): there exist a constant \( c_3 > 0 \) such that \( p'_\lambda(u) = 0 \) for \( u \in [c_3\lambda, +\infty) \). Suppose that the support of \( \beta^*_j \) can be partitioned into \( S_j = S_{1j} \cup S_{2j} \) where \( S_{1j} = \{ (j, k) : |\beta_{jk}| \geq (c_2 + c_3)\lambda \} \) and \( S_{2j} = S_j - S_{1j} \). We denote constants \( A_1 = 22/\rho_s, A_2 = 2.2c_2, B_1 = 32/\rho_s, B_2 = 3.2c_2, \gamma = 11c_2^{-1}\rho_s^{-1} < 1 \) 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:

\[
\|\hat{\beta}^{(\ell)}_j - \beta^*_j\|_2 \leq A_1 \left\{ \|\nabla S_{1j} L_j(\beta^*_j)\|_2 + a\sqrt{s^*_{2j}}\lambda \right\} + A_2\sqrt{s^*}\lambda^\ell \quad \text{and} \quad (4.7)
\]

\[
\|\hat{\beta}^{(\ell)}_j - \beta^*_j\|_1 \leq B_1 \left\{ \|\nabla S_{1j} L_j(\beta^*_j)\|_2 + a\sqrt{s^*_{2j}}\lambda \right\} + B_2 s^*\lambda^\ell. \quad (4.8)
\]

**Proof of Theorem 4.10.** See §A.5 for a detailed proof.

We note that the \( \ell_2 \)-rate of convergence (4.7) matches the optimal rate of convergence for parameter estimation discussed in Wang et al. (2014) (Theorem 4.8), therefore we claim that NOSI method attains optimality for parameter estimation.

### 4.2 A Theoretical Analysis of Pairwise Score Test

In our pairwise score test for the null hypothesis \( H_0 : \beta^*_{jk} = 0 \), we combine loss functions \( L_j(\cdot) \) and \( L_k(\cdot) \) together because \( \beta_{jk} \) appears in both \( L_j(\beta_j) \) and \( L_k(\beta_k) \) (recall that we use \( \beta_{jk} \) and \( \beta_{kj} \) interchangeably). Our test statistic is constructed using the gradient of \( L_j(\cdot) \) and \( L_k(\cdot) \) and can be viewed as a high-dimensional analogue of the classical score test in low-dimensional settings.

Recall that we define the pairwise score function \( S_{jk}(\beta^*_{j\ell}) \) and the pairwise score statistic \( \hat{S}_{jk} \) in (3.6) and (3.8) respectively. Both of them involve linear combinations of entries of \( \nabla L_j(\cdot) \) and \( \nabla L_k(\cdot) \). To characterize the limiting distribution of \( \hat{S}_{jk} \), we first derive the limiting distribution of \( S_{jk} \) and then show that \( \hat{S}_{jk} \) converges to \( S_{jk}(\beta^*_{j\ell}) \) in probability. For the asymptotic distribution of \( S_{jk}(\beta^*_{j\ell}) \), which involves both \( \nabla L_j(\beta^*_j) \) and \( \nabla L_k(\beta^*_k) \), we study the asymptotics of this pair of score functions together.
4.2.1 The Limiting Distribution the Score function

According to a pair of nodes \( j, k \in [d] \), entries in \( \beta_j \) and \( \beta_k \) can be categorized into three types: (i) \( \beta_{jk} \), (ii) \( \beta_{j\setminus k} = (\beta_{j\ell}; \ell \neq k)^T \) and (iii) \( \beta_{k\setminus j} = (\beta_{k\ell}; \ell \neq j)^T \). If we let

\[
L_{jk}(\beta_{jk}, \beta_{j\setminus k}, \beta_{k\setminus j}) := L_j(\beta_j) + L_k(\beta_j),
\]

then \( S_{jk} \) is a linear transformation of \( \nabla L_{jk}(\beta_{jk}^*, \beta_{j\setminus k}, \beta_{k\setminus j}) \). For notational simplicity, we denote \( \beta_{j\vee k} := (\beta_{jk}, \beta_{j\setminus k}^T, \beta_{k\setminus j}^T)^T \). By definition, the gradient of \( L_{jk}(\cdot) \) evaluated at \( \beta_{j\vee k} \) is given by

\[
\nabla j_k L_{jk}(\beta_{j\vee k}) = \nabla j_k L_j(\beta_j) + \nabla k_j L_k(\beta_k); \quad \nabla j_k L_{jk}(\beta_{j\vee k}) = \nabla j_k L_j(\beta_j) \quad \text{and} \quad \nabla k\vee j L_{jk}(\beta_{j\vee k}) = \nabla k\vee j L_k(\beta_k).
\]

By the definition of \( L_{jk} \), \( \nabla L_{jk}(\cdot) \) is a second-order U-statistic. We denote the kernel function of \( \nabla L_j(\beta_j) \), \( \nabla L_k(\beta_k) \) and \( \nabla L_{jk}(\beta_{jk}) \) as \( h^j_{ij}(\beta_j) \), \( h^k_{ij}(\beta_k) \) and \( h^{jk}_{ij}(\beta_{jk}) \) respectively, where the subscripts of these kernel functions indicates that they are constructed using \( X_i \) and \( X_i' \). As shown in Ning and Liu (2014), \( \mathbb{E}\{h^j_{ij}(\beta_j)\} = \mathbb{E}\{h^k_{ij}(\beta_k)\} = 0 \); hence \( h^{jk}_{ij}(\beta_{jk}) \) is also centered. Thanks to the U-statistic structure, we can characterize the limiting distribution of \( \nabla L_{jk}(\beta_{jk}^*) \) using the method of Hájek projection (Van der Vaart, 2000; DasGupta, 2008), which approximates a U-statistic with a sum of i.i.d. random variables. We define

\[
g_{jk}(X_i) := \frac{n}{2} \mathbb{E}\{\nabla L_{jk}(\beta_{jk}^*) \mid X_i\} = \mathbb{E}\{h^{jk}_{ij}(\beta_{jk}^*) \mid X_i\} \quad \text{and} \quad (4.9)
\]
\[
U_{jk} := \frac{2}{n} \sum_{i=1}^{n} g_{jk}(X_i) = \sum_{i=1}^{n} \mathbb{E}\{\nabla L_{jk}(\beta_{jk}^*) \mid X_i\}. \tag{4.10}
\]

Thus \( 2g_{jk}(X_i)/n \) is the projection of \( \nabla L_{jk}(\beta_{jk}^*) \) onto the \( \sigma \)-filed generated by \( X_i \) and \( U_{jk} \) is the Hájek projection of \( \nabla L_{jk}(\beta_{jk}^*) \). We show that \( U_{jk} \) is a good approximation of \( \nabla L_{jk}(\beta_{jk}^*) \) and enables us to characterize the limiting distribution of \( \nabla L_{jk}(\beta_{jk}^*) \). Before stating the main results, we first present an assumption that guarantees that \( g_{jk}(X_i) \) is not degenerate.

**Assumption 4.11.** Under Assumption 4.1, for \( g_{jk}(X_i) \) defined in (4.9), we denote the covariance matrix of \( g_{jk}(X_i) \) as \( \Sigma_{jk} := \mathbb{E}\{g_{jk}(X_i)g_{jk}(X_i)^T\} \). We assume that there exist a constant \( c_\Sigma > 0 \) such that \( \lambda_{\min}(\Sigma_{jk}) \geq c_\Sigma \) for all \( 1 \leq j < k \leq d \).

Assumption 4.11 requires that the minimum eigenvalue of \( \Sigma_{jk} \) to be bounded away from 0; which means that \( g_{jk}(X_i) \) is non-degenerate in the sense that for any \( v \in \mathbb{R}^{2d-3} \), \( ||v||_2 = 1 \) \( \text{Var}(v^T U_{jk}) \neq 0 \). This assumption is standard for the existence of asymptotic variance for U-statistics in low-dimensional models.

Our next lemma characterizes the limiting distribution of \( \nabla L_{jk}(\beta_{jk}^*) \) and is pivotal for establishing the asymptotic normality of \( \tilde{S}_{jk} \).

**Lemma 4.12.** For any \( b \in \mathbb{R}^{2d-3} \) with \( ||b||_2 = 1 \) and \( ||b||_0 \leq \tilde{s} \), if \( \lim_{n \to \infty} \tilde{s}/n = 0 \), we have

\[
\frac{\sqrt{n}}{2} b^T \nabla L_{jk}(\beta_{jk}^*) \sim N(0, b^T \Sigma_{jk} b). \tag{4.11}
\]
Proof of Lemma 4.12. See §B.1 for a detailed proof.

Lemma 4.12 shows that the asymptotic distribution of $\nabla L_{jk}(\beta^*_{jk})$ resembles a multivariate Gaussian distribution in some sense because the inner product of $\nabla L_{jk}(\beta^*_{jk})$ and any sparse vector converges to a univariate Gaussian distribution. This property is pivotal in our NOSI method because as we will see, it enables us to derive the asymptotic distribution of $\hat{S}_{jk}$.

Note that $b^T \nabla L_{jk}(\beta^*_{jk})$ is a univariate U-statistic with kernel $b^T h_{iv}^k(\beta^*_{jk})$. We denote $\Theta_{jk} := \mathbb{E}\{h_{iv}^k(\beta^*_{jk}) h_{iv}^k(\beta^*_{jk})^T\}$. By Hoeffding’s theorem, the variance of $b^T \nabla L_{jk}(\beta^*_{jk})$ is given by

$$\text{Var}(b^T \nabla L_{jk}(\beta^*_{jk})) = \frac{2}{n(n-1)} (2(n-2)\mu_1 + \mu_2),$$

where $\mu_1 := b^T \Sigma^k b$ and $\mu_2 := b^T \Theta_{jk} b$. Moreover, the variance of $b^T U_{jk}$ is $\text{Var}(b^T U_{jk}) = 4\mu_1/n$. Theory of Hájek projection implies that $\sqrt{n} b^T \nabla L_{jk}(\beta^*_{jk})$ and $\sqrt{n} b^T U_{jk}$ has the same asymptotic distribution if

$$\lim_{n \to \infty} \text{Var}(b^T \nabla L_{jk}(\beta^*_{jk})) / \text{Var}(b^T U_{jk}) = 1 + \lim_{n \to \infty} \mu_2/(2n) = 1.$$

Therefore we assume the sparsity of $b$ to be $o(n)$ to guarantee that $\mu_2/n \to 0$, which indicates the validity of Hájek projection in our high-dimensional setting.

4.2.2 Validity of Pairwise Score Test

We prove the validity of the pairwise score test. Let $\hat{\beta}_j$ and $\hat{\beta}_k$ be the estimators of $\beta^*_{j\cdot}$ and $\beta^*_{k\cdot}$ obtained from Algorithm 1; we assume that $H^j = \mathbb{E}[\nabla^2 L_j(\beta_{j\cdot})]$ is invertible and denote $w^*_{j,k} = H^j_{k,j\setminus k}(H^j_{j\setminus k,k\setminus j})^{-1}$. Note that we can write the pairwise score function $S_{jk}(\cdot)$ and the test statistic $\hat{S}_{jk}$, which are defined in (3.6) and (3.8) respectively, as

$$S_{jk}(\beta_{j\cdot}) = \nabla_{jk} L_{jk}(\beta_{j\cdot}) - w^*_{j,k} \nabla_{j\setminus k} L_{jk}(\beta_{j\cdot}) - w^*_{k,j} \nabla_{k\setminus j} L_{jk}(\beta_{j\cdot}) \quad \text{and} \quad \hat{S}_{jk} = \nabla_{jk} \hat{L}_{jk}(\beta^*_{j\cdot}) - \hat{w}^*_{j,k} \nabla_{j\setminus k} \hat{L}_{jk}(\beta^*_{j\cdot}) - \hat{w}^*_{k,j} \nabla_{k\setminus j} \hat{L}_{jk}(\beta^*_{j\cdot});$$

where $\hat{\beta}^*_{j\cdot} := (0, \hat{\beta}^T_{j\setminus k}, \hat{\beta}^T_{k\setminus j})^T$, $\hat{w}_{j,k}$ is obtained from the Dantzig-type problem (3.7) and $\hat{w}_{k,j}$ can be obtained similarly by solving

$$\hat{w}_{k,j} = \text{argmin} \|w\|_1 \quad \text{such that} \quad \|\nabla^2_{k,j\setminus k} L_k(0, \hat{\beta}^*_{k\setminus j}) - w^T \nabla^2_{k,j\setminus k} L_k(0, \hat{\beta}^*_{k\setminus j})\|_\infty \leq \lambda_D. \quad (4.15)$$

To derive the asymptotic distribution of $\hat{S}_{jk}$, we show that $\sqrt{n}(\hat{S}_{jk} - S_{jk}(\beta^*_{j\cdot})) = o_p(1)$ and obtain the limiting distribution of $S_{jk}(\beta^*_{j\cdot})$ from Lemma 4.12. We first present the required assumptions for the validity of our derivations.

Assumption 4.13. We assume that for all $j \in [d]$, $H^j$ is invertible. And for all $j,k$ satisfying $1 \leq j \neq k \leq d$, we denote $w^*_{j,k} = H^j_{j\setminus k,k\setminus j}(H^j_{j\setminus k,k\setminus j})^{-1}$. In addition, we assume that there exist an integer $s^*$ and a positive number $w_0$ such that $\|w^*_{j,k}\|_0 \leq s^* - 1$ and $\|w^*_{j,k}\|_1 \leq w_0$ with $s^* = o(n)$. The regularization parameter $\lambda_D$ in the Dantzig selector problem (3.7) satisfies $\lambda_D \gtrsim w_0 s^* \log^2 d$. Moreover, we assume that

$$\lim_{n \to \infty} \{(1 + w_0 + w_0^2) (s^* \lambda \log^8 d + s^* \lambda D)\} = 0 \quad \text{and} \quad \lim_{n \to \infty} \sqrt{n} (s^* + s) \lambda \lambda D = 0. \quad (4.16)$$
If we can treat \( w_0 \) as a constant, then Assumption 4.13 is reduced to \( \lambda_D \gtrsim s^* \lambda \log^2 d, s^* \lambda_D + s^* \lambda \log^8 d = o(1) \) and \((s^*+s^*) \lambda_D = o(n^{1/2})\). Note that \( \lambda \propto \sqrt{\log^5 d/n} \), we can choose \( \lambda_D = Cs^* \sqrt{\log^9 d/n} \) with a sufficiently large \( C \), provided \( s^* \sqrt{\log^2 d/n} = o(1) \) and \((s^*+s^*) \log^7 d/n = o(n^{-1/2})\). Hence the condition is fulfilled if \( \log d = o\left(\min\left\{\left(\frac{\sqrt{n}}{s^*}\right)^{2/21}, \left(\frac{\sqrt{n}}{s^*}\right)^{1/7}, \left(\frac{\sqrt{n}}{s^*}\right)^{1/7}\right\}\right) \).

The following assumption assumes the sparse eigenvalue condition for \( \nabla^2 L(\beta_j) \), which is similar to Assumption 4.4. Such condition is essential for the consistency of \( \hat{\beta}_{jk} \).

**Assumption 4.14.** We assume that the Assumptions 4.1, 4.4 and 4.13 hold. We denote the \( s \)-sparse eigenvalues \( \rho_{-} (\mathbb{E}\{\nabla^2 L_j(\cdot)\}, \beta^*_j; s, r) \) and \( \rho_{+} (\mathbb{E}\{\nabla^2 L_j(\cdot)\}, \beta^*_j; s, r) \) as \( \rho^*_-(s) \) and \( \rho^*_+(s) \) respectively, where the radius \( r \) is chosen to be the same as that in Assumption 4.4. Let \( s^* \) be defined in Assumption 4.13, we further assume that there exist an integer \( k^* \geq s^* \) and a positive number \( \tau_s \) such that \( \lim_{n \to \infty} k^* (\log^9 d/n)^{1/2} = 0 \) and

\[
0 < \tau_s \leq \rho^*_-(s^*+k^*) < \rho^*_+(k^*) \leq (1 + 0.5k^*/s^*)\tau_s, \quad 1 \leq j, k \leq d. \tag{4.17}
\]

Comparing with Assumption 4.4, this condition also requires that the \( \rho^*_+(s)/\rho^*_-(s) \) grows sub-linearly in \( s \). This assumption is pivotal for the analysis of the Dantzig-type estimator \( \hat{\beta}_{jk} \).

The following lemma, similar to Lemma 4.7, shows that the sparse eigenvalue condition also hold for \( \nabla^2 L_j(\cdot) \) with high probability for sufficiently large \( n \), which enables us to derive the rate of convergence for \( \hat{\beta}_{jk} \).

**Lemma 4.15.** Under the Assumptions 4.1, 4.4, 4.13 and 4.14, we denote the \( s \)-sparse eigenvalues of \( \nabla^2 L_j(\cdot) \) over the \( \ell_1 \)-ball centered at \( \beta^*_j \) with radius \( r \) as \( \rho^-_j(s) \) and \( \rho^+_j(s) \) respectively where \( r \) is defined in Assumption 4.4. If the number of observations \( n \) is sufficiently large such that \( \nu_s \gtrsim s^*(\log^9 d/n)^{1/2} \), with probability at least \( 1 - d^{-1} \), we have

\[
0.95 \tau_s \leq \rho^-_j(s^*+k^*) \leq \rho^+_j(k^*) \leq (1 + 0.5k^*/s^*)\tau_s + 0.05\tau_s,
\]

where the \( k^* \) and \( \nu_s \) the same as in Assumption 4.14. Moreover, we have

\[
\rho^+_j(k^*)/\rho^-_j(s^*+k^*) \leq 1 + 0.58k^*/s^*.
\]

**Proof of Lemma 4.15.** The proof is similar to that of Lemma 4.7, see §B.2 for a detailed proof.

Now we present the main theorem of pairwise score test.

**Theorem 4.16.** Under the Assumptions 4.1, 4.4, 4.13 and 4.14, for \( j, k \in [d] \) and \( j \neq k \), we denote \( \beta_{j\setminus k} = (\beta_{jk}, \ell \neq k)^T \), \( \beta_{k\setminus j} = (\beta_{kj}, \ell \neq j)^T \) and \( \beta_{jk\setminus k} = (\beta_{jk}, \beta_{jk}^T, \beta_{jk}^T)^T \). The pairwise score statistic and pairwise score function \( \hat{S}_{jk} \) and \( S_{jk} (\beta_{jk\setminus k}) \) are defined in (3.8) and (3.6) respectively. Then it holds uniformly for all \( j \neq k \) and \( j, k \in [d] \) that \( \sqrt{n} \hat{S}_{jk} = \sqrt{n} S_{jk} (\beta_{jk\setminus k}^*) + o_P(1) \). Furthermore, under the null hypothesis \( H_0: \beta_{jk}^* = 0 \), \( \sqrt{n} \hat{S}_{jk} \) converges to a Gaussian random variable in distribution:

\[
\sqrt{n} \hat{S}_{jk}/2 \sim N(0, \sigma^2_{jk}),
\]

where the asymptotic variance \( \sigma^2_{jk} \) is given by

\[
\sigma^2_{jk} = \Sigma_{jk, jk}^{jk} - 2\Sigma_{jk, jk}^{jk} \mathbf{w}_{j,k}^* - 2\Sigma_{jk, k\setminus j}^{jk} \mathbf{w}_{k,j}^* + \mathbf{w}_{j,k}^T \Sigma_{jk, k\setminus j}^{jk} \mathbf{w}_{j,k}^* + \mathbf{w}_{k,j}^T \Sigma_{j\setminus k, k\set\setminus k}^{jk} \mathbf{w}_{k,j}^*.
\]

19
Proof of Theorem 4.16. See §B.3 for a detailed proof.

By Theorem 4.16, to test the null hypothesis $H_0: \beta_{jk}^* = 0$ against the alternative hypothesis $H_1: \beta_{jk}^* \neq 0$, we reject $H_0$ if $\sqrt{n}S_{jk}$ is too extreme. If we have a consistent estimator of the asymptotic variance $\sigma_{jk}^2$, denoted as $\hat{\sigma}_{jk}^2$, we construct a studentized test statistic $\sqrt{n}S_{jk}/(2\hat{\sigma}_{jk})$. The test function of pairwise score test with significance level $\alpha$ is given by

$$
\psi_{jk}(\alpha) = \begin{cases} 
1 & \text{if } |\sqrt{n}S_{jk}/(2\hat{\sigma}_{jk})| > \Phi^{-1}(1-\alpha/2) \\
0 & \text{if } |\sqrt{n}S_{jk}/(2\hat{\sigma}_{jk})| \leq \Phi^{-1}(1-\alpha/2),
\end{cases}
$$

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

For a consistent estimator of $\sigma_{jk}^2$, we first define an estimator of $\Sigma_{jk}$ as

$$
\hat{\Sigma}_{jk}(\beta_{j\check{k}}) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{1}{n-1} \sum_{i' \neq i} h_{ii'}(\beta_{j\check{k}}) \right\}^{\otimes 2},
$$

where $h_{ii'}(\beta_{j\check{k}})$ is the kernel function of the second-order U-statistic $\nabla L_{jk}(\beta_{j\check{k}})$. Let $\hat{\beta}_{j\check{k}} := (0, \hat{\beta}_{j\check{k}}^{(1)}, \hat{\beta}_{j\check{k}}^{(2)})^T$ and $\hat{\Sigma}_{jk} := \hat{\Sigma}_{jk}(\hat{\beta}_{j\check{k}})$, a consistent estimator of $\sigma_{jk}^2$ can be constructed as follows:

$$
\hat{\sigma}_{jk}^2 := \hat{\Sigma}_{jk} = 2\hat{\Sigma}_{j,k;j,k} - 2\hat{\Sigma}_{j,k,j\check{k}} \hat{w}_{j,k} - 2\hat{\Sigma}_{j,k,k\check{j}} \hat{w}_{k,j} + \hat{w}_{j,k}^T \hat{\Sigma}_{j,k,j\check{k}} \hat{w}_{j,k} + \hat{w}_{k,j}^T \hat{\Sigma}_{j,k,k\check{j}} \hat{w}_{k,j}.
$$

The next corollary shows that under $H_0$, the type I error of $\psi_{jk}(\alpha)$ converges to the desired significance level and the associated p-value is asymptotically uniformly distributed in $[0, 1]$.

Corollary 4.17. Let the test function of pairwise score test with significance level $\alpha$ be defined in (4.18) and we reject the null hypothesis if $\psi_{jk}(\alpha) = 1$. The associated p-value is defined as $p_{\psi}^{jk} := 2\left(1 - \Phi\left(|\sqrt{n}S_{jk}/(2\hat{\sigma}_{jk})|\right)\right)$. Under the Assumptions 4.1, 4.4, 4.13 and 4.14, we have

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

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

Proof of Corollary 4.17. See §B.4 for a detailed proof.

4.3 Supergraph Testing via Gaussian Multiplier Bootstrap

We present the theoretical aspects of our super-graph test which tests whether a given graph contains the true graph. Let $G = (V, E)$ be a given graph and $G^* = (V, E^*)$ be the true graph structure of the graphical model, we test the null hypothesis $H_0: G^* \subseteq G$. We define the complement of the edge set $E$ as $E^c := \{(j,k): 1 \leq j, k \leq d, (j,k) \notin E\}$. As described in §3, we reject the null hypothesis if and only if we reject the hypothesis $H_0$: $\beta_{jk}^* = 0$ for some $(j,k) \in E^c$.

Because of such relationship between the global inference and the local inference, the test statistic for super-graph test, $T = \sqrt{n} \max_{E^c} |S_{jk}|/2$, is constructed based on pairwise score statistics $\hat{S}_{jk}$. We denote the $\alpha$-th quantile of $T$ as $c_T(\alpha)$, which is defined as $c_T(\alpha) = \inf\{t \in \mathbb{R}: \mathbb{P}(T \leq t) \leq \alpha\}$. Given quantiles $\{c_T(\alpha), \alpha \in [0,1]\}$, the test that rejects the null hypothesis when $T$ exceeds $c_T(1-\alpha)$
has significance level $\alpha$ asymptotically. As described in §3, we use the method of Gaussian multiplier bootstrap to obtain the quantiles of $T$. Recall that we write $\hat{S}_{jk}$ as a summation in (3.10) and define its bootstrap counterpart $\hat{S}_{jk}^B$ in (3.11). Given $n$ i.i.d. standard normal random variables $\xi_1, \ldots, \xi_n$, we use the conditional quantiles of $W := \sqrt{n}\max_{E^c}\hat{S}_{jk}^B/2$ to approximate $c_T(\alpha)$. The $\alpha$-th conditional quantile of $W$ given $\xi_1, \ldots, \xi_n$ is defined as $c_W(\alpha) := \inf\{t \in \mathbb{R}: \mathbb{P}_\xi(W \leq t) \leq \alpha\}$ where $\mathbb{P}_\xi$ is the probability measure induced by $\xi_1, \ldots, \xi_n$. In what follows, we first state the regularity condition for the validity of our method and then present the main result.

**Assumption 4.18.** Let $s^*$ be defined in Assumption 4.13 and $s^* = \max_{j \in [d]} \|\beta_j^*\|_0$. Under Assumption 4.13, we further assume that $w_0$ in Assumption 4.13 is bounded and that there exist a constant $c > 0$ such that $\sqrt{n}\log d(s^* + s^*)\lambda \lambda_D = O(n^{-c})$ and $\log^{13}d/n = O(n^{-c})$.

The following Theorem shows that the quantiles of $W$ computed conditional on $\xi_1, \ldots, \xi_n$ well approximate the quantile of $T$ uniformly.

**Theorem 4.19.** Let $G = (V, E)$ and $G^* = (V, E^*)$ be any given graph and the true graph corresponding to the graphical model, respectively. We let $E^c = \{(j, k): (j, k) \notin E\}$ be the complement of the edge set of $E$. For any $(j, k) \in E^c$, we define the pairwise score statistic $\hat{S}_{jk}$ and its bootstrap counterpart $\hat{S}_{jk}^B$ in (3.8) and (3.11) respectively. Moreover, we define $T = \sqrt{n}\max_{E^c}\hat{S}_{jk}/2$ and $W := \sqrt{n}\max_{E^c}\hat{S}_{jk}^B/2$ as our super-graph test statistic and its Bootstrap counterpart. The conditional $\alpha$-quantiles of $W$ given $\xi_1, \ldots, \xi_n \sim N(0, 1)$ are denoted as $c_W(\alpha)$. We assume that Assumptions 4.1, 4.11, 4.13 and 4.18 hold. Under the null hypothesis $H_0: G^* \subset G$, there exist generic constants $c$ and $C$ such that

$$\sup_{\alpha \in [0, 1]} |\mathbb{P}(T \leq c_W(\alpha)) - \alpha| \leq Cn^{-c}. \quad (4.20)$$

We define the test function of super-graph test with significance level $\alpha$ as

$$\psi_G(\alpha) = \begin{cases} 1 & \text{if } T > c_W(1 - \alpha) \\ 0 & \text{if } T \leq c_W(1 - \alpha), \end{cases}$$

and reject $H_0$ if $\psi_G(\alpha) = 1$. Then our super-graph test is level-$\alpha$ significant asymptotically:

$$\lim_{n \to \infty} \sup_{\alpha \in [0, 1]} |\mathbb{P}(\psi_G(\alpha) = 1 \mid H_0) - \alpha| = 0.$$ 

**Proof of Theorem 4.19.** See §B.5 for a detailed proof.

## 5 Numerical Experiments

In this section we study the finite-sample performance of the proposed graph inference methods on both simulated and real-world datasets. More specifically, for numerical simulations, we examine the validity of the proposed methods on Gaussian, Ising, and mixed graphical models. In addition, we analyze the music annotation dataset for an application to real data.
5.1 Numerical Simulations

We first examine the numerical performance of local graph inference using the proposed pairwise score tests. Recall that local graph inference tests the null hypothesis $H_0: \beta_{jk}^* = 0$. 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 $|j-k| = 1, 2, d-2, d-1$. An illustration of a 4-nearest-neighbor graph with 25 nodes and $s^* = 4$ is shown in Figure 1(a). More specifically, we sample $X_1, \ldots, X_n$ from a Gaussian distribution $N_d(0, \Sigma)$. For the precision matrix $\Theta = \Sigma^{-1}$, we set $\Theta_{jj} = 1$, $|\Theta_{jk}| = \mu \in [0, 0.25]$ for $|j-k| = 1, 2, d-2, d-1$ and $\Theta_{jk} = 0$ for $2 \leq |j-k| \leq d-2$. Note $\mu$ denotes the signal strength of the graph inference problem and $\mu \leq 0.25$ ensures that $\Theta$ is diagonal dominant and invertible.

(ii) Ising graphical model. We set $n = 100$ and $d = 200$. The graph structure is a $10 \times 20$ grid; hence the sparsity level $s^* = 4$. An illustration of an $5 \times 5$ grid is shown in Figure 1(b). We use Markov Chain Monte Carlo method (MCMC) to simulate $n$ data from an Ising model with joint distribution $p(x) \propto \exp \left( \sum_{j \neq k} \beta_{jk}^* x_j x_k \right)$ (using the package IsingSampler). We set $|\beta_{jk}^*| = \mu \in [0, 1]$ if 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; hence 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. An illustration of a $4 \times 4 \times 2$ grid graph is shown in Figure 1(c) where we display the Gaussian nodes and the binomial nodes as blue and red circles respectively. We set $|\beta_{jk}^*| = \mu \in [0, 1]$ if if there exists an edge connecting node $j$ and node $k$, and $\beta_{jk}^* = 0$ otherwise. We simulate data by a two-step method proposed in Lee and Hastie (2014).

We denote the true parameters of the graphical models as $\{\beta_{jk}^*, j \neq k\}$. We also denote $\beta_j^* = (\beta_{j1}^*, \ldots, \beta_{jd}^*)^T$. For the Gaussian graphical model, we have $\beta_{jk}^* = \Theta_{jk}$. We first obtain a point
estimate of $\beta_j^*$ by solving (3.4) using Algorithm 1 using the capped-$\ell_1$ penalty $p_\lambda(u) = \lambda \min\{u, \lambda\}$. The parameter $\lambda$ is chosen by 10-fold cross validation as suggested by Ning and Liu (2014).

Recall that the form of the loss function $L_j(\beta_j)$ is exactly the loss function for logistic regression, where we use Rademacher random variables $y_{ii'}$ as response and $y_{ii'}(x_{ij} - x_{i'j})\beta_j^T(x_i - x_{i'})$ as covariates, Algorithm 1 can be easily implemented by using the $\ell_1$-regularized logistic regression such as the glmnet package. In particular, the algorithm converges in only a few iterations, indicating that it attains a good balance between computational efficiency and statistical accuracy.

Once $\hat{\beta}_j$ is obtained, we solve the Dantzig-type problem (3.7) using $\hat{\beta}_j$ as input. We set the regularization parameter $\lambda_D$ to be 0.2. However, we find that the final performance of the proposed method is not quite sensitive to $\lambda_D$.

For local inference, we compare the method of pairwise score test with the desparsity method proposed in van de Geer et al. (2014). Although the desparsity method is only proposed for hypothesis tests in generalized linear models (GLMs), it can be used for graphical models by performing nodewise regression. Their method constructs confidence intervals for each parameter of the GLMs; thus yields a hypothesis test by rejecting the null hypothesis $H_0: \beta_{jk}^* = 0$ if the confidence interval fails to cover 0. To examine the validity of our method, we test the null hypothesis $H_0: \beta_{jk}^* = 0$ for both the cases where the null hypothesis does and does not hold. We report in Tables 1, 2 and 3 the type I errors and the powers of the hypothesis tests at the 0.05 significance level using different choices of $\mu \in [0, 1]$. To obtain the type I error and power, we repeat the whole procedure for 500 times and use the rejection rate to evaluate type I error and power. As revealed by these tables, our method achieves accurate type I errors, which is comparable to the desparsity method. In terms the power of the test, our method is nearly as powerful as the desparsity method. Such phenomenon suggests that we trade little testing efficiency for model generality.

Table 1: Local inference results from Gaussian graphical models: Type I errors and Powers of the pairwise score test and the desparsity method (van de Geer et al., 2014) for $H_0: \beta_{jk}^* = 0$ at the 0.05-significance level. “Type-I” denotes the type I errors and $\mu$ is the signal strength.

| $\mu$ | Pairwise Score Test | Desparsity Method |
|-------|---------------------|-------------------|
|       | Type-I | Power | Type-I | Power |
| 1     | 0.00   | 0.051 | 0.051 | 0.049 | 0.049 |
| 2     | 0.03   | 0.051 | 0.068 | 0.055 | 0.078 |
| 3     | 0.06   | 0.053 | 0.100 | 0.050 | 0.118 |
| 4     | 0.09   | 0.056 | 0.159 | 0.052 | 0.188 |
| 5     | 0.12   | 0.056 | 0.242 | 0.051 | 0.284 |
| 6     | 0.15   | 0.052 | 0.432 | 0.055 | 0.496 |
| 7     | 0.18   | 0.054 | 0.623 | 0.050 | 0.691 |
| 8     | 0.21   | 0.053 | 0.724 | 0.052 | 0.781 |
| 9     | 0.24   | 0.054 | 0.793 | 0.047 | 0.861 |

We then examine the numerical performance of global graph inference using the multiplier bootstrap method. Recall that for global inference, we aim to test the null hypothesis $H_0: G^* \subset G$ for a given graph $G$. We set the true graph to be a 6-nearest neighbor graph and set the true
Table 2: Local inference results from Ising graphical models: Type I errors and Powers of the pairwise score test and the desparsity method (van de Geer et al., 2014) for $H_0: \beta_{jk}^* = 0$ at the 0.05-significance level. “Type-I” denotes the type I errors and $\mu$ is the signal strength.

| $\mu$ | Pairwise Score Test | Desparsity Method |
|-------|---------------------|-------------------|
|       | Type-I | Power | Type-I | Power |
| 1     | 0.00   | 0.056 | 0.056  | 0.049 | 0.049 |
| 2     | 0.10   | 0.062 | 0.101  | 0.054 | 0.088 |
| 3     | 0.20   | 0.062 | 0.229  | 0.514 | 0.200 |
| 4     | 0.30   | 0.050 | 0.446  | 0.055 | 0.533 |
| 5     | 0.40   | 0.051 | 0.636  | 0.495 | 0.679 |
| 6     | 0.50   | 0.045 | 0.726  | 0.048 | 0.764 |
| 7     | 0.60   | 0.055 | 0.799  | 0.053 | 0.826 |
| 8     | 0.70   | 0.052 | 0.866  | 0.049 | 0.872 |
| 9     | 0.80   | 0.056 | 0.924  | 0.053 | 0.919 |
| 10    | 0.90   | 0.052 | 0.955  | 0.056 | 0.943 |
| 11    | 1.00   | 0.053 | 0.974  | 0.050 | 0.959 |

parameters $|\beta_{jk}^*| = \mu \in [0, 1]$ if there is an edge connecting node $j$ and node $k$, and set $\beta_{jk}^* = 0$ otherwise. We simulate data from the Gaussian graphical model, Ising model and mixed graphical model with $n = 100$ and $d = 200$. For the mixed graphical model, we set the first 100 nodes to be binomial and the rest 100 nodes to be Gaussian. Using a significance level $\alpha = 0.05$, we test whether a given $k$-nearest neighbor graph of 200 nodes contains the true graph for $k = 0, 2, 4, 6, 8, 10$. We report in Tables 4, 5, and 6 the empirical rejection rates based on 500 independent trials. Notice that the probability of rejecting null hypothesis $H_0: G^* \subset G$ corresponds to the type I error for $k \geq 6$ and corresponds to the power of the test for $k \leq 4$, thus the empirical rejection rate gives us a Monte-Carlo approximation of the type I error and power of the test. As shown in Table 4, 5, and 6, the type I error of our super-graph test is approximately 0.05, which corroborates with our theoretical analysis.

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
Table 3: Local inference results from mixed graphical models: Type I errors and Powers of the pairwise score test and the desparsity method van de Geer et al. (2014) for $H_0: \beta^*_{jk} = 0$ at the 0.05-significance level. “Type-I” denotes the type I errors and $\mu$ is the signal strength.

| $\mu$ | Pairwise Score Test | Desparsity Method |
|-------|---------------------|--------------------|
|       | Type-I | Power   | Type-I | Power   |
| 1     | 0.00   | 0.057   | 0.057  | 0.054   | 0.054   |
| 2     | 0.03   | 0.052   | 0.065  | 0.054   | 0.057   |
| 3     | 0.06   | 0.056   | 0.104  | 0.052   | 0.128   |
| 4     | 0.09   | 0.053   | 0.133  | 0.050   | 0.147   |
| 5     | 0.12   | 0.050   | 0.219  | 0.046   | 0.253   |
| 6     | 0.15   | 0.051   | 0.398  | 0.059   | 0.425   |
| 7     | 0.18   | 0.058   | 0.592  | 0.052   | 0.589   |
| 8     | 0.21   | 0.058   | 0.687  | 0.054   | 0.729   |
| 9     | 0.24   | 0.053   | 0.753  | 0.050   | 0.786   |

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 feature extraction can be found in Tzanetakis and Cook (2002). In addition to these continuous variables, binary variables in the CAL500 dataset are represented by 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). In summary, the CAL500 dataset consists of 502 instances with 68 continuous variables and 174 binary variables. Our goal is to infer the association relationships between these different types of variables using semiparametric exponential family graphical models. This dataset has been analyzed in Cheng et al. (2013) 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 connection of each pair of nodes.

We model the CAL500 dataset using the semiparametric exponential family graphical model. Similar to Turnbull et al. (2008) and Cheng et al. (2013), 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. (2013), we keep all the binary labels. Thus the processed dataset has $n = 502$ data points of dimension $d = 226$. We apply the pairwise score test with $\alpha = 0.05$ significance level to each pair of variables to determine the presence of an edge between them. We use Bonferroni correction to avoid spurious positives. The null hypothesis is that these two variables are conditionally independence given the rest of variables. We set the nonconvex penalty function in optimization problem (3.4) to be capped-\ell_1 penalty $p_\lambda(u) = \lambda \min\{u, \lambda\}$ with the regularization parameter $\lambda$ selected by 10-fold cross-validation as in the previous section. For numerical stability, in our analysis, we perform the proposed test 100 times on randomly drawn subsamples of size $n/2$ and report the edges selected at least 90 times. We present the fitted graph in Figure 2, where we plot the connected components and omit the singletons. To better display
Table 4: Global inference results for Gaussian graphical models: Rejection rates of the super-graph test for $H_0: G^* \subset G$ at the 0.05-significance level. The true graph is a 6-nearest-neighbor graph and $G$ is set to be a $k$-nearest neighbor graph with $k = 0, 2, 4, 6, 8$ and 10. Nonzero parameters have the same absolute value $\mu$.

| $\mu$ | $k = 0$ | $k = 2$ | $k = 4$ | $k = 6$ | $k = 8$ | $k = 10$ |
|-------|---------|---------|---------|---------|---------|---------|
| 1     | 0.00    | 0.056   | 0.056   | 0.052   | 0.052   | 0.052   |
| 2     | 0.03    | 0.068   | 0.068   | 0.054   | 0.052   | 0.052   |
| 3     | 0.06    | 0.132   | 0.084   | 0.060   | 0.056   | 0.052   |
| 4     | 0.09    | 0.356   | 0.134   | 0.106   | 0.062   | 0.058   | 0.005   |
| 5     | 0.12    | 0.624   | 0.324   | 0.182   | 0.060   | 0.054   | 0.054   |
| 6     | 0.15    | 0.804   | 0.542   | 0.348   | 0.062   | 0.058   | 0.052   |

the graphical structure, we use a square to represent each type of 13 MFCC features respectively. If a node connects to any node within the group of variables in a MFCC node, we connect it with this MFCC node with an edge. We use circles to represent the binary variables and use different colors to indicate their categories. The obtained graph has some interesting properties. First of all, the continuous features are densely connected within themselves, which is similar to the results in Cheng et al. (2013). For connections between continuous and binary variables, we find that the noisiness of the music (square 4) is connected with not cheerful emotions (circle 16), rock music (circle 70), the instrument electric guitar (circle 79), songs are very catchy (circle 96, negatively correlated) and very tonic songs (circle 119, negatively correlated). We also find that both the average amplitude (square 1) and periodic amplitude variation (square 2) are connected with unpleasant songs (circle 30). Moreover, edges connecting two binary variables also display interesting patterns. For instance, we find that passionate emotions (circle 17) are connected with not passionate emotions (circle 18), soft rock music (circle 143), and emotional vocals (circle 58); very danceable songs (circle 122) are connected with the usage “at a party” (circle 123) and songs with fast tempo (circle 100); the songs with heavy beats (circle 103) are connected with folk songs (circle 65), and not tender emotions (circle 38). In addition, we find edges between songs with positive feelings (circle 108) and the carefree (circle 13), cheerful and festive (circle 15), happy (circle 21) and optimistic (circle 31) emotions, which has intuitive explanations.

In summary, the application of the proposed method to the CAL500 dataset reveals some interesting associations between these variables and can be used as a useful complement for analyzing high dimensional datasets with more complex distributions.

6 Conclusion

We propose an integrated framework for estimating and uncertainty assessment of the high dimensional semiparametric exponential family graphical models. Unlike exiting works, an important feature of our approach is that it is invariant to the base measures of the nodewise conditional distributions. For parameter estimation, we adopt the adaptive multi-stage relaxation algorithm to solve the nonconvex-penalized optimization problem; This method is computationally efficient and attains the minimax optimal rates of convergence. For graph uncertainty assessment, we propose a
Table 5: Global inference results for Ising graphical models: Rejection rates of the super-graph test for $H_0: G^* \subset G$ at the 0.05-significance level. The true graph is a 6-nearest-neighbor graph and $G$ is set to be a $k$-nearest neighbor graph with $k = 0, 2, 4, 6, 8$ and 10. Nonzero parameters have the same absolute value $\mu$.

| $\mu$ | $k = 0$ | $k = 2$ | $k = 4$ | $k = 6$ | $k = 8$ | $k = 10$ |
|-------|---------|---------|---------|---------|---------|---------|
| 1     | 0.00    | 0.062   | 0.062   | 0.060   | 0.052   | 0.052   |
| 2     | 0.10    | 0.282   | 0.140   | 0.104   | 0.066   | 0.054   |
| 3     | 0.20    | 0.462   | 0.324   | 0.242   | 0.062   | 0.058   |
| 4     | 0.30    | 0.752   | 0.622   | 0.468   | 0.064   | 0.052   |
| 5     | 0.40    | 0.886   | 0.792   | 0.610   | 0.062   | 0.056   |
| 6     | 0.50    | 0.940   | 0.864   | 0.722   | 0.066   | 0.060   |
| 7     | 0.60    | 0.970   | 0.916   | 0.806   | 0.062   | 0.058   |
| 8     | 0.70    | 0.980   | 0.944   | 0.874   | 0.068   | 0.062   |
| 9     | 0.80    | 1.000   | 0.962   | 0.924   | 0.064   | 0.058   |
| 10    | 0.90    | 1.000   | 0.982   | 0.952   | 0.058   | 0.056   |
| 11    | 1.00    | 1.000   | 0.990   | 0.968   | 0.056   | 0.054   |

A pairwise score test for examining the existence of an edge of the graph and a multiplier bootstrap based method for testing whether a given graph contains the true graph. The proposed theory is backed up by numerical results on both simulated and real-world datasets.

**Acknowledgement**

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

**A Proof of the Estimation Results**

In this appendix, we prove the main results for estimation results presented in §4.1. We first prove Theorem 4.2, which bounds the gradient of the loss function. We proceed to prove Proposition 4.5 which establishes the sparse eigenvalue condition for Gaussian graphical models. Then we prove Lemma 4.7, which shows that the sparse eigenvalue condition holds for $\nabla L_j(\cdot)$ in an $\ell_1$-ball centered at $\beta^*_j$ when $n$ is sufficiently large. Finally we prove Theorem 4.8 and Theorem 4.10 which establishes the $\ell_1$ and $\ell_2$-rates of convergence for the estimators obtained from the adaptive multi-stage convex relaxation algorithm.
Table 6: Global inference results for mixed models: Rejection rates of the super-graph test for \( H_0: G^* \subset G \) at the 0.05-significance level. The true graph is a 6-nearest-neighbor graph and \( G \) is set to be a \( k \)-nearest neighbor graph with \( k = 0, 2, 4, 6, 8 \) and 10. Nonzero parameters have the same absolute value \( \mu \).

|   | \( k = 0 \) | \( k = 2 \) | \( k = 4 \) | \( k = 6 \) | \( k = 8 \) | \( k = 10 \) |
|---|---|---|---|---|---|---|
| 1 | 0.00 | 0.066 | 0.066 | 0.064 | 0.064 | 0.056 |
| 2 | 0.03 | 0.072 | 0.072 | 0.062 | 0.062 | 0.054 |
| 3 | 0.06 | 0.122 | 0.088 | 0.076 | 0.056 | 0.056 |
| 4 | 0.09 | 0.264 | 0.160 | 0.104 | 0.054 | 0.054 |
| 5 | 0.12 | 0.450 | 0.296 | 0.242 | 0.064 | 0.062 |
| 6 | 0.15 | 0.658 | 0.524 | 0.402 | 0.058 | 0.052 |

A.1 Proof of Theorem 4.2

**Proof of Theorem 4.2.** Recall that \( \nabla L_j(\beta_j^*) \) is a centered second-order U-statistic with kernel function \( h_{ij}^j(\beta_j^*) \in \mathbb{R}^{d-1} \), whose entries are given by

\[
[h_{ij}^j(\beta_j^*)]_{jk} = \frac{R_{ij}(X_{ij} - X_{i'j})(X_{ik} - X_{i'k})}{1 + R_{ij}^j(\beta_j^*)}.
\]

By the independence between \( X_i \) and \( X_{i'} \), Assumption 4.1 implies that

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

which further implies that for any \( x > 0 \)

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

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

\[
P\left(\left|X_{ij} - X_{i'j}\right| > x^2\right) \leq P\left(\left|X_{ij} - X_{i'j}\right| > x\right) + P\left(\left|X_{ik} - X_{i'k}\right| > x\right) \\
\leq 4\exp(2\kappa_m + \kappa_h)\exp(-x), \tag{A.1}
\]

Letting \( c_0 = \exp(2\kappa_m + \kappa_h) \), by (A.1) we have

\[
P\left(\left|[h_{ij}^j(\beta_j^*)]_{jk}\right| > x^2\right) \leq P\left(\left|X_{ij} - X_{i'j}\right| > x\right) + c_0 \exp(-x), \quad x > 0. \tag{A.2}
\]

Thanks to the U-statistic structure of \( \nabla L_j(\beta_j^*) \), we obtain its tail probability by the concentration of U-statistics. We use the following concentration inequality presented in Ning and Liu (2014).

**Lemma A.1** (Concentration of U-statistics). Let \( X_1, ..., X_n \) be independent random variables. Consider the following U-statistics of order \( m \)

\[
U_n = \sum u(X_{i1}, ..., X_{im}) / \binom{n}{m},
\]



28
where the sum is over all \( i_1 < \ldots < i_m \) selected from \( \{1, \ldots, n\} \), and \( \mathbb{E}\{u(X_{i_1}, \ldots, X_{i_m})\} = 0 \) for all \( i_1 < \ldots < i_m \). If there exist positive constants \( L_1, L_2 \) and \( q \), such that

\[
P\left( |u(X_{i_1}, \ldots, X_{i_m})| \geq x \right) \leq L_1 \exp(-L_2 x^q),
\]  

(A.3)
for all $i_1 < ... < i_m$ and $x > 0$. Then for $x \geq \sqrt{8\mathbb{E}(U_n^2)}$,
\[
\mathbb{P}(|U_n| \geq x) \leq 4 \exp\left(-k^{q/(2+q)}x^{2q/(2+q)} / 8\right) + 4 \binom{n}{m} L_1 \exp\left(-L_2 k^{q/(2+q)}x^{2q/(2+q)} / 2^q\right),
\]
where $k = [n/m]$ is the largest integer less than $n/m$.

**Proof of Lemma A.1.** See the proof of Lemma 8.2 in Ning and Liu (2014) for a detailed proof.

Notice that the concentration inequality only holds for $x \geq \sqrt{8\mathbb{E}(U_n^2)} \propto n^{-1/2}$. Since $\mathbb{E}\{\nabla L_j(\beta_j^*)\} = 0$, by setting $L_1 = 4c_0$ $L_2 = 1$ and $q = 1/2$ in Lemma A.1 we have the following bound for tail probability of $\nabla L_j(\beta_j^*)$:
\[
\mathbb{P}\left(|\nabla_{jk} L_j(\beta_j^*)| > x\right) \leq 4 \exp\left(-m^{1/5}x^{2/5} / 8\right) + 16c_0 n^2 \exp\left(-m^{1/5}x^{2/5} / \sqrt{2}\right)/2,
\]
where $x \geq n^{-1/2}$ and $m = [n/2]$ is the largest integer less than $n/2$. Taking a union bound over \{(j, k) : k \in [d], k \neq j\} we obtain
\[
\mathbb{P}\left(\|\nabla L_j(\beta_j^*)\|_\infty > x\right) \lesssim n^2 d \exp\left(-m^{1/5}x^{2/5} / \sqrt{2}\right) \leq n^2 d \exp\left(-n^{1/5}x^{2/5} / 2\right),
\]
where we use $m = [n/2] \geq n/4$. By setting $x = (2\delta)^{5/2}(\log^5 d/n)^{1/2}$, we have
\[
\mathbb{P}\left(\|\nabla L_j(\beta_j^*)\|_\infty > (2\delta)^{5/2}(\log^5 d/n)^{1/2}\right) \lesssim n^2 d \exp(-\delta \log d), \quad \delta > 0.
\]
Moreover, since (A.4) holds for all $j, k \in [d]$, we conclude that (A.5) is also true for any $j \in [d]$. Therefore by choosing a sufficiently large $\delta > 0$, say $\delta = K_1^{2/5}/2$, we have
\[
\|\nabla L_j(\beta_j^*)\|_\infty \leq K_1 \sqrt{\log^5 d/n} \quad \forall j \in [d]
\]
holds with probability greater than $1 - (4d)^{-1}$.

Similarly, for any index set $S \subset \{(j, k) : k \in [d], k \neq j\}$ with $|S| \leq s^*$, a union bound over $S$ yields
\[
\mathbb{P}\left(\|\nabla S L_j(\beta_j^*)\|_2^2 > s^* x^2\right) \leq \sum_{(j, k) \in S} \mathbb{P}\left(|\nabla_{jk} L_j(\beta_j^*)| > x\right).
\]
Then by tail probability (A.4) we have
\[
\mathbb{P}\left(\|\nabla S L_j(\beta_j^*)\|_2^2 > s^* x^2\right) \lesssim s^* n^2 \exp\left(-n^{1/5}x^{2/5} / 2\right),
\]
By setting $x = (2\delta)^{5/2}(\log^5 n/n)^{1/2}$ in (A.7) we obtain
\[
\mathbb{P}(\|\nabla S L_j(\beta_j^*)\|_2^2 > s^*(2\delta)^{5/2}\log^5 n/n) \lesssim s^* n^2 \exp(-\delta \log n), \quad \delta > 0.
\]
Note that $\lim_{n \to \infty} s^*/n = 0$, by choosing a sufficiently large $\delta$, we conclude that there exists a constant $K_2$ such that with probability at least $1 - (4d)^{-1}$
\[
\|\nabla S L_j(\beta_j^*)\|_2^2 \leq K_2 s^* \log^5 n/n, \quad \forall j \in [d].
\]
Moreover, this is true for any fixed \( j \in [d] \) and any index set \( S \subset \{(j,k) : k \in [d], k \neq j\} \) as long as \( |S| \leq s^* \) because \((A.7)\) only depends on the size of the index set and holds for all \( j \in [d] \). Finally by taking a union bound we conclude that with probability at least \( 1 - (2d)^{-1} \), for all \( j \in [d] \) and \( S \subset \{(j,k) : k \in [d], k \neq j\} \) with \( |S| \leq s^* \), we have

\[
\|\nabla L_j(\beta_j^*)\|_{\infty} \leq K_1 \sqrt{\log^5 d/n} \quad \text{and} \quad \|\nabla S L_j(\beta_j^*)\|_2^2 \leq K_2 \sqrt{s^* \log^5 n/n}.
\]

\[\square\]

**A.2 Proof of Proposition 4.5**

*Proof of Proposition 4.5.* We prove this lemma in two steps. For any \( \beta_j \in \mathbb{R}^{d-1} \) such that \( \|\beta_j - \beta_j^*\|_1 \leq r \) and any \( v \in \mathbb{R}^{d-1} \) such that \( \|v\|_2 = 1 \), we first give a lower bound for \( v^T \mathbb{E}\{\nabla^2 L_j(\beta_j)\} v \) by truncation. Then we give an upper bound in the second step.

**Step (i): Lower Bound of** \( v^T \mathbb{E}\{\nabla^2 L_j(\beta_j)\} v \). For two truncation levels \( \tau > 0 \) and \( R > 0 \) and for any \( \beta_j \in \mathbb{R}^{d-1} \) such that \( \|\beta_j - \beta_j^*\|_1 \leq r \) we denote \( A_{ii'} := \{|X_{ij}| \leq \tau\} \cap \{|X_{i'j}| \leq \tau\}, B_i := \{|X^T_{ij} \beta_j^*| \leq R\} \cap \{|X^T_{i'j} \beta_j^*| \leq R\} \) and \( B_{ii'} := \{|X^T_{ij} \beta_j^*| \leq R\} \cap \{|X^T_{i'j} \beta_j^*| \leq R\} \). The values of \( R \) and \( \tau \) will be determined later. By the definition of \( \nabla^2 L_j(\cdot) \), for any \( v \in \mathbb{R}^{d-1} \) with \( \|v\|_2 = 1 \), we have

\[
v^T \nabla^2 L_j(\beta_j) v \geq \frac{2C_1(R, \tau)}{n(n-1)} \sum_{i < i'} (X_{ij} - X_{i'j})^2 \left\{ (X_{ij} - \Theta_{ij} \beta_j^*)^T v \right\}^2 I(B_i) I(B_{ii'}) I(A_{ii'}),
\]

where \( C_1(R, \tau) := \exp(-4R\tau)(1 + \exp(-4R\tau))^{-2} \). For notational simplicity, we denote the right-hand side of \((A.8)\) as \( C_1(R, \tau) v^T \Delta v \). By the properties of Gaussian graphical models, the conditional density of \( X_{ij} \) given \( \mathcal{I} := \{X_{i'j} = x_{i'j}\} \cap B_i \) is

\[
p(x_{ij} | \mathcal{I}) = p(x_{ij} | B_i) / \int_{\mathbb{R}} p(x_{ij} | B_i) dx_{ij} = p(x_{ij} | x_{i'j}),
\]

where we use the fact that \( p(x_{ij} | B_i) = p(x_{ij}) / \mathbb{P}(B_i) \) and that \( \mathbb{P}(B_i) \) is a constant. Recall that

\[
p(x_{ij} | X_{i'j}) = \sqrt{\Theta_{jj}/(2\pi)} \exp\left\{ -\Theta_{jj}/2 \left( x_{ij} - X_{i'j} \beta_j^* \right)^2 \right\}
\]

where \( \beta_j^* = \Theta_{j'j} \),

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

\[
\mathbb{E}\left[(X_{ij} - X_{i'j})^2 I(A_{ii'}) | \mathcal{I}_i \cap \mathcal{I}_i' \right] = \Theta_{jj}/(2\pi) \int_{-\tau}^{\tau} \int_{-\tau}^{\tau} (x_{ij} - x_{i'j})^2 \exp\left\{ -\Theta_{jj}/2 \left( x_{ij} - \Theta_{j'j} x_{i'j} \beta_j^* \right)^2 + (x_{i'j} - \Theta_{j'j} x_{i'j} \beta_j^*)^2 \right\} dx_{ij} dx_{i'j}.
\]

Note that on event \( \mathcal{I}_i, |\beta_j^T X_{i'j}| \leq R \), hence the expression above can be lower-bounded by

\[
\mathbb{E}\left[(X_{ij} - X_{i'j})^2 I(A_{ii'}) | \mathcal{I}_i \cap \mathcal{I}_i' \right] \geq \Theta_{jj}/(2\pi) \int_{-\tau}^{\tau} \int_{-\tau}^{\tau} (x_{ij} - x_{i'j})^2 \exp\left\{ -\Theta_{jj}/2 \left( x_{ij}^2 + x_{i'j}^2 + 2R^2 + 2R(|x_{ij}| + |x_{i'j}|) \right) \right\} dx_{ij} dx_{i'j}.
\]
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
\[
v^T \mathbb{E}(\Delta) v = v^T \mathbb{E} \left\{ \mathbb{E}(\Delta | \cap_{i=1}^n I_i) \right\} v \geq C_2(R, \tau) \mathbb{E} \left\{ (X_{i \cap j} - X_{i' \cap j'})^T v \right\}^2 I(B_i) I(B_j).
\]

By Cauchy-Schwarz inequality we have
\[
\mathbb{E} \left\{ ((X_{i \cap j} - X_{i' \cap j'})^T v)^2 (1 - I(B_i) I(B_j)) \right\} \leq \sqrt{\mathbb{E} \left\{ (X_{i \cap j} - X_{i' \cap j'})^T v \right\}^4} \sqrt{\mathbb{P}(B_i^c \cup B_j^c)}.
\]

(A.9)

Note that for Gaussian graphical model, the marginal distribution of \( X_{i \cap j} \) is \( N(0, \Sigma_{i \cap j}) \). If we denote \( \Sigma_{i \cap j} \) as \( \Sigma_j \), we have \( (X_{i \cap j} - X_{i' \cap j'})^T v \sim N(0, \sigma_i^2) \), \( X_{i \cap j}^T \beta_j^* \sim N(0, \sigma_j^2) \) and \( X_{i \cap j}^T \beta \sim N(0, \sigma_j^2) \) where \( \sigma_j^2 = 2v^T \Sigma_j v \), \( \sigma_j^2 = \beta_j^T \Sigma_j \beta_j^* \) and \( \sigma_j^2 = \beta_j^T \Sigma_j \beta_j \). Hence we have
\[
\mathbb{E} \left\{ (X_{i \cap j} - X_{i' \cap j'})^T v \right\}^4 = 3\sigma_j^4.
\]

Because the maximum eigenvalue of \( \Sigma_j \) is upper bounded by \( C_\lambda \), we have \( \sigma_j^2 \leq C_\lambda \sigma_j^2 \) and \( \sigma_j^2 \leq 2C_\lambda \). Note that \( \sigma_j^2 - \sigma_j^2 = \beta_j^T \Sigma_j \beta_j - \beta_j^T \Sigma_j \beta_j^* \), the following lemma in linear algebra bounds this type of error:

Lemma A.2. Let \( M \in \mathbb{R}^{d \times d} \) be a symmetric matrix and vectors \( v_1 \) and \( v_2 \in \mathbb{R}^d \), then
\[
\| v_1^T M v_1 - v_2^T M v_2 \| \leq \| M \| \| v_1 - v_2 \|^2 + 2 \| M \| \| v_1 - v_2 \|.
\]

Proof of Lemma A.2. Note that \( v_1^T M v_1 - v_2^T M v_2 = (v_1 - v_2)^T M (v_1 - v_2) + 2 v_2^T M (v_1 - v_2) \), Hölder’s inequality implies
\[
\| v_1^T M v_1 - v_2^T M v_2 \| \leq \| (v_1 - v_2)^T M (v_1 - v_2) \| + 2 \| v_2^T M (v_1 - v_2) \|
\]
\[
\leq \| M \| \| v_1 - v_2 \|^2 + 2 \| M \| \| v_1 - v_2 \|.
\]

By Lemma A.2, we have
\[
\sigma_j^2 - \sigma_j^2 \leq \| \Sigma_j \| \| \beta_j - \beta_j^* \|^2 + 2 \| \Sigma_j \beta_j^* \| \| \beta_j - \beta_j^* \|.
\]

(A.10)

By Hölder’s inequality and the relation between \( \ell_1 \)-norm and \( \ell_2 \)-norm of a vector, we have \( \| \Sigma_j \beta_j^* \| \leq \| \Sigma_j \| \| \beta_j^* \| \leq \sqrt{s\sigma} C_\beta D \). Therefore the right-hand side of \( (A.10) \) can be bounded by
\[
\sigma_j^2 - \sigma_j^2 \leq r^2 D + 2\sqrt{s\sigma} C_\beta D,
\]

which shows that \( \sigma_j^2 \) is also bounded because \( r = \mathcal{O}(1/\sqrt{s\sigma}) \). In addition, by the bound \( 1 - \Phi(x) \leq \exp(-x^2/2)/(x/\sqrt{2\pi}) \) for the standard normal distribution function, we obtain
\[
\mathbb{P}(B_i^c) \leq \mathbb{P}(X_{i \cap j}^T \beta_j^* > R) + \mathbb{P}(X_{i \cap j}^T \beta_j > R) \leq c \sigma_j \exp\left\{ -R^2/(2\sigma_j^2) \right\} / R + c \sigma_j \exp\left\{ -R^2/(2\sigma_j^2) \right\} / R,
\]

where the constant \( c = 1/\sqrt{2\pi} \). We denote the last expression as \( C_3(R) \), then the right-hand side of (A.9) can be upper-bounded by \( \sqrt{3\sigma_j^2} \sqrt{2C_3(R)} \leq 2 \sqrt{6C_3(R)} C_\lambda \). Hence we can choose a sufficiently large \( R \) such that \( 2 \sqrt{6C_3(R)} C_\lambda = \lambda_{\min}(\Sigma) \) and we denote this particular choice of \( R \) as \( R_0 \).
Now we have
\[ E \left\{ (\mathbf{X}_{i,j} - \mathbf{X}_{i',j})^T \mathbf{v} \right\}^2 \leq \lambda_{\min}(\Sigma) \]
Note that \( E \left\{ (\mathbf{X}_{i,j} - \mathbf{X}_{i',j})^T \mathbf{v} \right\}^2 = \sigma_v^2 \geq 2\lambda_{\min}(\Sigma) \), we obtain that
\[ \mathbf{v}^T E(\nabla^2 L_j(\beta_j)) \mathbf{v} \geq C_1(R_0, \tau)C_2(R_0, \tau)\lambda_{\min}(\Sigma) \]
for all \( \tau \in \mathbb{R} \).
Therefore we conclude that for all \( \beta_j \in \mathbb{R}^{d-1} \) such that \( \|\beta_j - \beta_j^*\|_1 \leq r \),
\[ \mathbf{v}^T E(\nabla^2 L_j(\beta_j)) \mathbf{v} \geq \max_{\tau \in \mathbb{R}} \left\{ C_1(R_0, \tau)C_2(R_0, \tau) \right\} \lambda_{\min}(\Sigma). \tag{A.11} \]

**Step (ii): Upper Bound of** \( \mathbf{v}^T E(\nabla^2 L_j(\beta_j)) \mathbf{v} \). For any \( \beta_j \in \mathbb{R}^{d-1} \) such that \( \|\beta_j - \beta_j^*\|_1 \leq r \) and for any \( \mathbf{v} \in \mathbb{R}^{d-1} \) with \( \|\mathbf{v}\|_2 = 1 \), by the definition of \( \nabla^2 L_j(\beta_j) \) we have
\[ \mathbf{v}^T \nabla^2 L_j(\beta_j) \mathbf{v} \leq (X_{ij} - X_{i',j})^2((\mathbf{X}_{i,j} - \mathbf{X}_{i',j})^T \mathbf{v})^2. \tag{A.12} \]
Notice that conditioning on \( \mathbf{X}_{i,j}, \mathbf{X}_{i',j} \sim N(\mathbf{X}_{i,j}^T \beta_j^*, \Sigma_j^{-1}) \), hence
\[ E((X_{ij} - X_{i',j})^2|\mathbf{X}_{i,j}, \mathbf{X}_{i',j}) = ((\mathbf{X}_{i,j} - \mathbf{X}_{i',j})^T \beta_j^*)^2 + 2\Theta_j^{-1}. \tag{A.13} \]
Combining (A.12) and (A.13) we obtain
\[ E(\mathbf{v}^T \nabla^2 L_j(\beta_j) \mathbf{v}) \leq E\left\{ E((\mathbf{X}_{i,j} - \mathbf{X}_{i',j})^2|\mathbf{X}_{i,j}, \mathbf{X}_{i',j}) \cdot ((\mathbf{X}_{i,j} - \mathbf{X}_{i',j})^T \mathbf{v})^2 \right\}
\[ \leq 2\Theta_j^{-1} E((\mathbf{X}_{i,j} - \mathbf{X}_{i',j})^T \mathbf{v})^2 + E\left\{ ((\mathbf{X}_{i,j} - \mathbf{X}_{i',j})^T \beta_j^*)^2 ((\mathbf{X}_{i,j} - \mathbf{X}_{i',j})^T \mathbf{v})^2 \right\}. \tag{A.14} \]
Because \( \mathbf{X}_{i,j} \sim N(\mathbf{0}, \Sigma_1) \) where \( \Sigma_1 := \Sigma_{i,j,\cdot,j} \), and also note that the maximum eigenvalue of \( \Sigma_1 \) is upper bounded by \( C_\lambda \), we have
\[ E((\mathbf{X}_{i,j} - \mathbf{X}_{i',j})^T \mathbf{v})^2 = 2\mathbf{v}^T \Sigma_1 \mathbf{v} \leq 2C_\lambda. \]
Moreover, by inequality \( 2ab \leq a^2 + b^2 \) we obtain
\[ 2E\left\{ ((\mathbf{X}_{i,j} - \mathbf{X}_{i',j})^T \beta_j^*)^2 ((\mathbf{X}_{i,j} - \mathbf{X}_{i',j})^T \mathbf{v})^2 \right\} \leq E((\mathbf{X}_{i,j} - \mathbf{X}_{i',j})^T \beta_j^*)^4 + E((\mathbf{X}_{i,j} - \mathbf{X}_{i',j})^T \mathbf{v})^4. \]
Since \( (\mathbf{X}_{i,j} - \mathbf{X}_{i',j})^T \mathbf{v} \sim N(0, \sigma_v^2) \) and \( (\mathbf{X}_{i,j} - \mathbf{X}_{i',j})^T \beta_j^* \sim N(0, 2\sigma_0^2) \) where \( \sigma_0^2 \) and \( \sigma_j^2 \) are defined as \( 2\mathbf{v}^T \Sigma_1 \mathbf{v} \) and \( \beta_j^T \Sigma_1 \beta_j^* \) respectively, we obtain
\[ E((\mathbf{X}_{i,j} - \mathbf{X}_{i',j})^T \beta_j^*)^4 = 3\sigma_v^4 \leq 12C_\lambda^2 \]
and \( E((\mathbf{X}_{i,j} - \mathbf{X}_{i',j})^T \mathbf{v})^4 = 12\sigma_0^4 \leq 12C_\lambda^2 C_\beta^2. \)
Therefore we can bound the right-hand side of (A.14) by
\[ E(\mathbf{v}^T \nabla^2 L_j(\beta_j) \mathbf{v}) \leq 4\Theta_j^{-1}C_\lambda + 6C_\lambda^2 + 6C_\lambda C_\beta^2. \tag{A.15} \]
Combining (A.11) and (A.15) we conclude that Proposition 4.5 holds with
\[ \rho_\star = \max_{\tau \in \mathbb{R}} \{ C_1(R_0, \tau)C_2(R_0, \tau) \} \lambda_{\min}(\Sigma) \] and \( \rho_\star^* = 4\Theta_j^{-1}C_\lambda + 6C_\lambda^2 + 6C_\lambda C_\beta^2. \)

□
A.3 Proof of lemma 4.7

Proof of lemma 4.7. Under Assumption 4.4, for any fixed \( \beta \in \mathbb{R}^{d-1} \) such that \( \| \beta - \beta^* \|_2 \leq r \) and any \( v \in \mathbb{R}^{d-1} \) with \( \| v \|_0 \leq 2s^* + k^* \), we denote the deviation of Hessian \( \nabla^2 L_j(\beta) - \mathbb{E}\{\nabla^2 L_j(\beta)\} \) as \( \Lambda \), our goal is to show that \( v^T \Lambda v \) is negligible. Hölder’s inequality implies that
\[
|v^T \Lambda v| \leq \| v \|_1 \| \Lambda v \|_\infty \leq \| v \|_2 \| \Lambda \|_\infty.
\]

Then we use the following lemma to control \( \| \Lambda \|_\infty \), which gives us an upper-bound for \( v^T \Lambda v \).

Lemma A.3. Under Assumption 4.1, for all \( j \in [d] \) and \( \beta_j \in \mathbb{R}^{d-1} \) there exist a constant \( C_h > 0 \) such that with probability at least \( 1 - (4d)^{-1} \),
\[
\| \nabla^2 L_j(\beta_j) - \mathbb{E}\{\nabla^2 L_j(\beta_j)\} \|_\infty \leq C_h \sqrt{\log d/n} \quad \text{for all } j \in [d].
\]

Moreover, for any \( j \in [d] \), let \( \beta \in \mathbb{R}^{d-1} \) such that \( \| \beta - \beta^* \|_1 \leq r_1(s^*, n, d) \) where \( s^* = \max_{j \in [d]} \| \beta^*_j \|_0 \) and \( r_1(s^*, n, d) \) satisfies \( \lim_{n \to \infty} r_1(s^*, n, d) \log^2 d = 0 \), it holds with probability at least \( 1 - (2d)^{-1} \) that
\[
\| \nabla^2 L_j(\beta_j) - \nabla^2 L_j(\beta^*_j) \|_\infty \leq C r_1(s^*, n, d) \log^2 d \quad \text{for all } j \in [d],
\]
where \( C \) is a constant that does not involve \( s^*, n \) or \( d \).

Proof of Lemma A.3. See §C.1.1 for a detailed proof. \( \square \)

By the relation between \( \ell_1 \)-norm and \( \ell_2 \)-norms and Lemma A.3, we have
\[
|v^T \Lambda v| \leq (2s^* + k^*) \| v \|_2 \| \Lambda \|_\infty \leq C_h(2s^* + k^*) \sqrt{\log d/n} \| v \|_2^2
\]
holds with probability at least \( 1 - (4d)^{-1} \). Since under Assumption 4.4 we have \( \lim_{n \to \infty} (2s^* + k^*) \sqrt{\log d/n} = 0 \), if \( n \) is large enough such that \( C_h(2s^* + k^*) \sqrt{\log d/n} \leq 0.05 \rho_* \), then we have
\[
0.95 \rho_* \leq \rho_{j-}(2s^* + k^*) - 0.05 \rho_* \leq \rho_{j-}(2s^* + k^*) < \rho_{j+}(k^*) \leq \rho_{j+}(k^*) + 0.05 \rho_*
\]
where we denote the \( s \)-sparse eigenvalues \( \rho_{j-}(\nabla^2 L_j(\cdot), \beta^*_j, s, r) \), \( \rho_{j+}(\nabla^2 L_j(\cdot), \beta^*_j, s, r) \) and \( \rho_+(\nabla^2 L_j(\cdot), \beta^*_j, s, r) \) as \( \rho_{j-}(s) \), \( \rho_{j-}(s) \), \( \rho_{j+}(s) \) and \( \rho_{j+}(s) \) respectively. Under Assumption 4.4, \( \rho_{j+}(k^*) / \rho_{j-}(2s^* + k^*) \leq 1 + 0.2k^*/s^* \) and \( k^* \geq 2s^* \), simple computation yields that
\[
\frac{\rho_{j+}(k^*)}{\rho_{j-}(2s^* + k^*)} \leq \frac{\rho_{j+}(k^*) + 0.05 \rho_*}{\rho_{j-}(2s^* + k^*)} \leq \frac{\rho_{j+}(k^*) + 0.05 \rho_{j-}(2s^* + k^*)}{0.95 \rho_{j-}(2s^* + k^*)} \leq 1 + 0.27 k^*/s^*.
\]
\( \square \)

34
A.4 Proof of Theorem 4.8

Proof of Theorem 4.8. 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 plays a significant role in our analysis. For all $j \in [d]$, we let $S_j := \{(j,k) : \beta_{jk}^* \neq 0\}$ be the support of $\beta_j^*$. For the number of iterations $\ell = 1, 2, \ldots$, let $G_\ell^j := \{(j,k) : \lambda_j^{(\ell-1)}(c_2 \lambda) \geq p'(c_2 \lambda), k \in [d]\}$ By condition (C.3) of the penalty function $p(\lambda(u)$ (see §3.2), we have $p'(c_2 \lambda) \geq 0.91 \lambda$. In addition, we let $J_j^\ell$ be the largest $k^*$ components of $[\hat{\beta}_j^{(\ell)}]G_\ell^j$ in absolute value where $k^*$ is defined in Assumption 4.4. In addition, we let $I_j^\ell = (G_\ell^j)^c \cup J_j^\ell$. Moreover, for notational simplicity, we denote $[\beta_j]_{G_j^\ell}^s$, $[\beta_j]_{G_j^\ell}$ and $[\beta_j]_{I_j^\ell}$ as $\beta_{G_j^\ell}, \beta_{J_j^\ell}$ and $\beta_{I_j^\ell}$ respectively when no ambiguity arises.

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

As assumed in the theorem, the regularization parameter satisfy $\lambda \geq 25K_1 \sqrt{\log^5 d/n}$. By Theorem 4.2 we have

$$\lambda \geq 25 \| \nabla L_j(\beta_j^*) \|_\infty \quad \text{for all } s \in V$$

(A.18)

with probability at least $1 - (4d)^{-1}$. We will also denote the $s$-sparse eigenvalues $\rho_-(\nabla^2 L_j(\cdot), \beta_j^*; s, r)$ and $\rho_+(\nabla^2 L_j(\cdot), \beta_j^*; s, r)$ as $\rho_j^-(s)$ and $\rho_j^+(s)$ respectively, where the radius $r$ of the $\ell_1$-ball centered at $\beta_j^*$ appears in assumption (4.4). Then lemma 4.7 implies that with probability at least $1 - 2d^{-1}$ we have

$$0.95 \rho \leq \rho_j^-(2s^* + k^*) \leq \rho_j^+(k^*) < +\infty \quad \text{and}$$

$$\rho_j^+(k^*) / \rho_j^-(2s^* + k^*) \leq 1 + 0.27k^*/s^* \quad \text{for all } j \in [d].$$

(A.19) (A.20)

Hereafter we condition on the event where (A.19),(A.20) and (4.2) hold. Theorem 4.2 and Lemma 4.7 implies that such event appears with probability at least $1 - d^{-1}$.

As stated above, the main point is to show that $|(G_\ell^j)^c| \leq 2s^*$ for each $\ell \geq 1$, the proof is based on induction. For $\ell = 1$, we have $G_1^j = S_j$ because $\lambda_j^{(0)} = \lambda$ for all $j, k \in [d]$. Hence $|(G_1^j)^c| \leq s^*$. Now we assume that $|(G_\ell^j)^c| \leq 2s^*$ for some integer $\ell$ and our goal is to prove that $|(G_{\ell+1}^j)^c| \leq 2s^*$. Our proof is based on two technical lemmas. Our first lemma bounds the $\ell_1$-norm and $\ell_2$-norm of $\hat{\beta}_j^{(\ell)} - \beta_j^*$ using the its subvector under the induction assumption that $|(G_\ell^j)^c| \leq 2s^*$.

Lemma A.4. Letting the index sets $S_j, G_\ell^j, J_j^\ell$ and $I_j^\ell$ be defined as above, we denote $\widetilde{G}_j^\ell := (G_\ell^j)^c$, under the assumption that $|G_\ell^j| \leq 2s^*$, we have

$$\|\hat{\beta}_j^{(\ell)} - \beta_j^*\|_2 \leq 2.2\|\hat{\beta}_j^{(\ell)} - \beta_j^*\|_2$$

and

$$\|\hat{\beta}_j^{(\ell)} - \beta_j^*\|_1 \leq 2.2\|\hat{\beta}_j^{(\ell)} - \beta_j^*\|_1,$$

(A.21)

Proof of Lemma A.4. See §C.1.2 for a detailed proof.

\[\square\]
Our second lemma guarantees that \( \hat{\beta}^{(t)} \) stays in the \( \ell_1 \)-ball centered at \( \beta^*_j \) with radius \( r \) for \( \ell \geq 1 \) where \( r \) appears in Assumption 4.4. 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 A.5.** For \( \ell \geq 1 \) and \( j \in [d] \), assuming that \( |(G^*_j)^c| \leq 2s^* \), the estimators \( \hat{\beta}^{(t)}_j \) obtained in each iteration of Algorithm 1 satisfies \( \| \hat{\beta}^{(t)}_j - \beta^*_j \|_1 \leq r \) where \( r \) appears in Assumption 4.4. Moreover, we denote \( \lambda^{(t)}_{S_j} = (\lambda^{(t)}_{jk}, (j, k) \in S_j)^T \), then we have

\[
\| \hat{\beta}^{(t)}_{I^j} - \beta^*_j \|_2 \leq 10r^{-1}_s \left( \| \nabla \tilde{\psi}^{(t)}_{L_j}(\beta^*_j) \|_2 + \| \lambda^{(t-1)}_{S_j} \|_2 \right), \quad \tilde{G}^t := (G^t)^c. \tag{A.22}
\]

This implies the following crude rates of convergence for \( \hat{\beta}^{(t)}_j \):

\[
\| \hat{\beta}^{(t)}_j - \beta^*_j \|_2 \leq 24r^{-1}_s \sqrt{s^*} \lambda \quad \text{and} \quad \| \hat{\beta}^{(t)}_j - \beta^*_j \|_1 \leq 33r^{-1}_s s^* \lambda. \tag{A.23}
\]

**Proof of Lemma A.5.** See §C.1.3 for a detailed proof. \( \square \)

Now we show that \( \tilde{G}^t_{jk} = (G^t_{jk})^c \) satisfies \( |\tilde{G}^t_{jk}| \leq 2s^* \), which concludes our induction. Letting \( A := (G^t_{jk})^c - S_j \), by the definition of \( G^t_{jk} \), \((j, k) \in A \) implies that \( (j, k) \notin S_j \) and \( p'_A(|\tilde{G}^{(t)}_{jk}|) \leq p'_A(c_2\lambda) \). Hence by the concavity of \( p_A(\cdot) \), for any \((j, k) \in A\), \( |\tilde{\beta}^{(t)}_{jk}| \geq c_2 \lambda \). Therefore we have

\[
\sqrt{|A|} \leq \| \hat{\beta}^{(t)}_{A} - \beta^*_A \|_2 / (c_2 \lambda) = \| \hat{\beta}^{(t)}_{A} - \beta^*_A \|_2 / (c_2 \lambda) \leq 24r^{-1}_s \sqrt{s^*} / c_2 \leq \sqrt{s^*}, \tag{A.24}
\]

where the first inequality follows from \( |A| \leq \sum_{(j, k) \in A} |\tilde{\beta}^{(t)}_{jk}|^2 / (c_2 \lambda)^2 \). Note that (A.24) implies that \( |(G^t_{jk})^c| \leq 2s^* \). Therefore by induction, \( |(G^t_{jk})^c| \leq 2s^* \) for any \( \ell \geq 1 \).

Now we have shown that for \( \ell \geq 1 \) and \( j \in [d], \) \( |(G^t_j)^c| \leq 2s^* \) and the crude statistical rates (A.23) hold. In what follows, we derive the more refined rates (4.3) and (4.4).

**A refined bound for** \( \| \hat{\beta}^{(t)}_j - \beta^*_j \|_2 \) **and** \( \| \hat{\beta}^{(t)}_j - \beta^*_j \|_1 \): For notational simplicity, we let \( \delta^{(t)} = \hat{\beta}^{(t)}_j - \beta^*_j \) and omit subscript \( j \) in \( S_j, G^*_j, J^*_j \) and \( I^j \). We denote \( \tilde{G}^t := (G^t)^c \). We will first derive a recursive bound that links \( \| \delta^{(t)}_{I^t} \|_2 \) to \( \| \delta^{(t-1)}_{I^{t-1}} \|_2 \). Note that by (A.21), \( \| \delta^{(t)} \|_1 \leq 2.2 \| \delta^{(t)} \|_1 \leq 2.2 \sqrt{2s^*} \| \delta^{(t)} \|_1 \) we only need to control \( \| \delta^{(t)}_{I^t} \|_2 \) to obtain the rates of convergence for \( \hat{\beta}^{(t)}_j \). By triangle inequality,

\[
\| \nabla \tilde{G}^t_{L_j}(\beta^*_j) \|_2 \leq \| \nabla S_{L_j}(\beta^*_j) \|_2 + \sqrt{|\tilde{G}^t - S|} \| \nabla L_j(\beta^*_j) \|_\infty.
\]

Since \( \lambda > 25 \| \nabla L_j(\beta^*_j) \|_\infty \), (A.24) implies that

\[
\| \nabla \tilde{G}^t_{L_j}(\beta^*_j) \|_2 \leq \| \nabla S_{L_j}(\beta^*_j) \|_2 + \| \delta^{(t-1)}_{A} \|_2 / (25c_2), \tag{A.25}
\]

where \( A := (G^t)^c - S \subset I^t \). Thus (A.25) can be written as

\[
\| \nabla \tilde{G}^t_{L_j}(\beta^*_j) \|_2 \leq \| \nabla S_{L_j}(\beta^*_j) \|_2 + \| \delta^{(t-1)}_{I^t} \|_2 / (25c_2). \tag{A.26}
\]
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 \( p'_\lambda(|\beta_{jk}|) \leq p'_\lambda(|\beta_{jk}^*| - c_2 \lambda) \) by the concavity of \( p_\lambda(\cdot) \). Hence the following inequality always holds:

\[
p'_\lambda(|\beta_{jk}|) \leq p'_\lambda(|\beta_{jk}^*| - c_2 \lambda) + |\beta_{jk} - \beta_{jk}^*|/c_2.
\] (A.27)

Applying (A.27) to \( \tilde{\beta}^{(\ell-1)}_j \) we have

\[
\|\lambda^{(\ell-1)}_S\|_2 \leq \left( \sum_{(j,k) \in \mathcal{S}} p'_\lambda(|\beta_{jk}^*| - c_2 \lambda)^2 \right)^{1/2} + \left( \sum_{(j,k) \in \mathcal{S}} |\tilde{\beta}^{(\ell-1)}_{jk} - \beta_{jk}^*|^2 \right)^{1/2} / c_2,
\]

which leads to

\[
\|\lambda^{(\ell-1)}_S\|_2 \leq \left( \sum_{(j,k) \in \mathcal{S}} p'_\lambda(|\beta_{jk}^*| - c_2 \lambda)^2 \right)^{1/2} + \|\delta^{(\ell-1)}_{\ell-1}\|_2 / c_2.
\] (A.28)

By (A.22), (A.26) and (A.28) we obtain

\[
\|\delta^{(\ell)}_{\ell}\|_2 \leq 10 \rho^{-1}_* \left\{ \|\nabla_S L_j(\beta_{j})\|_2 + \Upsilon_j \right\} + \gamma \|\delta^{(\ell-1)}_{\ell-1}\|_2,
\]

where \( \gamma := 11(c_2 \rho^*_s)^{-1} \) and we define \( \Upsilon_j := \left( \sum_{(j,k) \in \mathcal{S}} p'_\lambda(|\beta_{jk}^*| - c_2 \lambda)^2 \right)^{1/2} \) for notational simplicity. Note that since \( c_2 \geq 24 \rho^{-1}_s \), we have \( \gamma < 1 \). By recursion we obtain

\[
\|\delta^{(\ell)}_{\ell}\|_2 \leq 10 \rho^{-1}_* \left\{ \|\nabla_S L_j(\beta_{j})\|_2 + \Upsilon_j \right\} + \gamma^{\ell-1} \|\delta^{(1)}_{1}\|_2.
\] (A.29)

Using \( \|\tilde{\beta}^{(\ell)}_j - \beta_{j}^*\|_2 \leq 2.2 \|\tilde{\beta}^{(\ell)}_{I_j} - \beta_{I_j}^*\|_2 \), we can bound \( \|\tilde{\beta}^{(\ell)}_j - \beta_{j}^*\|_2 \) by

\[
\|\tilde{\beta}^{(\ell)}_j - \beta_{j}^*\|_2 \leq 22 \rho^{-1}_* \left\{ \|\nabla_S L_j(\beta_{j})\|_2 + \Upsilon_j \right\} + 2.2 \gamma^{\ell-1} \|\delta^{(1)}_{1}\|_2.
\]

Note that for \( \ell = 1 \), by (A.22) we have

\[
\|\delta^{(1)}_{1}\|_2 \leq 10 \rho^{-1}_* \sqrt{s^2} (\lambda + \sqrt{2} \|\nabla L_j(\beta_{j})\|_\infty) \leq 11 \rho^{-1}_* \sqrt{s^2} \lambda \leq c_2 \gamma \sqrt{s^2} \lambda.
\] (A.30)

then we establish the following bound for \( \|\tilde{\beta}^{(\ell)}_j - \beta_{j}^*\|_2 \):

\[
\|\tilde{\beta}^{(\ell)}_j - \beta_{j}^*\|_2 \leq 22 \rho^{-1}_* \left\{ \|\nabla_S L_j(\beta_{j})\|_2 + \Upsilon_j \right\} + 2.2 c_2 \sqrt{s^2} \lambda \gamma^{\ell}.
\] (A.31)

Similarly, by \( \|\tilde{\beta}^{(\ell)}_j - \beta_{j}^*\|_1 \leq 2.2 \sqrt{2 s^3} \|\tilde{\beta}^{(\ell)}_{I_j} - \beta_{I_j}^*\|_2 \), we obtain a bound on \( \|\tilde{\beta}^{(\ell)}_j - \beta_{j}^*\|_1 \):

\[
\|\tilde{\beta}^{(\ell)}_j - \beta_{j}^*\|_1 \leq 32 \sqrt{s^3} \rho^{-1}_* \left\{ \|\nabla_S L_j(\beta_{j})\|_2 + \Upsilon_j \right\} + 2.2 \gamma^{\ell-1} \sqrt{2 s^3} \|\delta^{(1)}_{1}\|_2.
\] (A.32)
By (A.30) we have $2.2 \sqrt{2s^*} \| \delta_{j}^{(t)} \|_2 \leq 3.2c_2 \gamma s^* \lambda$, then the right-hand side of (A.32) can be bounded by

$$
\| \tilde{\beta}_j^{(t)} - \beta_j^* \|_1 \leq \frac{32s^*-1}{\rho_s} \left\{ \| \nabla S_j L_j(\beta_j^*) \|_2 + \gamma_j \right\} + 3.2c_2 \gamma s^* \lambda^\ell. \tag{A.33}
$$

Note that $\| \nabla S_j L_j(\beta_j^*) \|_2 \leq \sqrt{s^* \log^5 n/n}$, (4.3) and (4.4) can be implied by (A.31) and (A.33) respectively. Recall that we condition on the event where (A.19), (A.20) and (4.2) hold, which implies that (4.4) and (4.3) hold with probability at least $1 - d^{-1}$ for $j = 1, \ldots, d$.

A.5 Proof of Theorem 4.10

Proof of Theorem 4.10. Let $S_j = \{(j, k): \beta^*_j k \neq 0, k \in [d]\}$ be the support of $\beta_j^*$ and let index set $G^\ell, J^\ell_j$ and $I^\ell_j$ be the same as defined in the proof of Theorem 4.8. 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 $G^\ell, J^\ell$ and $I^\ell$. Moreover, we let $\delta^{(t)} = \beta_j^{(t)} - \beta_j^*$, it is shown in Lemma A.5 that

$$
\| \delta^{(t)}_{\ell} \|_2 \leq 10\rho_s^{-1} \left( \| \nabla \tilde{G}_\ell L_j(\beta_j^*) \|_2 + \| \lambda_s^{(\ell-1)} \|_2 \right); \quad \tilde{G}^\ell = (G^\ell)^c. \tag{A.34}
$$

In the proof of Theorem 4.8, we show that $|\tilde{G}^\ell| \leq 2s^*$ for all $j \in [d]$ and $\ell \geq 1$. Because $S_j = S_{1j} \cup S_{2j}$ where $S_{1j} = \{(j, k): |\beta^*_j k| \geq (c_2 + c_3)\lambda \}$ and $S_{2j} = S_j - S_{1j}$, then by triangle inequality we have

$$
\| \nabla S_j L_j(\beta_j^*) \|_2 \leq \| \nabla S_{1j} L_j(\beta_j^*) \|_2 + \sqrt{s_{2j}^2} \| \nabla S_{2j} L_j(\beta_j^*) \|_\infty.
$$

Since $\lambda > 25 \| \nabla L_j(\beta_j^*) \|_\infty$, by (A.26), we further have

$$
\| \nabla \tilde{G}_\ell L_j(\beta_j^*) \|_2 \leq \| \nabla S_{1j} L_j(\beta_j^*) \|_2 + \sqrt{s_{2j}^2} \lambda / 25 + \| \delta^{(\ell-1)}_{\ell} \|_2 / (25c_2). \tag{A.35}
$$

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

$$
Y_j := \lambda \left( \sum_{(j, k) \in S_j} p'_\lambda(|\beta^*_j k| - c_2 \lambda)^2 \right)^{1/2} = \lambda \left( \sum_{(j, k) \in S_{2j}} p'_\lambda(|\beta^*_j k| - c_2 \lambda)^2 \right)^{1/2} \leq \sqrt{s_{2j}^2} \lambda.
$$

Therefore (A.28) is reduced to

$$
\| \lambda^{(\ell-1)}_s \|_2 \leq Y_j + \| \delta^{(\ell-1)}_{\ell-1} \|_2 / c_2 \leq \sqrt{s_{2j}^2} \lambda + \| \delta^{(\ell-1)}_{\ell-1} \|_2 / c_2. \tag{A.36}
$$

Combining (A.34), (A.35) and (A.36) we obtain

$$
\| \delta^{(t)}_{\ell} \|_2 \leq 10\rho_s^{-1} \left\{ \| \nabla S_j L_j(\beta_j^*) \|_2 + 1.04 \sqrt{s_{2j}^2} \lambda + 1.04 \| \delta^{(\ell-1)}_{\ell-1} \|_2 / c_2 \right\}.
$$

38
Then by recursion, we obtain the following estimation error:
\[
\|\delta^{(t)}_{Jt}\|_2 \leq 10\rho_s^{-1}\left\{\|\nabla S_{1j} L_j(\beta^*_j)\|_2 + 1.04\sqrt{s_{2j}^*}\lambda\right\} + \gamma^{t-1}\|\delta^{(1)}_{J1}\|_2, \quad \gamma := 11c^{-1}\rho_s^{-1}.
\]
Note that we assume \(c_2 \geq 24\rho_s^{-1}\), for \(k = 1\) by (A.30) we have
\[
2.2\|\delta^{(1)}_{J1}\|_2 \leq 2.2c_2\gamma\sqrt{s^*}\lambda \quad \text{and} \quad 2.2\sqrt{2s^*}\|\delta^{(1)}_{J1}\|_2 \leq 3.2c_2\gamma s^*\lambda.
\]
Therefore using the original notation, we can obtain the refined rates of convergence by (A.21)
\[
\|\widehat{\beta}^{(t)}_j - \beta^*_j\|_2 \leq 22\rho_s^{-1}\left\{\|\nabla S_{1j} L_j(\beta^*_j)\|_2 + 1.04\sqrt{s_{2j}^*}\lambda\right\} + 2.2c_2\gamma^t\sqrt{s^*}\lambda \quad \text{and}
\]
\[
\|\widehat{\beta}^{(t)}_j - \beta^*_j\|_1 \leq 32\sqrt{s^*}\rho_s^{-1}\left\{\|\nabla S_{1j} L_j(\beta^*_j)\|_2 + 1.04\sqrt{s_{2j}^*}\lambda\right\} + 3.2c_2\gamma^t s^*\lambda,
\]
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 \(j \in [d]\). \(\square\)

## B Proof of Inferential Results

In this appendix, we prove the inferential results. More specifically, we first prove Lemma 4.12, which is pivotal for deriving the limiting distribution of the pairwise score statistic. Then we prove Lemma 4.15, which derives sparse eigenvalue condition for the Dantzig-type problem (3.7). We prove Theorem 4.16 and Corollary 4.17 next, which justify the validity of our pairwise score test. Finally, we prove Theorem 4.19; our proof applies the method of Gaussian multiplier bootstrap to the structure of second-order U-statistics.

### B.1 Proof of Lemma 4.12

**Proof of Lemma 4.12.** Before proving this lemma, we first let \(\nabla^2 L_{jk}(\beta_{jk})\) be the Hessian of \(L_{jk}(\beta_{jk})\) and define \(H^{jk} := E\{\nabla^2 L_{jk}(\beta_{jk})\}\). We also define
\[
\Sigma^{jk} := E\{g_{jk}(X_i)g_{jk}(X_i)^T\} \quad \text{and} \quad \Theta^{jk} := E\{h^{jk}_{ii}(\beta^*)h^{jk}_{ii'}(\beta^*)^T\}.
\]
Under assumption 4.1, we first show that there exist a positive constant \(D\) such that for any \(j, k \in d, j \neq k\),
\[
\max\left\{\left\|\Sigma^{jk}\right\|_\infty, \left\|H^{jk}\right\|_\infty, \left\|\Theta^{jk}\right\|_\infty\right\} \leq D.
\]
The reason is as follows.

Note that Hölder’s inequality implies
\[
\left\|H^{jk}\right\|_\infty \leq \max_{j \in [d]} E|X_{ij} - X_{ij'}|^4 \leq \max_{j \in [d]} E|X_j|^4 \quad \text{for any} \quad j, k \in [d], j \neq k.
\]
Similarly, for \(\Theta^{jk}\), we also have \(\left\|\Theta^{jk}\right\|_\infty \leq \max_{j \in [d]} E|X_j|^4\). By (4.1) we have
\[
E|X_j|^4 = \int_0^\infty P(|X_j| > t)dt \leq \int_0^\infty c \exp(-t^{1/4})dt = 24c, \quad c = 2 \exp(\kappa_m + \kappa_h/2).
\]

39
Moreover, note that by the law of total variance, the diagonal elements of \( \Sigma_{jk} \) are no larger than the corresponding diagonal elements of \( \Theta_{jk} \); then by Cauchy-Schwarz inequality, \( \|\Sigma_{jk}\|_\infty \leq \|\Theta_{jk}\|_\infty \). Therefore there exist a constant \( D \) that does not depend on \((s^*, n, d)\) such that
\[
\max \left\{ \|H_{jk}\|_\infty, \|\Sigma_{jk}\|_\infty, \|\Theta_{jk}\|_\infty \right\} \leq D, \quad 1 \leq j < k \leq d. \tag{B.1}
\]

Now we are really to prove this lemma.

Recall that \( \nabla L_{jk}(\beta_{j\vee k}) \) is a U-statistic with kernel function \( h_{i'j}^{jk}(\beta_{j\vee k}) \). Because \( h_{i'j}^{jk}(\beta_{j\vee k}) \) is centered, the law of total expectation implies that \( \mathbb{E}\{g_{jk}(X_i)\} = 0 \). Note that the left-hand side of (4.11) can be written as
\[
\frac{\sqrt{n}}{2} b^T \nabla L_{jk}(\beta_{j\vee k}) = \frac{\sqrt{n}}{2} b^T U_{jk} + \frac{\sqrt{n}}{2} b^T (\nabla L_{jk}(\beta_{j\vee k}) - U_{jk})
\]
\[
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[ \frac{b^T g_{jk}(X_i)}{I_1} + \frac{\sqrt{n}}{2} b^T (\nabla L_{jk}(\beta_{j\vee k}) - U_{jk}) \right].
\]

Notice that \( I_1 \) is a weighted sum of i.i.d. random variables with the mean and variance given by
\[
\mathbb{E}(b^T g_{jk}(X_i)) = 0 \quad \text{and} \quad \text{Var}(b^T g_{jk}(X_i)) = b^T \Sigma_{jk} b.
\]

Central limit theorem implies that \( I_1 \sim N(0, b^T \Sigma_{jk} b) \). In what follows we use \( h_{i'i'} \) and \( h_{i'i'|i} \) to denote \( h_{i'i'}^{jk}(\beta_{j\vee k}) \) and \( \mathbb{E}\{h_{i'i'}^{jk}(\beta_{j\vee k})|X_i\} = g_{jk}(X_i) \). Thus we can write \( I_2 \) as
\[
I_2 = \frac{1}{\sqrt{n}(n-1)} \sum_{i<i' \atop s<s'} b^T \chi_{i'i'}, \quad \text{where} \quad \chi_{i'i'} = (h_{i'i'} - h_{i'i'|i} - h_{i'i'|i'}). \]

Then \( \mathbb{E}(I_2^2) \) can be expanded as
\[
\mathbb{E}(I_2^2) = \frac{1}{n(n-1)^2} \sum_{i<i' \atop s<s'} b^T \mathbb{E}(\chi_{i'i'} \chi_{s's'})^T b. \tag{B.2}
\]

By the definition of \( \chi_{i'i'} \), we have
\[
\mathbb{E}(\chi_{i'i'} \chi_{s's'}^T) = \mathbb{E}(h_{i'i'} h_{s's'}^T) - \mathbb{E}(h_{i'i'} h_{s's'}^T|s) - \mathbb{E}(h_{i'i'} h_{s's'}^T|s') + \mathbb{E}(h_{i'i'} h_{s's'}^T|s' s')
\]
\[
+ \mathbb{E}(h_{i'i'} h_{s's'}^T|s) + \mathbb{E}(h_{i'i'} h_{s's'}^T|s') - \mathbb{E}(h_{i'i'} h_{s's'}^T) + \mathbb{E}(h_{i'i'|i} h_{s's'}^T) + \mathbb{E}(h_{i'i'|i} h_{s's'}^T|s) + \mathbb{E}(h_{i'i'|i} h_{s's'}^T|s'). \tag{B.3}
\]

Therefore, for \( i \neq s, s' \) and \( i' \neq s, s' \), law of total expectation implies that \( \mathbb{E}(\chi_{i'i'} \chi_{s's'}) = 0 \). Similarly, if exactly one of \( i, i' \) is identical to one of \( s, s' \), say \( i = s \), then (B.3) becomes
\[
\mathbb{E}(\chi_{i'i'} \chi_{i'i''}) = \mathbb{E}(h_{i'i'} h_{i'i''}^T) - \mathbb{E}(h_{i'i'} h_{i'i''}^T|i) - \mathbb{E}(h_{i'i'} h_{i'i''}^T|i'') + \mathbb{E}(h_{i'i'|i} h_{i'i''}^T|i'), \quad i \neq i' \neq i''.
\]

Note that by the law of total expectation, for each term in (B.3) we have
\[
\mathbb{E}(h_{i'i'} h_{i'i''}^T) = \mathbb{E}(h_{i'i'} h_{i'i''}^T|i) = \mathbb{E}(h_{i'i'} h_{i'i''}^T|i'') = \mathbb{E}(h_{i'i'} h_{i'i''}^T). \]

40
Therefore, $E(\chi_{ii'}\chi_{ii'}^T) = 0$. Finally, if $i = s$ and $i' = s'$, by the law of total expectation, (B.3) can be further reduced to $E(\chi_{ii'}\chi_{ii'}^T) = E(h_{ii'}h_{ii'}^T) − E(h_{ii'}|h_{ii'}^T) − E(h_{ii'}|h_{ii'}^T) = \Theta_{jj}^k − 2\Sigma_{jk}^k$. Thus by triangle inequality we have
\[
\|E(\chi_{ii'}\chi_{ii'}^T)\|_\infty \leq \|E(h_{ii'}h_{ii'}^T)\|_\infty + \|E(h_{ii'}|h_{ii'}^T)\|_\infty + \|E(h_{ii'}|h_{ii'}^T)\|_\infty \leq 3D,
\]
where the last inequality follows from Assumption 4.11. Then equation (B.2) can be reduced to
\[
E(I_2^2) = \frac{1}{n(n-1)^2} \sum_{i<i', s<s'} b^T E(\chi_{ii'}\chi_{ss}^T)b = \frac{1}{n(n-1)^2} \sum_{i<i'} b^T E(\chi_{ii'}\chi_{ii'}^T)b.
\]
By Hölder’s inequality we obtain
\[
E(I_2^2) \leq \frac{1}{2(n-1)} \|b\|_1 \|E(\chi_{ii'}\chi_{ii'}^T)b\|_\infty \leq \frac{1}{2(n-1)} \|b\|_1 \|E(\chi_{ii'}\chi_{ii'}^T)\|_\infty \leq \frac{3D}{2(n-1)} \|b\|_1^2.
\]
Since $\|b\|_0 \leq \tilde{s}$, by the relationship between $\ell_1$-norm and $\ell_2$-norm, we can further bound the right-hand side of (B.4) by $E(I_2^2) \leq 1.5\tilde{s}D/(n-1) \rightarrow 0$, where we use the condition that $\lim_{n \rightarrow \infty} \tilde{s}/n = 0$.  

B.2 Proof of lemma 4.15

Proof of lemma 4.15. Similar to the proof of Lemma 4.7, for any $\beta_j \in \mathbb{R}^{d−1}$ such that $\|\beta_j - \beta_j^s\|_1 \leq r$, we denote the deviation of Hessian $\nabla^2 L_j(\beta_j) - \mathbb{E}\nabla^2 L_j(\beta_j)$ as $\Lambda$. Lemma A.3 implies that $\|\Lambda\| \leq C_h\sqrt{\log d}/n$ holds for all $j \in [d]$, then we can choose $n$ to be sufficiently large such that $k^*\sqrt{\log d}/n \leq 0.05\nu_s$, this is true because $\lim_{n \rightarrow \infty} k^*\sqrt{\log d}/n = 0$. Therefore with probability at least $1 - (2d)^{-1}$ we have
\[
0.95\nu_s \leq \rho_{j−}(s^*+k*) - 0.05\nu_s \leq \rho_{j−}(s^*) \leq \rho_{j+}(s^*) \leq \rho_{j+}(k^*) \leq \rho_{j+}(k^*) + 0.05\nu_s \text{ for all } j \in [d],
\]
where we use $\rho_{j−}(s), \rho_{j−}(s), \rho_{j+}(s)$ and $\rho_{j+}(s)$ to denote the s-sparse eigenvalues $\rho_−(\mathbb{E}\nabla^2 L_j(\cdot), \beta_j^s; r)$, $\rho_−(\nabla^2 L_j(\cdot), \beta_j^s; s, r)$, $\rho_+(\mathbb{E}\nabla^2 L_j(\cdot), \beta_j^s; r)$, and $\rho_+(\nabla^2 L_j(\cdot), \beta_j^s; s, r)$ respectively, where $r$ appears in Assumption 4.4. Then by assumption 4.14 we have
\[
0.95\nu_s \leq \rho_{j−}(s^*+k*) \leq \rho_{j+}(k^*) \leq (1.05 + 0.5k^*/s^*)\nu_s.
\]
Moreover, we can also obtain that
\[
\rho_{j+}(k^*)/\rho_{j−}(s^*+k^*) \leq (1.05 + 0.5k^*/s^*)/0.95 \leq 1 + 0.58k^*/s^*,
\]
where we use the assumption that $k^* \geq s^*$.  

B.3 Proof of Theorem 4.16

Proof of theorem 4.16. We first remind the reader that for $1 \leq j \neq k \leq d$, we denote $\beta_j = (\beta_{jk}, k \neq j)^T \in \mathbb{R}^{d−1}$, $\beta_{j,k} = (\beta_{j,k}, \ell \neq j, k)^T \in \mathbb{R}^{d−2}$, $\beta_{k,j} = (\beta_{k,j}, \ell \neq j, k)^T$ and $\beta_{j,k,v} = (\beta_{j,k,v})^T \in \mathbb{R}^{2d−3}$. And we also denote $\tilde{\beta}_{j,v,k} = (0, \tilde{\beta}_{j,k}, \tilde{\beta}_{k,j})^T$. Our goal is to prove that
\[
\lim_{n \rightarrow \infty} \max_{j<k} \sqrt{n} \left| \hat{S}_{jk} - S_{jk}(\beta_{j,k,v}) \right| = 0.
\]
By the expression of $S_{jk}(\beta_{j,k}^*)$ and $\hat{S}_{jk}$ in (4.13) and (4.14), under null hypothesis, for a fixed pair of nodes $j$ and $k$, we have $S_{jk} - S_{jk}(\beta_{j,k}^*) = I_{1j} + I_{2j} + I_{1k} + I_{2k}$ where $I_{1j}$ and $I_{2j}$ are defined as

\[
I_{1j} := (\nabla_{jk}L_j(\tilde{\beta}_j) - \nabla_{jk}L_j(\beta_{j,k}^*)) - \tilde{w}_{j,k}^T(\nabla_{jk}L_j(\tilde{\beta}_j) - \nabla_{jk}L_j(\beta_{j,k}^*)) \quad \text{and}
\]
\[
I_{2j} := (\hat{w}_{j,k} - \tilde{w}_{j,k})^T \nabla_{jk}L_j(\beta_{j,k}^*);
\]

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

\[
I_{1k} := (\nabla_{kj}L_k(\tilde{\beta}_k) - \nabla_{kj}L_k(\beta_{j,k}^*)) - \tilde{w}_{k,j}^T(\nabla_{kj}L_k(\tilde{\beta}_k) - \nabla_{kj}L_k(\beta_{j,k}^*)) \quad \text{and}
\]
\[
I_{2k} := (\hat{w}_{k,j} - \tilde{w}_{k,j})^T \nabla_{kj}L_j(\beta_{j,k}^*).
\]

We first bound $I_{1j}$. Recall that $\tilde{\beta}_j = (0, \tilde{\beta}_{j,k})^T$. Note that under the null hypothesis, $\beta_{j,k}^* = 0$, by the Mean-Value Theorem, there exists a $\tilde{\lambda}_{j,k} \in \mathbb{R}^{d-2}$ in the line segment between $\tilde{\beta}_{j,k}$ and $\beta_{j,k}^*$ such that

\[
I_{1j} = \left\{ \tilde{\lambda}_{j,k} \tilde{w}_{j,k}^T \tilde{\beta}_{j,k} - w_{j,k}^T \beta_{j,k} \right\} (\tilde{\beta}_{j,k} - \beta_{j,k}^*),
\]

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

\[
|I_{1j}| \leq \left| \frac{\|\Lambda_{j,k,j,k} - \tilde{w}_{j,k}^T \Lambda_{j,k,j,k} \|_\infty \|\delta_{j,k}\|_1}{I_{11}} + \frac{\|\Lambda_{j,k,j,k} - \tilde{\lambda}_{j,k,j,k} \|_\infty \|\delta_{j,k}\|_1}{I_{12}} 
\right.
\]
\[
\left. + \frac{\|\tilde{w}_{j,k}^T (\Lambda_{j,k,j,k} - \tilde{\lambda}_{j,k,j,k}) \|_{\infty} \|\delta_{j,k}\|_1}{I_{13}} \right|
\]

Theorem 4.8 implies that $\|\delta\|_1 \leq C s^* \lambda$ with probability tending to 1 for some constant $C > 0$. Then by the definition of Dantzig selector, $I_{11} \leq C s^* \lambda D$, with high probability. Moreover, the constant $C$ is the same for all $(j, k)$. By assumption 4.13, $I_{11} = o(n^{-1/2})$ with probability tending to 1.

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

\[
I_{12} \leq \|\Lambda_{j,k,j,k} - \tilde{\lambda}_{j,k,j,k} \|_\infty \|\delta_{j,k}\|_1.
\]

By Lemma A.3 we obtain

\[
\|\Lambda - \tilde{\lambda}\|_\infty \leq \|\Lambda - \Lambda^*\|_\infty + \|\Lambda^* - \tilde{\lambda}\|_\infty \leq 2Cs^* \lambda \log^2 d.
\]

Therefore we have

\[
I_{12} \leq 2Cs^2 \lambda^2 \log^2 d \lesssim s^* \lambda D \quad \text{uniformly for } \{(j, k): 1 \leq j \neq k \leq d\}.
\]

Similarly by Hölder’s inequality, we have

\[
I_{13} \leq \|\tilde{w}_{j,k} \|_1 \|\Lambda - \tilde{\lambda}\|_\infty \|\delta\|_1.
\]
Notice that by the optimality of $\hat{w}_{j,k}$, $\|\hat{w}_{j,k}\|_1 \leq \|w_{j,k}^*\|_1 \leq w_0$. Combining (B.7) and (B.6) we have

$$I_{13} \leq C w_0 s^2 \lambda^2 \log^2 d \lesssim s^* \lambda \lambda_D$$

uniformly for $\{(j,k) : 1 \leq j \neq k \leq d\}$.

where we use the fact that $\lambda_D \gtrsim w_0 s^* \lambda \log^2 d$. Therefore we can conclude that

$$|I_{1j}| \leq s^* \lambda \lambda_D + s^2 \lambda^2 \log^2 d + w_0 s^2 \lambda^2 \log^2 d \lesssim s^* \lambda \lambda_D = o_P(n^{-1/2}).$$

For $I_{2j}$, Hölder’s inequality implies that $|I_2| \leq \|w_{j,k}^* - \hat{w}_{j,k}\|_1 \|\nabla L_j(\beta_j^*)\|_\infty$. To control $I_{2j}$, we need to the following lemma to obtain the estimation error of the Dantzig selector $\hat{w}_{j,k}$.

**Lemma B.1.** For $1 \leq j \neq k \leq d$, let $\hat{w}_{j,k}$ be the solution of the Dantzig-type optimization problem 3.7 and let $w_{j,k}^* = H_{j,k,j,k}^j H_{j,k,j,k}^j \Sigma_j^{-1}, w_{j,j}^* = H_{j,k,j,k}^j H_{j,k,j,k}^j \Sigma_j^{-1}$. Under the Assumptions 4.1, 4.4, 4.13 and 4.14, with probability tending to one, we have

$$\|\hat{w}_{j,k} - w_{j,k}^*\|_1 \leq 37 \nu_s^{-1} s^* \lambda \lambda_D \text{ for all } 1 \leq j \neq k \leq d.$$ 

*Proof of Lemma B.1.* See §C.2.1 for a detailed proof.

Now combining Lemma B.1 and Theorem 4.2 we obtain that

$$|I_2| \leq 37 \nu_s^{-1} K_1 s^* \lambda \lambda_D \sqrt{\log^5 d/n} \times s^* \lambda \lambda_D = o(n^{-1/2}).$$

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 $(j,k)$, we conclude that

$$\sqrt{n} \left( \hat{S}_{jk} - S_{jk}(\beta_{j,vk}^*) \right) = o_P(1) \text{ uniformly for } 1 \leq j < k \leq d.$$ 

The rest of the theorem follows from Lemma 4.12. By Lemma 4.12,

$$\frac{\sqrt{n}}{2} S(\beta_{j,vk}^*) = \nabla_j L_{j,k}(\beta_{j,vk}^*) - w_{j,k}^* \nabla_j L_{j,k}(\beta_{j,vk}^*) - w_{j,k}^* \nabla_j L_{j,k}(\beta_{j,vk}^*) - w_{j,k}^* \nabla_j L_{j,k}(\beta_{j,vk}^*) \sim N(0, \sigma^2_{jk}),$$

where the asymptotic variance $\sigma^2_{jk}$ is given by

$$\sigma^2_{jk} = \Sigma_{jk,j,k}^j - 2 \Sigma_{jk,j,k}^j w_{j,k}^* - 2 \Sigma_{jk,j,k}^j w_{j,k}^* + w_{j,k}^* \Sigma_{jk,j,k}^j w_{j,k}^* + w_{j,k}^* \Sigma_{jk,j,k}^j w_{j,k}^*.$$ 

For a more accurate estimation of $\hat{S}_{jk} - S_{jk}(\beta_{j,vk}^*)$, we have

$$\sqrt{n} \left| \hat{S}_{jk} - S_{jk}(\beta_{j,vk}^*) \right| \leq \sqrt{n} (|I_1| + |I_2|) \lesssim \sqrt{n} (s^* + s^*) \lambda \lambda_D.$$ 

We will use inequality (B.8) to prove Theorem (4.19).
B.4 Proof of Corollary 4.17

Proof of Corollary 4.17. We only need to show that $\hat{\sigma}^2_{jk}$ is a consistent estimator of $\sigma^2_{jk}$, which is equivalent to showing that $\lim_{n \to \infty} |\hat{\sigma}^2_{jk} - \sigma^2_{jk}| = 0$. To begin with, triangle inequality implies that

$$
|\sigma^2_{jk} - \sigma^2_{jk}| \leq \left| \hat{\Sigma}^k_{j,k,jk} - \Sigma^k_{j,k,jk} \right| + 2 \left| \hat{\Sigma}^k_{j,k,jk} - \Sigma^k_{j,k,jk} \right| + \left| \hat{\Sigma}^k_{j,k,jk} - \Sigma^k_{j,k,jk} \right|
$$

where $\hat{\Sigma}^k = \hat{\Sigma}^k(\hat{\beta}'_{j,k})$ and $\hat{\Sigma}^k(\beta_{j,k})$ is defined in (4.19). To prove the consistency of $\hat{\sigma}^2_{jk}$, we need the following theorem to show that $\hat{\Sigma}^k$ is a consistent estimator of $\Sigma^k$ in the sense that $\|\hat{\Sigma}^k - \Sigma^k\|_\infty$ is negligible.

Lemma B.2. For $1 \leq j < k \leq d$, let $\hat{\Sigma}^k(\beta_{j,k})$ be defined as (4.19). Suppose $\hat{\beta}_j$ and $\hat{\beta}_k$ are the estimators of $\beta_j^*$ and $\beta_k^*$ obtained from Algorithm 1 and we denote $\hat{\Sigma}_{jk} = (\hat{\beta}_{jk}, \hat{\beta}_j^T k, \hat{\beta}_k^T j)^T$. Then $\hat{\Sigma}^k(\hat{\beta}_{j,k})$ is a consistent estimator of $\Sigma^k$. There exist a constant $C_\Sigma$ that does not depend on $(j, k)$ such that

$$
\|\hat{\Sigma}^k(\hat{\beta}_{j,k}) - \Sigma^k\|_\infty \leq C_\Sigma(s^* \lambda \log^8 d + \sqrt{\log^9 d/n}) \quad \text{for } 1 \leq j < k \leq d.
$$

Proof of Lemma B.2. See §C.2.2 for a detailed proof. \hfill \Box

In the rest of the proof, we will omit the superscript in both $\hat{\Sigma}^k$ and $\Sigma^k$ for notational simplicity. By Lemma B.2,

$$
I_1 \leq \|\hat{\Sigma} - \Sigma\|_\infty \leq O_p\left(s^* \lambda \log^8 d + \sqrt{\log^9 d/n}\right).
$$

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

$$
I_2 = \left| (\hat{\Sigma}_{j,k} - \Sigma_{j,k}) \right| + \left| (\hat{\Sigma}_{j,k} - \Sigma_{j,k}) \right| + \left| (\hat{\Sigma}_{j,k} - \Sigma_{j,k}) \right|.
$$

By Hölder’s inequality, Lemma B.2 and the estimation error of $\hat{w}_{j,k}$, we obtain an upper-bound for $I_{21}$ as follows:

$$
I_{21} = \|\hat{\Sigma}_{j,k} - \Sigma_{j,k}\|_\infty = O_p\left(s^* \lambda \log^8 d + \sqrt{\log^9 d/n}\right).
$$

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

$$
I_{22} \leq \|\hat{\Sigma}_{j,k} - \Sigma_{j,k}\|_\infty = O_p\left(s^* \lambda \log^8 d + \sqrt{\log^9 d/n}\right),
$$

where the constant $D$ appears in (B.1). For $I_{23}$, by Hölder’s inequality and B.2 we obtain

$$
I_{23} \leq \|\hat{\Sigma}_{j,k} - \Sigma_{j,k}\|_\infty = O_p\left(w_0\left(s^* \lambda \log^8 d + \sqrt{\log^9 d/n}\right)\right).
$$
Combining (B.10), (B.11) and (B.12) we have
\[ I_{2j} \lesssim (w_0 + s^* \lambda_D)(s^* \lambda \log^8 d + \sqrt{\log^9 d/n}) + s^* \lambda_D D \lesssim w_0 s^* \lambda \log^8 d + s^* \lambda_D. \] (B.13)

For term $I_{3j}$, by triangle inequality we have
\[ I_{3j} \leq \left\| \hat{\Sigma}_{j,k} - \Sigma_{j,k,j,k} \right\|_1 \left\| \hat{\Sigma}_{j,k,j,k} \right\|_\infty + \left\| \hat{\Sigma}_{j,k,j,k} - \Sigma_{j,k,j,k} \right\|_1 \lesssim (D \omega_0 s^* \lambda_D + D s^* \lambda_D^2), \] (B.15)

where we use Hölder’s inequality $\| \Sigma_{j,k,j,k} \|_\infty \leq \| \Sigma_{j,k,j,k} \|_1 \leq \| \Sigma \|_\infty \leq Dw_0$. By (B.14), (B.15) and $\lambda_D \gtrsim w_0 s^* \lambda \log^2 d$, we obtain
\[ I_{3j} \lesssim w_0^2 \left( s^* \lambda \log^8 d + \sqrt{\log^9 d/n} \right) + (D \omega_0 s^* \lambda_D + D s^* \lambda_D^2) \lesssim w_0^2 s^* \lambda \log^8 d + w_0 s^* \lambda_D. \] (B.16)

Therefore combining (B.9), (B.13) and (B.16) we obtain $I_1 + I_{2j} + I_{3j} = o_P(1)$. We can show similarly that $I_{2k} + I_{3k} = o_P(1)$. Thus $|\hat{\sigma}_{jk}^2 - \sigma_{jk}^2| \to 0$ with probability converging to 1.

\[ \square \]

### B.5 Proof of Theorem 4.19

**Proof of Theorem 4.19.** Recall that we denote $\beta_{j,k} = (\beta_{j,\ell} \neq k)^T$, $\beta_{j,k} = (\beta_{j,k}, \beta_{j,k}^T)^T$ and $L_j(\beta_{j,k}) = L_j(\beta_j) + L_k(\beta_k)$. We also denote the kernel functions of second-order U-statistic $\nabla L_{jk}(\beta_{j,k})$ as $\mathbf{h}_{ij,k}^T(\beta_{j,k})$. Note that $\hat{S}_{jk}$ and $S_{jk}(\beta_{j,k})$ can be written as
\[ \hat{S}_{jk} = \frac{2}{n(n-1)} \sum_{i \neq i'} \left\{ (1, \hat{\Sigma}_{j,k}^T, \hat{\Sigma}_{k,j}^T) \mathbf{h}_{ij,k}^T(\beta_{j,k}) \right\} \]
and\[ S_{jk}(\beta_{j,k}) = \frac{2}{n(n-1)} \sum_{i \neq i'} \left\{ (1, \Sigma_{j,k}^T, \Sigma_{k,j}^T) \mathbf{h}_{ij,k}^T(\beta_{j,k}) \right\} \]
where $\hat{\beta}_{j,k} = (0, \hat{\beta}_{j,k}, \hat{\beta}_{j,k})$. For notational simplicity, we define
\[ \tilde{\Omega}_{jk} := (1, -\hat{\Sigma}_{j,k}^T, -\hat{\Sigma}_{k,j}^T)^T \quad \text{and} \quad \Omega_{jk} := (1, -\Sigma_{j,k}^T - \Sigma_{k,j}^T)^T. \] (B.17)

Note that we show in Lemma B.1 that $\| \tilde{\Omega}_{jk} - \Omega_{jk} \|_1 \leq Cs^* \lambda_D$ for $1 \leq j < k \leq d$. Recall that we define $g_{jk}(X_i) = \mathbf{h}_{ij,k}^T(\beta_{j,k}^*) = \mathbb{E}(\mathbf{h}_{ij,k}^T(\beta_{j,k}^*)|X_i)$; the Hájek projection of $S_{jk}^*$ is given by
\[ \hat{S}_{jk}^* := \frac{2}{n} \sum_{i=1}^n \tilde{\Omega}_{jk}^T g_{jk}(X_i) = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{1}{n-1} \sum_{i' \neq i} \tilde{\Omega}_{jk}^T \mathbf{h}_{ij,k}^T(\beta_{j,k}^*) \right\}. \] (B.18)
For $\xi_1, \ldots, \xi_n \sim N(0,1)$, we define the bootstrap counterpart of $\hat{S}^*_j$ as
\[
\hat{S}^*_{Bjk} := \frac{2}{n} \sum_{i=1}^{n} \Omega_T^{jk} \mathbf{g}_{jk}(X_i) \xi_i = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{1}{n-1} \sum_{i' \neq i} \Omega_{j}^{jk} \mathbf{h}_{jk}(\beta_{j}^{\psi_{jk}})(\xi_i + \xi_j) \right\}.
\]
(B.19)

Finally, we define $T_0 := \sqrt{n} \max_{ij} |\hat{S}^*_{jk}|/2$ and $W_0 := \sqrt{n} \max_{ij} |\hat{S}^*_{jk}|/2$. To show the validity of the super-graph test, we need the following two auxiliary lemmas: the former shows that $T - T_0$ converges to zero in probability and the latter shows that $W$ is well approximated by $W_0$.

**Lemma B.3.** Let $T = \sqrt{n} \max_{ij} |\hat{S}^*_{jk}|/2$ and $T_0 = \sqrt{n} \max_{ij} |\hat{S}^*_{jk}|/2$ where $S_{jk}$ and $S^*_{jk}$ are defined in (3.8) and (B.18) respectively. Under the null hypothesis $H_0: G^* \subset G$, we have $|T - T_0| \leq \sqrt{n}(s^* + s^*) \lambda \lambda_D$.

**Proof of Lemma B.3.** See §C.2.3 for a detailed proof.

**Lemma B.4.** Let $W = \sqrt{n} \max_{ij} |\hat{S}^*_{jk}|/2$ and $W_0 = \sqrt{n} \max_{ij} |\hat{S}^*_{jk}|/2$ where $\hat{S}^*_{jk}$ and $S^*_{jk}$ are defined in (3.11) and (B.19) respectively. Then we conclude that $|W - W_0| \leq (\log^{11} d/n)^{1/4}$ under the null hypothesis.

**Proof of Lemma B.4.** See §C.2.4 for a detailed proof.

We denote $T := \sqrt{n}(s^* + s^*) \lambda \lambda_D$ and $W := (\log^{11} d/n)^{1/4}$. By Lemma B.3 and Lemma B.4, we have $|T - T_0| \leq T$ and $|W - W_0| \leq W$. By choosing $\zeta_1^2 \asymp \max\{T, W\}/\sqrt{\log d}$ and $\zeta_2 \asymp \max\{T, W\}/\zeta_1$, we obtain from assumption 4.18 and Markov’s inequality that
\[
\zeta_1 \sqrt{\log d} \asymp \zeta_2 \leq Cn^{-c}
\]
for some constant $c > 0$ and $C > 0$. Therefore we conclude that there exists $\zeta_1$ and $\zeta_2$ that both converges to 0 with $\zeta_1 \sqrt{\log d} + \zeta_2 \leq n^{-c}$ such that
\[
P\left(|T - T_0| > \zeta_1\right) < \zeta_2 \quad \text{and} \quad P\left(P_{\xi}(|W - W_0| > \zeta_1) > \zeta_2\right) \leq \zeta_2.
\]

Finally, we show that $T_0$ and $W_0$ satisfy the regularity conditions listed in Chernozhukov et al. (2013), which guarantees that the conditional quantiles of $W$ converges to the quantiles of $T$ uniformly. The theorem is as follows.

**Theorem B.5** (Validity of Gaussian Multiplier Bootstrap). Let $Z_1, \ldots, Z_n$ be $n$ independent $d$-dimensional random vector with $\mathbb{E}(Z_i) = 0$ and let $\xi_1, \ldots, \xi_n$ be i.i.d. $N(0,1)$ random variables. We define
\[
T_0 := n^{-1/2} \max_{j \in [d]} \sum_{i=1}^{n} Z_{ij} \quad \text{and} \quad W_0 := n^{-1/2} \max_{j \in [d]} \sum_{i=1}^{n} Z_{ij} \xi_i.
\]

We assume that there exists two constants $0 < c_1 < C_1$ such that
\[
c_1 \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[Z_{ij}^2\right] \leq C_1 \quad \text{for all} \ 1 \leq j \leq d.
\]
(B.20)
and that there is a sequence of constants \( \{B_n\} \) and two constants \( c_2 > 0 \) and \( C_2 > 0 \) such that
\[
|Z_{ij}| \leq B_n \quad \text{and} \quad B_n^2 \log^7(nd)/n \leq C_2n^{-c_2}.
\] (B.21)

Moreover, we assume that the statistic of interest \( T \) and the corresponding multiplier bootstrap statistic \( W \) approximates \( T_0 \) and \( W_0 \) respectively in the sense that there exist \( \zeta_1 \geq 0 \) and \( \zeta_2 \geq 0 \) depending on \( n \) such that \( \zeta_1 \sqrt{\log d} + \zeta_2 \leq C_3n^{-c_3} \) for some \( c_3, C_3 > 0 \) and
\[
\mathbb{P}(|T - T_0| > \zeta_1) < \zeta_2,
\]
\[
\mathbb{P}\left\{\mathbb{P}\{(W - W_0| > \zeta_1) > \zeta_2\} < \zeta_2.\right\}
\]

Then there exists two constant \( c_2 > 0 \) and \( C_2 > 0 \) depending only on \( c_1, C_1, c_2 \) and \( C_2 \) such that
\[
\rho := \sup_{\alpha \in [0,1]} |\mathbb{P}(T_0 \leq cW_0(\alpha)) - \alpha| \leq Cn^{-c}.
\]

**Proof of Theorem B.5.** See the proof of Theorem 3.2 and Corollary 3.1 in Chernozhukov et al. (2013) for a detailed proof.

Now we verify that condition (B.20) and (B.21) are satisfied. Recall that we denote
\[
g_{jk}(X_i) = \mathbb{E}\{h_{ii'}(\beta_{j'\vee k})|X_i\} = h_{ii'}^{jk}(\beta_{j'\vee k}),
\]
we verify the existence of positive constants \( c_1, C_1, c_2 \) and \( C_2 \) and a sequence of constants \( \{B_n\} \) such that
\[
(i): c_1 \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\Omega_{jk}^Tg_{jk}(X_i)\right]^2 \leq C_1 \quad \text{for all} \quad (j, k) \in E^c \quad \text{and} \quad (B.22)
\]
\[
(ii): \left|\Omega_{jk}^Tg_{jk}(X_i)\right| \leq B_n \quad \text{and} \quad B_n^2 \log^7(nd)/n \leq C_2n^{-c_2}. \quad (B.23)
\]

For (i) in (B.22), note that \( g_{jk}(X_i) \) are i.i.d. random vectors and by the definition of \( \Sigma_{jk} \) we have
\[
\mathbb{E}\left\{\Omega_{jk}^Tg_{jk}(X_i)\right\}^2 = \Omega_{jk}^T\mathbb{E}\left\{g_{jk}(X_i)g_{jk}(X_i)^T\right\}\Omega_{jk} = \Omega_{jk}^T\Sigma_{jk}\Omega_{jk}. \quad (B.24)
\]

Note that by Assumption 4.11, \( \lambda_{\min}(\Sigma_{jk}) \geq c_\Sigma \). We also show that \( \|\Sigma_{jk}\|_{\infty} \leq D \) in (B.1). Note that \( \|\Omega_{jk}\|_2 \geq 1 \) and \( \|\Omega_{jk}\|_1 \leq (1 + w_0)^2 \); the right-hand side of (B.24) can be upper and lower bounded by
\[
c_\Sigma \leq \lambda_{\min}(\Sigma_{jk})\|\Omega_{jk}\|_2 \leq \Omega_{jk}^T\Sigma_{jk}\Omega_{jk} \leq D\|\Omega_{jk}\|_1 \leq (1 + w_0)^2D.
\]

For (ii) in (B.23), by definition, \( g_{jk}(X_i) = h_{ii'}^{jk}(\beta_{j'\vee k}) \); Assumption 4.1 indicates that \( \|g_{jk}(X_i)\|_{\infty} \leq C\log^2 d \) for some generic constant \( C \) that is independent of \( (j, k) \) with probability tending to one. By Hölder’s Inequality, it holds with probability tending to one that
\[
\left|\Omega_{jk}g_{jk}(X_i)\right| \leq \|\Omega_{jk}\|_1\|g_{jk}(X_i)\|_{\infty} \leq C(2w_0 + 1)\log^2 d.
\]

47
Therefore we can choose $c_1 = c_{\Sigma}$, $C_1 = (1 + w_0)^2 D$, $B_n = C(2w_0 + 1) \log^2 d$, Assumption 4.18 implies that there exist a constant $c_2$ such that

$$B_n^2 \log^7 (nd)/n \lesssim \log^{11} d/n \leq n^{-c_2}.$$ 

Therefore (ii) is also satisfied. By Theorem B.5, we conclude that there exists two generic constants $C$ and $c$ such that

$$\sup_{\alpha \in [0,1]} |\mathbb{P}(T' \leq c_{W'}(\alpha)) - \alpha| \leq Cn^{-c},$$

(B.25)

where $T' := \sqrt{n} \max_{E^c} \tilde{S}_{jk}/2$ and $W' := \sqrt{n} \max_{E^c} \tilde{S}^B_{jk}/2$. Note that (B.25) also holds for $T'' := \sqrt{n} \min_{E^c} \tilde{S}_{jk}/2$ and $W'' := \sqrt{n} \min_{E^c} \tilde{S}^B_{jk}/2$, we conclude that (4.20) holds for our test statistic $T = \max\{T', |T''|\}$ and its Bootstrap counterpart $W$. Therefore the super-graph test attains its nominal significance level asymptotically:

$$\lim_{n \to \infty} \sup_{\alpha \in [0,1]} |\mathbb{P}(\psi_G(\alpha) = 1 | H_0) - \alpha| = 0.$$

\[\square\]

C Proof of Auxiliary Results

In this appendix, we prove the auxiliary results. We first prove the auxiliary lemmas presented in Appendix A and Appendix B. Then we give detailed proofs of the technical lemmas.

C.1 Proof of Auxiliary Results for Appendix A

We prove the auxiliary results presented in Appendix A. We first prove Lemma A.3, which bounds the deviation of the Hessian $\nabla^2 L_j(\cdot)$. Then we prove Lemma A.4 and Lemma C.1.3; the former shed light on bounding the $\ell_1$ and $\ell_2$-rates of convergence for estimator $\hat{\beta}_j$ and the latter shows that, in each iteration, the solution of convex relaxation problem remains in the $\ell_1$-ball centered at $\beta_j^*$ where sparse eigenvalue condition holds.

C.1.1 Proof of Lemma A.3

Proof of Lemma A.3. Note that $\nabla^2 L_j(\beta_j)$ is a second-order U-statistic, so is $\nabla^2 L_j(\beta_j) - \mathbb{E}\{\nabla^2 L_j(\beta_j)\}$. We denote its kernel as $T_{ii'}(\beta_j)$, then

$$\nabla^2 L_j(\beta_j) - \mathbb{E}\{\nabla^2 L_j(\beta_j)\} = \frac{2}{n(n-1)} \sum_{i<i'} T_{ii'}(\beta_j).$$

Note that $\left\| \mathbb{E}\{T_{ii'}(\beta_j)\} \right\|_\infty$ is bounded because

$$\left\| \mathbb{E}\{T_{ii'}(\beta_j)\} \right\|_\infty \leq \max_{j \in [d]} \mathbb{E}|X_{ij} - X_{i'j}|^4 \leq \max_{j \in [d], i \in [n]} \mathbb{E}|X_{ij}|^4 \leq \int_0^\infty c \exp(-t^{1/4}) dt = 24c,$$
where \( c = 2\exp(\kappa_m + \kappa_h/2) \). Let \( \nabla^2_{jk,j\ell}L_j(\beta_j) = \partial^2L_j(\beta_j)/\partial\beta_{jk}\partial\beta_{j\ell} \) and let \( [T_{ik}](\beta_j) \) be the corresponding kernel functions. That is, \( \nabla^2_{jk,j\ell}L_j(\beta_j) = \binom{n-1}{2} \sum_{i<i'} [T_{ik}](\beta_j) \). For \( x > 0 \) such that \( x^4 > \|E\{T_{ik} \beta_j)\|_\infty \) and \( k, \ell \neq j \), we have

\[
\mathbb{P}(\|T_{ik} \beta_j)\|_\infty > x^4) \leq \mathbb{P}(\|X_{ij} - X_{ij'}\|_\infty > x^4)
\]

As a direct implication of Assumption 4.1, we have \( \mathbb{P}(\|X_{ij} - X_{ij'}\|_\infty > x) \leq 2\exp(2\kappa_m + \kappa_h)\exp(-x) \) for all \( j \in [d] \). Then we can bound the right-hand side of (C.1) by

\[
\mathbb{P}(\|T_{ik} \beta_j)\|_\infty > x^4) \leq 6\exp(2\kappa_m + \kappa_h)\exp(-x). 
\]

Letting \( C_T = \max\{6\exp(2\kappa_m + \kappa_h), 1\} \), it holds that

\[
\mathbb{P}(\|T_{ik} \beta_j)\|_\infty > x) \leq C_T \exp(-2^{-4/4}x^{1/4}) \quad \text{for all} \ x > 0. 
\]

Now by choosing \( L_1 = C_T, L_2 = 2^{-1/4}, m = \lfloor n/2 \rfloor \) and \( r = 1/4 \), condition (A.3) in Lemma A.1 is satisfied. Then for any \( x \geq 8\text{Var}\left(\nabla^2_{jk,j\ell}L_j(\beta_j)\right)^{1/2} \), we have

\[
\mathbb{P}\left(\|\nabla^2_{jk,j\ell}L_j(\beta_j) - \mathbb{E}\{\nabla^2_{jk,j\ell}L_j(\beta_j)\}\|_\infty > x\right) \leq 4\exp(-m^{1/9}x^{2/9}/8) + 4C_T\left(\frac{n}{2}\right)\exp(-m^{1/9}x^{2/9}/\sqrt{2}). 
\]

Because \( [T_{ik} \beta_j)\|_\infty \) is \( \ell_2 \)-integrable, by the asymptotic normality of U-statistics,

\[
\text{Var}\left(\nabla^2_{jk,j\ell}L_j(\beta_j)\right) = \mathcal{O}(n^{-1}). 
\]

Therefore we let \( x = C(\log^9 d/n)^{1/2} \) and take a union bound over indices \( k \) and \( \ell \) to obtain that

\[
\mathbb{P}\left(\|\nabla^2L_j(\beta_j) - \mathbb{E}\{\nabla^2L_j(\beta_j)\}\|_\infty > C\sqrt{\log^9 d/n}\right) \leq C_1 \exp\left(-C_2 \log d + 2 \log n + 2 \log d\right),
\]

where \( C_1 \) and \( C_2 \) are constants determined by \( C \). Then by choosing a sufficiently large \( C = C_h \) we can show that \( \|\nabla^2L_j(\beta_j) - \mathbb{E}\nabla^2L_j(\beta_j)\|_\infty \leq C_h\sqrt{\log^9 d/n} \) with probability at least \( 1 - (4d)^{-1} \).

For the second argument (A.17), let \( \Delta = \beta_j - \beta^*_j \) where \( \beta_j \in \mathbb{R}^{d-1} \) lies in the \( \ell_1 \)-ball centered at \( \beta^*_j \) with radius \( r_1(s^*, n, d) \), that is, \( \|\beta_j - \beta^*_j\|_1 \leq r_1(s^*, n, d) \). Recall that in (A.2) we show that for any \( j, k \in [d], j \neq k \) and for any \( x > 0 \)

\[
\mathbb{P}(\|X_{ij} - X_{ij'}\|_\infty > x^2) \leq 4\exp(2\kappa_m + \kappa_h)\exp(-x),
\]

taking a union bound over \( 1 \leq j < k \leq d \) and \( 1 \leq i < i' \leq n \) we obtain that

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

$$R_{i'j'}^2 (\beta_j) = \exp \left\{ -(X_{ij} - X_{i'j}) (\Delta + \beta_j^*)^T (X'_{ij} - X'_{i'j}) \right\} = \omega_{i'j'} R_{i'j'}^2 (\beta_j^*).$$

Thus we can write $\nabla^2 L_j (\beta_j)$ as:

$$\nabla^2 L_j (\beta_j) = \frac{2}{n(n-1)} \sum_{i < i'} R_{i'j'}^2 (\beta^*) (X_{ij} - X_{i'j})^2 (X'_{ij} - X'_{i'j}) \omega_{i'j'} (1 + R_{i'j'}^2 (\beta^*))^2 \omega_{i'j'} \left( 1 + \omega_{i'j'} R_{i'j'}^2 (\beta^*) \right)^2. \quad \text{(C.2)}$$

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

$$\min \{ \omega_{i'j'}, 1/\omega_{i'j'} \} \leq \frac{\omega_{i'j'} (1 + R_{i'j'}^2 (\beta^*))^2 \omega_{i'j'} \left( 1 + \omega_{i'j'} R_{i'j'}^2 (\beta^*) \right)^2}{\min \{ \omega_{i'j'}, 1/\omega_{i'j'} \}} \leq \max \{ \omega_{i'j'}, 1/\omega_{i'j'} \}. \quad \text{(C.3)}$$

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

$$\exp(-b) \leq \min \{ \omega_{i'j'}, 1/\omega_{i'j'} \} \leq \max \{ \omega_{i'j'}, 1/\omega_{i'j'} \} \leq \exp(b). \quad \text{(C.4)}$$

Combining (C.2), (C.3) and (C.4) we obtain

$$\exp(-b) \nabla^2 L_j (\beta_j^*) \leq \nabla^2 L_j (\beta_j) \leq \exp(b) \nabla^2 L_j (\beta_j^*). \quad \text{(C.5)}$$

Then by (C.5) we have

$$\| \nabla^2 L_j (\beta_j) - \nabla^2 L_j (\beta_j^*) \| \leq \max \left\{ 1 - \exp(-b), \exp(b) - 1 \right\} \| \nabla^2 L_j (\beta_j^*) \| \leq b \| \nabla^2 L_j (\beta_j^*) \|.$$

Notice that under Assumption 4.1, as shown in §B.1, we can assume that $\| \mathbb{E} \{ \nabla^2 L_j (\beta_j^*) \} \| \leq D$ where $D$ appears in (B.1). By triangle inequality,

$$\| \nabla^2 L_j (\beta_j^*) \| \leq \| \nabla^2 L_j (\beta_j^*) - \mathbb{E} \{ \nabla^2 L_j (\beta_j^*) \} \| + \| \mathbb{E} \{ \nabla^2 L_j (\beta_j^*) \} \| \leq D + C_h \sqrt{\log^9 d/n} \leq 2D$$

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

$$\| \nabla^2 L_j (\beta_j) - \nabla L_j (\beta_j^*) \| \leq 2 C r_1(s^*, n, d) \log^2 d$$

holds with probability at least $1 - (2d)^{-1}$. \qed
C.1.2 Proof of Lemma A.4

Proof of Lemma A.4. In what follows, for notational simplicity and readability, we omit \( j \) in the subscript and \( \ell \) in the superscript by simply writing \( S_j, G_j^\ell, J_j^\ell \) and \( I_j^\ell \) as \( S, G, J \) an \( I \) respectively. By the definition of \( G \), \( \| \lambda_{C}^{(\ell-1)} \|_{\infty} \geq p_j^{(\ell)}(\theta) \geq 0.91A > 22.75\| \nabla L_j(\beta_j^*) \|_{\infty} \). We prove this lemma in two steps. In the first step we show that \( \| \tilde{\beta}_j^{(\ell)} - \beta_j^* \|_{1} \leq 2.2\| \tilde{\beta}_G^{\ell} - \beta_G^* \|_{1} \). Suppose that \( \tilde{\beta}_j^{(\ell)} \) is the solution in the \( \ell \)-th iteration and we denote \( \nabla_{jk} L_j(\beta_j) = \partial L_j(\beta_j)/\partial \beta_j^{jk} \), the Karush-Kuhn-Tucker condition implies that

\[
\nabla_{jk} L_j(\tilde{\beta}_j^{(\ell)}) + \lambda_{jk}^{(\ell-1)} \text{sign}(\tilde{\beta}_j^{(\ell)}) = 0 \quad \text{if} \quad \tilde{\beta}_j^{(\ell)} \neq 0;
\]

\[
\nabla_{jk} L_j(\tilde{\beta}_j^{(\ell)}) + \lambda_{jk}^{(\ell-1)} \xi_j^{(\ell)} = 0, \quad \xi_j^{(\ell)} \in [-1, 1] \quad \text{if} \quad \tilde{\beta}_j^{(\ell)} = 0.
\]

The Karush-Kuhn-Tucker conditions can be written in a compact form as

\[
\nabla L_j(\tilde{\beta}_j^{(\ell)}) + \lambda_j^{(\ell-1)} \circ \xi_j^{(\ell)} = 0,
\]

where \( \xi_j^{(\ell)} \in \partial \| \tilde{\beta}_j^{(\ell)} \|_1 \) and \( \lambda_j^{(\ell-1)} = (\lambda_{jk}^{(\ell-1)}; k \neq j)^T \).

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

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

\[
\nabla L_j(\tilde{\beta}_j) - \nabla L_j(\beta_j^*) = \nabla^2 L_j(\tilde{\beta}_j) \delta.
\]

Then (C.6) implies that

\[
0 \leq \delta^T \nabla^2 L_j(\tilde{\beta}_j) \delta = -\left( \delta, \lambda \circ \xi \right) - \left( \nabla L_j(\beta_j^*), \delta \right). \tag{C.7}
\]

For term (ii) in (C.7), Hölder’s inequality implies that

\[
(ii) \geq -\| \nabla L_j(\beta_j^*) \|_{\infty} \| \delta \|_1. \tag{C.8}
\]

For term (i) in (C.7), recall that we denote \( |v| \) as the vector that takes entrywise absolute value for \( v \); by the fact that \( \xi_j^{(\ell)} \tilde{\beta}_j^{(\ell)} = |\tilde{\beta}_j^{(\ell)}| \), we have \( \xi_G \circ \delta_G = |\delta_G| \) and \( \xi_H \circ \delta_H = |\delta_H| \). Since \( \delta_S = \tilde{\beta}_j^{(\ell)} \). Hölder’s inequality implies that

\[
\langle \delta, \lambda \circ \xi \rangle = \langle \delta_S, (\lambda \circ \xi)_S \rangle + \langle |\delta_H|, \lambda_H \rangle + \langle |\delta_G|, \lambda_G \rangle \geq -\| \delta_S \|_1 \| \lambda_S \|_{\infty} + \| \delta_G \|_1 \| \lambda_G \|_{\min} + \| \delta_H \|_1 \| \lambda_H \|_{\min}. \tag{C.9}
\]

Combining (C.7), (C.8) and (C.9) we have

\[
-\| \delta_S \|_1 \| \lambda_S \|_{\infty} + \| \delta_G \|_1 \| \lambda_G \|_{\min} + \| \delta_H \|_1 \| \lambda_H \|_{\min} - \| \nabla L_j(\beta_j^*) \|_{\infty} \| \delta \|_1 \leq 0. \tag{C.10}
\]

51
By the definition of \( G \), we have \( \| \lambda G \|_{\infty} \geq p'_{\lambda}(c_2 \lambda) \geq 0.91 \lambda \). Rearranging terms in (C.10) we have
\[
\frac{\lambda G}{\lambda} = p'_{\lambda}(c_2 \lambda) \leq \| \delta G \|_{1} \leq \| \nabla L_j(\beta^*_j) \|_{\infty} \| \delta \|_{1} + \| \delta s \|_{1} \| \lambda s \|_{\infty}.
\]
Using the decomposability of \( \ell_1 \)-norm, we have
\[
\left( \frac{p'_{\lambda}(c_2 \lambda)}{p'_{\lambda}(\theta)} - \| \nabla L_j(\beta^*_j) \|_{\infty} \right) \| \delta G \|_{1} \leq \| \lambda s \|_{\infty} + \| \nabla L_j(\beta^*_j) \|_{\infty} \| \delta G^c \|_{1}.
\]
Recall that \( \lambda > 25 \| \nabla L_j(\beta^*_j) \|_{\infty} \) and \( p'_{\lambda}(\theta) \geq 0.91 \lambda \), (C.11) implies
\[
\| \delta G \|_{1} \leq \frac{\lambda + \| \nabla L_j(\beta^*_j) \|_{\infty}}{p'_{\lambda}(c_2 \lambda) - \| \nabla L_j(\beta^*_j) \|_{\infty}} \| \delta G^c \|_{1} \leq 1.2 \| \delta G^c \|_{1},
\]
where we use the fact that
\[
\frac{\lambda + \| \nabla L_j(\beta^*_j) \|_{\infty}}{p'_{\lambda}(c_2 \lambda) - \| \nabla L_j(\beta^*_j) \|_{\infty}} \leq \frac{\lambda + 0.04 \lambda}{0.91 \lambda - 0.04 \lambda} \leq 1.2.
\]
Going back to the original notation, (C.12) is equivalent to
\[
\| \tilde{\beta}^{(f)}_j - \beta^*_j \|_{1} \leq 1.2 \| \tilde{\beta}^{(f)}_j - \beta^*_j \|_{1}.
\]
Now we show in the second step that \( \| \tilde{\beta}^{(f)}_j - \beta^*_j \|_{2} \leq 2.2 \| \tilde{\beta}^{(f)}_j - \beta^*_j \|_{2} \). Recall that \( J \) is the largest \( k^* \) components of \( \tilde{\beta}^{(f)}_G \) in absolute value where we omit the subscript \( j \) and superscript \( \ell \) in the sets \( G^f_j, J^f_j \) and \( I^f_j \); by the definition of \( J \) we obtain that
\[
\| \delta I^c \|_{\infty} \leq \| \delta I \|_{1}/k^* \leq \| \delta G \|_{1}/k^*, \quad \text{where } \delta = \tilde{\beta}^{(f)}_j - \beta^*_j.
\]
By inequality (C.12) and the fact that \( G^c \subset I \), we further have
\[
\| \delta I^c \|_{\infty} \leq 1.2/k^* \| \delta G^c \|_{1} \leq 1.2/k^* \| \delta I \|_{1}.
\]
Then by Hölder’ inequality and (C.13) we obtain that
\[
\| \delta I^c \|_{2} \leq \left( \| \delta I^c \|_{1} \| \delta I^c \|_{\infty} \right)^{1/2} \leq (1.2/k^*)^{1/2} \left( \| \delta I \|_{1} \| \delta I \|_{1} \right)^{1/2}.
\]
By the definition of index sets \( G \) and \( I \), we have \( I^c \subset G \) and \( G^c \subset I \). Then by (C.12) and (C.14) we obtain
\[
\| \delta I^c \|_{2} \leq (1.2/k^*)^{1/2} \left( \| \delta G^c \|_{1} \| \delta G \|_{1} \right)^{1/2} \leq 1.2 \| \delta G^c \|_{1}/\sqrt{k^*}.
\]
By the norm inequality between \( \ell_1 \)-norm and \( \ell_2 \)-norm, we have
\[
\| \delta I^c \|_{2} \leq 1.2 \| \delta G^c \|_{1}/\sqrt{k^*} \leq 1.2 \sqrt{2s^*/k^*} \| \delta G^c \|_{2},
\]
where we use \( k^* \geq 2s^* \) and the induction assumption that \( |G| \leq 2s^* \). Then triangle inequality for \( \ell_2 \)-norm yields that
\[
\| \delta \|_{2} \leq \| \delta I^c \|_{2} + \| \delta I \|_{2} \leq 2.2 \| \delta I \|_{2}.
\]
Note that (C.12) and (C.15) are equivalent to
\[
\| \tilde{\beta}^{(f)}_j - \beta^*_j \|_{2} \leq 2.2 \| \tilde{\beta}^{(f)}_j - \beta^*_j \|_{2} \quad \text{and} \quad \| \tilde{\beta}^{(f)}_j - \beta^*_j \|_{1} \leq 2.2 \| \tilde{\beta}^{(f)}_j - \beta^*_j \|_{1},
\]
where \( \tilde{G}^f_j = (G^f_j)^c \), which concludes the proof.
C.1.3 Proof of Lemma A.5

Proof of Lemma A.5. We first show that \( \hat{\beta}_j^{(t)} \) stays in the \( \ell_1 \)-ball centered at \( \beta_j^* \) with radius \( r \) where \( r \) appears in Assumption 4.4. For notational simplicity, we denote \( \delta = \hat{\beta}_j^{(t)} - \bar{\beta}_j \) and write \( S_j, G_j^t, J_j^t \) and \( I_j^t \) as \( S, G, J \) an I respectively. We prove by contradiction. Suppose that \( \|\delta\|_1 > r \), then we can define \( \bar{\beta}_j = \beta_j^* + t(\hat{\beta}_j^{(t)} - \beta_j^*) \in \mathbb{R}^{d-1} \) with \( t \in [0,1] \) such that \( \|\bar{\beta}_j - \beta_j^*\|_1 \leq r \). Let \( \bar{\delta} := \bar{\beta}_j - \beta_j^* \), then by (C.15) we obtain

\[
\|\bar{\delta}\|_2 = t\|\delta\|_2 \leq 2.2t\|\delta_i\|_2 = 2.2\|\bar{\delta}_i\|_2. \tag{C.16}
\]

Moreover, by Lemma (A.4) and the relation between \( \ell_1 \)-norm and \( \ell_2 \)-norm we have

\[
\|\bar{\delta}\|_1 = t\|\delta\|_1 \leq 2.2t\|\delta G^c\|_1 \leq 2.2\sqrt{2s^*}\|\bar{\delta}_i\|_2, \tag{C.17}
\]

where we use the fact that \( G^c \subseteq I \) and the induction assumption that \( |G^c| \leq 2s^* \). By Mean-Value theorem, there exist a \( \gamma \in [0,1] \) such that \( \nabla L_j(\bar{\beta}_j) - \nabla L_j(\beta_j^*) = \nabla^2 L_j(\beta_1)\bar{\delta} \), where \( \beta_1 := \gamma \beta_j^* + (1 - \gamma)\bar{\beta}_j \in \mathbb{R}^{d-1} \). In what follows we will derive an upper bound for \( \|\bar{\delta}_j\|_2^2 \) from \( \bar{\delta}_i^T \nabla^2 L_j(\beta_1)\bar{\delta} \). Before doing that, we present two lemmas that analyze the property of restricted correlation coefficients. These two lemmas also appear in Zhang (2010) and Zhang et al. (2013) for \( \ell_2 \)-loss. The first lemma shows that the restricted correlation coefficients defined as follows are closely related to the sparse eigenvalues.

**Lemma C.1.** Let \( M : \mathbb{R}^m \rightarrow S^m \) be a mapping from \( \mathbb{R}^m \) to the space of \( m \times m \)-symmetric matrices. We define the restricted correlation coefficients of \( M \) over the \( \ell_1 \)-ball centered at \( u_0 \) with radius \( r \) as

\[
\pi(M(\cdot), u_0; s, k, r) := \sup_{v \cdot w \in \mathbb{R}^m} \left\{ \frac{v^T M(u) w}{v^T M(u) v} \right\} : I \cap J = \emptyset, |I| \leq s, |J| \leq k, \|u - u_0\|_1 \leq r \}
\]

Suppose that the local sparse eigenvalue \( \rho_-(M(u_0; s+k, r) > 0 \), then we have the following upper bound on the restricted correlation coefficient \( \pi(M(\cdot), u_0; s, k) \):

\[
\pi(M(\cdot), u_0; s, k, r) \leq \frac{\sqrt{k}}{2} \sqrt{\rho_+(M(\cdot), u_0; k, r) \rho_- (M(\cdot), u_0; s+k, r) - 1}.
\]

**Proof of Lemma C.1.** See §C.3.1 for a detailed proof. \( \square \)

We denote the restricted correlation coefficients of \( \nabla^2 L_j(\cdot) \) over the \( \ell_1 \)-ball centered at \( \beta_j^* \) with radius \( r \) as \( \pi_j(s_1, s_2) := \pi(\nabla^2 L_j(\cdot), \beta_j^*; s_1, s_2, r) \) and denote the \( s \)-sparse eigenvalues \( \rho_-(\nabla^2 L_j(\cdot), \beta_j^*; s, r) \) and \( \rho_+(\nabla^2 L_j(\cdot), \beta_j^*; s, r) \) as \( \rho_j-(s) \) and \( \rho_j+(s) \) respectively. Applying Lemma C.1 to \( \pi_j(2s^*+k^*, s^*) \) we obtain

\[
\pi_j(2s^*+k^*, s^*) \leq k^{1/2} \sqrt{\rho_j+(s^*)/\rho_j-(2s^*+k^*) - 1/2}, \tag{C.18}
\]

then Assumption 4.4 implies

\[
\pi_j(2s^*+k^*, s^*) \leq 0.5 \sqrt{0.27k^*/s^*}. \tag{C.19}
\]

53
By (C.12), (C.19) and $G^c \subset I$ we obtain
\[
1 - 2\pi_j(2s^*+k^*, k^*)k^* - 1 \|\tilde{\delta}_G\|_1 / \|\tilde{\delta}_I\|_2 \geq 1 - 1.2\sqrt{0.54} := \kappa_1,
\]
where we denote $\kappa_1 := 1 - 1.2\sqrt{0.54} \geq 0.11$. Now we use the second lemma to get an lower bound of $\tilde{\delta}^T \nabla^2 L_j(\beta_1)\tilde{\delta}$, which implies an upper bound for $\|\tilde{\delta}_I\|_2$.

**Lemma C.2.** Let $M : \mathbb{R}^m \rightarrow \mathbb{S}^m$ be a mapping from $\mathbb{R}^m$ to the space of $m \times m$-symmetric matrices. Suppose that the sparse eigenvalue $\rho_-(M(\cdot), u_0; s+k, r) > 0$, let the restricted correlation coefficients of $M(\cdot)$ be defined in Lemma C.1. We denote the restricted correlation coefficients $\pi(M(\cdot), u_0; s, k, r)$ and $s$-sparse eigenvalue $\rho_-(M(\cdot), u_0; s, r)$ as $\pi(s, k)$ and $\rho_-(s)$ respectively for notational simplicity. For any $v \in \mathbb{R}^d$, let $F$ be any index set such that $|F^c| \leq s$, let $J$ be the set of indices of the largest $k$ entries of $v_F$ in absolute value and let $I = F^c \cup J$. For any $u \in \mathbb{R}^d$ such that $\|u - u_0\|_2 \leq r$ and any $v \in \mathbb{R}^d$ satisfying $1 - 2\pi(s+k, k)\|v_F\|_1 / \|v_F\|_2 > 0$ we have
\[
v^T M(u) v \geq \rho_- (s+k) \left( \|v_I\|_2 - 2\pi(s+k, k) \|v_F\|_1 / \|v_F\|_2 \right) \|v_I\|_2.
\]

**Proof of Lemma C.2.** See §C.3.2 for a detailed proof. ☐

Now applying Lemma C.2 to $\nabla^2 L_j(\cdot)$ with $F=G$, $s=2s^*$ and $k=k^*$ we obtain
\[
\tilde{\delta}^T \nabla^2 L_j(\beta_1)\tilde{\delta} \geq \rho_j - (2s^*+k^*)\|\tilde{\delta}_I\|_2 \|\tilde{\delta}_I\|_2 - 2\pi_j(2s^*+k^*, k^*) / \|\tilde{\delta}_G\|_1
\]
\[
(C.21)
\]
Then by (C.20), the right-hand side of (C.21) can be lower bounded by
\[
\tilde{\delta}^T \nabla^2 \ell_j(\delta) \geq \kappa_1 \rho_j - (2s^*+k^*) \|\tilde{\delta}_I\|_2^2 \geq 0.95\kappa_1 \rho_j \|\tilde{\delta}_I\|_2^2 = \kappa_2 \rho_j \|\tilde{\delta}_I\|_2^2
\]
where we let $\kappa_2 := 0.95\kappa_1 \geq 0.1$. Now we derive an upper bound for $\tilde{\delta}^T \nabla^2 L_j(\beta_1)\tilde{\delta}$. We define the symmetric Bregman divergence of $L_j(\beta)$ as $D_j(\beta_1, \beta_2) := \langle \beta_1 - \beta_2, \nabla L_j(\beta_1) - \nabla L_j(\beta_2) \rangle$, where $\beta_1, \beta_2 \in \mathbb{R}^{d-1}$. Then by definition, $\tilde{\delta}^T \nabla^2 \ell_j(\delta) = D_j(\tilde{\beta}_j, \beta_j^*)$. The following lemma relates $D_j(\tilde{\beta}_j, \beta_j^*)$ with $D_j(\tilde{\beta}_j, \beta_j^*)$.

**Lemma C.3.** Let $D_j(\beta_1, \beta_2) := \langle \beta_1 - \beta_2, \nabla L_j(\beta_1) - \nabla L(\beta_2) \rangle$, $\beta(t) = \beta_1 + t(\beta_2 - \beta_1)$, $t \in (0, 1)$ be any point on the line segment between $\beta_1$ and $\beta_2$. Then we have
\[
D_j(\beta(t), \beta_1) \leq t D_j(\beta_2, \beta_1)
\]

**Proof of Lemma C.3.** See §C.3.3 for a detailed proof. ☐

By Lemma C.3 and (C.7),
\[
D_j(\tilde{\beta}_j, \beta_j^*) \leq t D_j(\tilde{\beta}_j, \beta_j^*) \leq -t \langle \nabla L_j(\beta_j^*), \delta \rangle - t \langle \delta, \lambda_j \circ \xi_j \rangle \tag{i}
\]
\[
(C.23)
\]
For term (i) in (C.23), by Hölder’s inequality we have
\[
-t \langle \nabla L_j(\beta_j^*), \delta \rangle \leq t \|\nabla G \cdot L_j(\beta_j^*)\|_2 \|\delta_G\|_2 + t \|\nabla G L_j(\beta_j^*)\|_\infty \|\delta_G\|_1
\]
\[
\leq \|\nabla G \cdot L_j(\beta_j^*)\|_2 \|\delta_I\|_2 + \|\nabla G L_j(\beta_j^*)\|_\infty \|\delta_G\|_1,
\]
\[
(C.24)
\]
where the inequality follows from $G^c \subset I$. For term (ii) in (C.23), by (C.9) and Hölder’s inequality we have
\[ -t\langle \delta, \lambda_j \circ \xi_j \rangle \leq -\langle \delta_S, (\lambda_j \circ \xi_j)_S \rangle - \langle |\tilde{G}_G|, \lambda_G \rangle \leq \|\lambda_S\|_2 \|\tilde{\delta}_I\|_2 - p'_\lambda(c_2 \lambda) \|\tilde{G}_G\|_1, \]
(C.25)
where we use the Hölder’s inequality and the definition of $G$. Combining (C.22),(C.24) and (C.25) we obtain
\[ \kappa_2 \rho_* \|\tilde{\delta}_I\|_2^2 \leq \left( \|\nabla_{G^c} L_j(\beta_j^*)\|_2 + \|\lambda_S\|_2 \right) \|\tilde{\delta}_I\|_2 + \left( \|\nabla L_j(\beta_j^*)\|_\infty - p'_\lambda(c_2 \lambda) \right) \|\tilde{G}_G\|_1 \]
\[ \leq \left( \|\nabla_{G^c} L_j(\beta_j^*)\|_2 + \|\lambda_S\|_2 \right) \|\tilde{\delta}_I\|_2, \]
where the second inequality follows from $p'_\lambda(c_2 \lambda) > \|\nabla L_j(\beta_j^*)\|_\infty$. From the inequality above and the induction assumption $|G^c| \leq 2s^*$ we obtain
\[ \|\tilde{\delta}_I\|_2 \leq 10\rho_*^{-1} \left( \|\nabla_{G^c} L_j(\beta_j^*)\|_2 + \|\lambda_S\|_2 \right) \leq 10\rho_*^{-1} \sqrt{s^*} \left( \sqrt{2} \|\nabla L_j(\beta_j^*)\|_\infty + \lambda \right). \]
(C.26)
Thus (C.17), (C.26) and the fact that $25\|\nabla L_j(\beta_j^*)\|_\infty \leq \lambda$ imply that
\[ \|\tilde{\delta}\|_1 \leq 22 \sqrt{2}\rho_*^{-1} (1 + \sqrt{2}/25) s^* \lambda < 33\rho_*^{-1} s^* \lambda < r, \]
(C.27)
where the last inequality follows from the definition of $\lambda$. Notice that (C.27) contradicts our assumption that $\|\tilde{\delta}\|_1 = r$, the reason for this contradiction is because we assume that $\|\tilde{\beta}_j^{(\ell)} - \beta_j^*\|_1 > r$, hence $\|\tilde{\beta}_j^{(\ell)} - \beta_j^*\|_1 \leq r$ and $\tilde{\beta}_j = \tilde{\beta}_j^{(\ell)}$. This means that $\tilde{\beta}_j^{(\ell)}$ stays in the $\ell_1$-ball centered at $\beta_j^*$ with radius $r$ in each iteration.

Moreover, by (C.15) and (C.26), we obtain the following upper bound for $\|\tilde{\delta}_I\|_2$:
\[ \|\tilde{\delta}_I\|_2 \leq 22 \rho_*^{-1} \left( \|\nabla_{G^c} L_j(\beta_j^*)\|_2 + \|\lambda_S\|_2 \right) \leq 24 \rho_*^{-1} \sqrt{s^*} \lambda, \]
where we use the condition that $\lambda \geq 25\|\nabla L_j(\beta_j^*)\|_\infty$. In addition, by (C.12) and (C.26) we obtain the following bound on $\|\tilde{\delta}\|_1$
\[ \|\tilde{\delta}\|_1 \leq 2.2 \|\tilde{\delta}_{G^c}\|_1 \leq 22 \sqrt{2s^*} \rho_*^{-1} \left( \|\nabla_{G^c} L_j(\beta_j^*)\|_2 + \|\lambda_S\|_2 \right) \leq 33\rho_*^{-1} s^* \lambda, \]
(C.28)
Therefore going back to the original notations, note that $\kappa_2 \geq 0.1$, we establish the following crude rates of convergence for $\ell \geq 1$:
\[ \|\tilde{\beta}_j^{(\ell)} - \beta_j^*\|_2 \leq 24 \rho_*^{-1} \sqrt{s^*} \lambda \quad \text{and} \quad \|\tilde{\beta}_j^{(\ell)} - \beta_j^*\|_1 \leq 33\rho_*^{-1} s^* \lambda. \]
And (C.26) is equivalent to
\[ \|\tilde{\beta}_j^{(\ell)} - \beta_j^*\|_2 \leq 10\rho_*^{-1} \left( \|\nabla_{G} L_j(\beta_j^*)\|_2 + \|\lambda_{S_j}^{(\ell-1)}\|_2 \right), \quad \tilde{G}_j^{\ell} := (G_j^c)^c. \]
C.2 Proof of Auxiliary Results for Appendix B

In this subsection, we prove the auxiliary results presented in Appendix B. We first prove Lemma B.1, which establishes the rate of convergence for \( \hat{\omega}_{j,k} \). Then we prove Lemma B.2, which is essential for deriving a consistent estimator of \( \sigma^2_{jk} \), the asymptotic variance of the pairwise score statistic \( \hat{S}_{jk} \). Finally we prove Lemma B.3 and Lemma B.4, which bound \( |T - T_0| \) and \( |W - W_0| \) respectively.

C.2.1 Proof of Lemma B.1

Proof of Lemma B.1. By the definition of \( \omega^*_{j,k} \) we have \( H^j_{jk,j\backslash k} = w^*_{j,k} H^j_{jk,j\backslash k} \). We let \( \beta^*_j = (0, \hat{\beta}^*_j) \) and denote \( \nabla^2 L_j(\hat{\beta}') \) and \( \nabla^2 L_j(\beta^*_j) \) as \( \Lambda \) and \( \Lambda^* \) respectively. In addition, we write \( H^j, w^*_{j,k} \) and \( \hat{\omega}_{j,k} \) as \( H, w^* \) and \( \hat{\omega} \) respectively for notational simplicity. Triangle inequality implies that

\[
\|A_{jk,j\backslash k} - w^* H_{j,k\backslash j}\|_2 \leq \|H_{j,k\backslash j} - A_{jk,j\backslash k}\|_2 + \|w^* H_{j,k\backslash j} - A_{jk,j\backslash k}\|_2.
\]

Hölder’s inequality implies that

\[
\|A_{jk,j\backslash k} - w^* H_{j,k\backslash j}\|_2 \leq \|A - H\|_\infty (1 + \|w^*\|_1).
\]  

(C.29)

Under null hypothesis, \( \beta^*_j = 0 \). By Lemma A.3, we have \( \|A - H\|_\infty \leq s^* \lambda \log^2 d \). Then the right-hand side of (C.29) is bounded by

\[
\|A_{jk,j\backslash k} - w^* H_{j,k\backslash j}\|_2 \lesssim s^* w_0 \lambda \log^2 d.
\]

Therefore, by the assumption that \( \lambda D \geq s^* w_0 \lambda \log^2 d \) we can ensure that \( w^* \) is in the feasible region of the Dantzig selector problem (3.7), hence we have \( \|w^*\|_1 \leq w_0 \) by the optimality of \( \hat{\omega} \). Let \( J \) be the support set of \( w^* \), that is, \( J := \{j, \ell\} : [w^*_{j,k}]_{j\ell} \neq 0, \ell \in [d], \ell \neq j\}; the optimality of \( w^* \) is equivalent to \( \|\hat{\omega}_{j^c} - 1\| + \|\hat{\omega}_j\|_1 \leq \|w^*_{j}\|_1 \). By triangle inequality, we have

\[
\|\hat{\omega}_{j^c} - w^*_{j}\|_1 = \|\hat{\omega}_{j^c}\|_1 - \|\hat{\omega}_j\|_1 \leq \|\hat{\omega}_j - w^*_{j}\|_1,
\]  

(C.30)

where \( j^c := \{(j, \ell): (j, \ell) \notin J, j \text{ fixed}\} \). Letting \( \tilde{\omega} = \hat{\omega} - w^* \), inequality (C.30) is equivalent to \( \|\tilde{\omega}_{j^c}\|_1 \leq \|\tilde{\omega}_j\|_1 \). Moreover, triangle inequality yields that

\[
\|A_{j,k\backslash j}\|_\infty \leq \|A_{j,k\backslash j} - A_{j,k\backslash j}\|_\infty + \|A_{j,k\backslash j} - A_{j,k\backslash j} w^*\|_\infty \leq 2 \lambda D,
\]

where the last inequality follows from that both \( w^* \) and \( \hat{\omega} \) are feasible for the Dantzig selector problem 3.7. Then triangle inequality implies

\[
|\tilde{\omega}^T A_{j,k\backslash j}\| \leq |\tilde{\omega}^T A_{j,k\backslash j}\|_{A_1} + |\tilde{\omega}^T A_{j,k\backslash j}\|_{A_2}.
\]

By Hölder’s inequality and inequality between \( \ell_1 \)-norm and \( \ell_2 \)-norms, we obtain that

\[
A_1 \leq 2 \lambda D \|\tilde{\omega}_j\|_1 \leq 2 s^* \lambda D \|\tilde{\omega}_j\|_2 \quad \text{and} \quad A_2 \leq 2 \lambda D \|\tilde{\omega}_j\|_1 \leq 2 \lambda D \|\tilde{\omega}_j\|_1 \leq 2 s^* \lambda D \|\tilde{\omega}_j\|_2.
\]
Hence we conclude that $|\hat{\omega}^T \Lambda_{j \setminus k, j} \hat{\omega}| \leq 4\sqrt{s^*} \lambda_D \|\hat{\omega}_J\|_2$.

We let $J_1$ be the set of indices of the largest $k^*$ component of $\hat{\omega}_{J_1}$ in absolute value and let $I = J_1 \cup J$, then $|I| \leq s^*+k^*$. Under the null hypothesis, $\|\hat{\beta}_j - \beta_j^*\|_1 = \|\hat{\beta}_{j \setminus k} - \beta_{j \setminus k}^*\|_1 = O(s^* \lambda) \leq r$.

We denote the $s$-sparse eigenvalue of $\nabla^2_{j \setminus k, j} L_j(\beta_j)$ over the $\ell_1$-ball centered at $\beta_j^*$ 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 $\nabla^2 L_j(\beta_j^*)$ as $\pi_j(-s), \rho_{j^+}(s)$ 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 C.2 we have

$$\left|\hat{\omega}^T \Lambda_{j \setminus k, j} \hat{\omega}\right| \geq \rho_{j^+}^s(\lambda^*+s^*) \left(\|\hat{\omega}_I\|_2 - 2\pi^s_j(\lambda^*+s^*) \|\hat{\omega}_{J_1^*}\|_2\right)\|\hat{\omega}_I\|_2.$$  \hspace{1cm} (C.31)

By $\|\hat{\omega}_{J^*}\|_1 \leq \|\hat{\omega}_J\|_1 \leq \sqrt{s^*} \|\hat{\omega}_J\|_2$ and Lemma 4.15, the right-hand side of (C.31) can be reduced to

$$\left|\hat{\omega}^T \Lambda_{j \setminus k, j} \hat{\omega}\right| \geq 0.95 \nu_s \left(\|\hat{\omega}_I\|_2 - 2\pi^s_j(\lambda^*+s^*) \|\hat{\omega}_{J_1^*}\|_2\sqrt{s^*/k^*}\right)\|\hat{\omega}_I\|_2.$$  \hspace{1cm} (C.32)

Using Lemma C.1 we obtain

$$2\pi^s_j(\lambda^*+s^*) \sqrt{s^*/k^*} \leq \sqrt{s^*/k^*} \sqrt{\rho_{j^+}^s(\lambda^*)/\rho_{j^-}^s(\lambda^*+s^*) - 1} \leq \sqrt{s^*/k^*} \sqrt{0.58k^*/s^*} \leq 0.76.$$  \hspace{1cm} (C.33)

Thus the right-hand side of (C.32) can be reduced to

$$\left|\hat{\omega}^T \Lambda_{j \setminus k, j} \hat{\omega}\right| \geq 0.95 \nu_s (1 - 0.76) \left(\|\hat{\omega}_I\|_2 - 2\pi^s_j(\lambda^*+s^*) \|\hat{\omega}_{J_1^*}\|_2\right)\|\hat{\omega}_I\|_2 \geq \nu_s \kappa \|\hat{\omega}_I\|_2^2,$$

where $\kappa = 0.22$. This inequality holds because $J \subset I$. By (C.33) we have

$$\nu_s \kappa \|\hat{\omega}_I\|_2^2 \leq 4 \sqrt{s^*} \lambda_d \|\hat{\omega}_J\|_2 \leq 4 \sqrt{s^*} \lambda_d \|\hat{\omega}_I\|_2,$$  which implies $\|\hat{\omega}_I\|_2 \leq 4 \nu_s^{-1} \kappa^{-1} \sqrt{s^*} \lambda_D$.

Therefore the estimation error of $\hat{\omega}_{j,k}$ can be bounded by

$$\|\hat{\omega}_I\|_2^2 \leq 2 \|\hat{\omega}_J\|_2 \leq 2 \sqrt{s^*} \|\hat{\omega}_J\|_2 \leq 8 \nu_s^{-1} \kappa^{-1} s^* \lambda_D \leq 37 \nu_s^{-1} s^* \lambda_D.$$  \hspace{1cm} (C.34)

Returning to the original notations, we conclude that $\|\hat{\omega}_{j,k} - \omega^*_{j,k}\|_1 \leq 37 \nu_s^{-1} s^* \lambda_D$ for all $(j,k)$ such that $j,k \in [d]$, $j \neq k$. \hfill \Box

C.2.2 Proof of lemma B.2

Proof of lemma B.2. Recall that we denote $\beta_{j \setminus k} = (\beta_{j \setminus \ell} \neq k)^T$ and $\beta_{j \setminus k} = (\beta_{j \setminus k}, \beta_{\setminus j \setminus k})$. Also recall that $L_{j,k}(\beta_{j \setminus k})$ is defined to be $L_j(\beta_j) + L_k(\beta_k)$. We denote the kernel function of the second-order U-statistic $\nabla L_{j,k}(\beta_{j \setminus k})$ as $h_{j,k}^{i,i'}(\beta_{j \setminus k})$ where subscript $i,i'$ indicates that $h_{j,k}^{i,i'}(\cdot)$ depends on $X_i$ and $X_{i'}$. We define $V_{j,k}^{i,i'}(\beta_{j \setminus k}) := h_{j,k}^{i,i'}(\beta_{j \setminus k}) h_{j,k}^{i,i'}(\beta_{j \setminus k})^T$. Then by definition, $\hat{\Sigma}_{j,k}(\beta_{j \setminus k})$ can be written as

$$\hat{\Sigma}_{j,k}(\beta_{j \setminus k}) = \frac{1}{n(n-1)^2} \sum_{i=1}^n \sum_{i' \neq i \neq i''} V_{j,k}^{i,i''}(\beta_{j \setminus k}).$$

57
Note that \( \hat{\Sigma}_{jk}(\hat{\beta}_{j\lor k}) - \Sigma_{jk} = \hat{\Sigma}_{jk}(\hat{\beta}_{j\lor k}) - \Sigma_{jk}(\beta^*_j\lor k) \). For any entry of \( \hat{V}^{jk}_{ii'i''}(\beta_{j\lor k}) \), since \( X_{ij} = \mathcal{O}_p(\log d) \), simple computation shows that with probability tending to 1, there exist a constant \( C_v \) such that
\[
\left\| \nabla_{\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d}} V^{jk}_{ii'i''}(\beta_{j\lor k}) \right\|_{\infty} \leq C_v \log^8 d \quad \text{for all } j, k \in [d], j \neq k \tag{C.34}
\]
and for any \((a, b), (c, d) \in \{(p, q) : p, q \in \{j, k\}, p \neq q\}\). Moreover, the constant \( C_v \) does not depend on \((j, k)\) or \(\beta_{j\lor k}\). Thus by the mean-value theorem, there exists a \( \hat{\beta} \) in the line segments between \( \beta_{j\lor k} \) and \( \beta^*_{j\lor k} \) such that
\[
[V^{jk}_{ii'i''}(\beta_{j\lor k})]_{ab,cd} - [V^{jk}_{ii'i''}(\beta^*_{j\lor k})]_{ab,cd} = \left\| \nabla_{\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d}} V^{jk}_{ii'i''}(\hat{\beta}_{j\lor k}) \right\|_{\infty} \left\| \hat{\beta}_{j\lor k} - \beta^*_{j\lor k} \right\|_1 = O_p(s^* \lambda \log^8 d),
\]
which yields that
\[
\|I_1\|_{\infty} = O_p(s^* \lambda \log^8 d) \quad \text{uniformly for all } (j, k). \tag{C.35}
\]

Now we turn to \( I_2 \). For notational simplicity, we use \( h_{ii'} \) and \( h_{ii'|i} \) to denote \( h^{jk}_{ij}(\beta^*_{j\lor k}) \) and \( h^{jk}_{ii'|i}(\beta^*_{j\lor k}) := \mathbb{E}\{h^{jk}_{ij}(\beta^*_{j\lor k}) | X_i\} \) respectively. As shown in §B.1, for \( i \neq i' \neq i'' \)
\[
\mathbb{E}(h_{ii'}h_{ii'}^T | X_i) = \mathbb{E}(h_{ii'|i}h_{ii'|i}^T) = \Sigma^{jk} \quad \text{and} \quad \mathbb{E}(h_{ij}h_{ij}^T) = \Theta^{jk},
\]
we can write \( I_2 \) as
\[
I_2 = \frac{n-2}{n-1} \left\{ \frac{n}{3} \sum_{i < i' < i''} \left[ V^{jk}_{ii'i''} - \mathbb{E}(V^{jk}_{ii'i''}) \right] \right\} + \frac{1}{n-1} \left\{ \binom{n}{2} \sum_{i < i'} \left[ V^{jk}_{ii'i'} - \mathbb{E}(V^{jk}_{ii'i'}) \right] \right\} + \frac{1}{n-1} \left( \Theta^{jk} - \Sigma^{jk} \right),
\]
where we use \( V^{jk}_{ii'i''} \) to denote \( V^{jk}_{ii'i''}(\beta^*_{j\lor k}) \). Observing that \( I_{21} \) is a centered third order U-statistic, for \( x \) large enough such that \( x^4 \geq \left\| \mathbb{E}\{V^{jk}_{i\lor j}(\beta^*_{j\lor k})\} \right\|_{\infty} \) and for any \((a, b), (c, d) \in \{(p, q) : p, q \in \{j, k\}\}\) we have
\[
\mathbb{P}\left( \left\| V^{jk}_{ii'i''}(\beta_{j\lor k}) \right\|_{ab,cd} > 2x^4 \right) \leq \mathbb{P}\left( (X_{ia} - X_{i'a})(X_{ib} - X_{i'b})(X_{ic} - X_{i'c})(X_{id} - X_{i'd}) > x^4 \right) \leq 8 \exp(2\kappa_m + \kappa_h) \exp(-x).
\]

Thus there exist constants \( c_1 \) and \( C_1 \) that does not depend on \( n \) or \( d \) or \((j, k)\) such that for any \( x \in \mathbb{R} \)
\[
\mathbb{P}(\left\| V^{jk}_{ii'i''}(\beta^*_{j\lor k}) \right\|_{ab,cd} > x) \leq C_1 \exp(c_1 x^{1/4}). \tag{C.36}
\]
By lemma A.1, with \( L_1 = C_1, \ L_2 = c_1, \ m = \lceil n/3 \rceil \) and \( q = 1/4 \) for any entry of \( I_{21}, \) (C.36) implies
\[
\mathbb{P}(\left\| I_{21} \right\|_{ab,cd} > x) \lesssim n^3 \exp(-2^{-1/4} c_1 m^{1/9} x^{2/9}).
\]

58
Taking a union bound over all the indices of $I_{21}$ we obtain
\[
\mathbb{P}(\|I_{21}\|_\infty > x) \lesssim \exp(-2^{-1/4}m^{1/9}x^{2/9} + 3 \log n + 2 \log d). \tag{C.37}
\]
Choosing $x = C\sqrt{\log^9 d/n}$ with a sufficiently large $C > 0$, by (C.37) we have
\[
\|I_{21}\|_\infty = \mathcal{O}_\mathbb{P}(\sqrt{\log^9 d/n}) \text{ uniformly for all } (j, k). \tag{C.38}
\]
For the second part $I_{22}$, note that it is a U-statistic of order 2, because (C.36) also holds for $\mathbb{V}_{ii'ii'}(\tilde{\beta}_{j\vee k})$. By first applying Lemma A.1 then taking a union bound, we can conclude that for all $(j, k)$, we have
\[
\|I_{21}\|_\infty = \mathcal{O}_\mathbb{P}(\sqrt{\log^9 d/n}). \tag{C.39}
\]
Combining (C.35)(C.38) and (C.39), we have the following error bound for $(\tilde{\Sigma}^j_{jk}(\tilde{\beta}_{j\vee k}))$:
\[
\|\tilde{\Sigma}^j_{jk}(\tilde{\beta}_{j\vee k}) - \Sigma^j_{jk}\|_\infty = \mathcal{O}_\mathbb{P}\left(s^*\lambda\log^8 d + \sqrt{\log^9 d/n}\right) \text{ for all } (j, k).
\]

C.2.3 Proof of Lemma B.3

Proof of Lemma B.3. By triangle inequality, $|\hat{S}_{jk} - \hat{S}^*_{jk}| \leq |\hat{S}_{jk} - \hat{S}_{jk}^*| + |\hat{S}_{jk}^* - \hat{S}^*_{jk}|$. Theorem 4.16 concludes that $|\hat{S}_{jk} - \hat{S}_{jk}^*| \lesssim (s^* + s^*)\lambda D$ for all $1 \leq j < k \leq d$; hence we only need to bound $|\hat{S}_{jk}^* - \hat{S}_{jk}^*|$. For notational simplicity, we denote $h_{ii'} = h_{ii'}^j(\beta_{j\vee k}^*)$ and $h_{ij|j} = h_{ij|j}^j(\beta_{j\vee k}^*)$ we have
\[
\hat{S}_{jk} - \hat{S}^*_{jk} = \frac{2}{n(n-1)} \sum_{i < j'} \Omega^T_{jk} \chi_{ii'} \quad \text{where } \chi_{ii'} = h_{ii'} - h_{ii'|j} - h_{ii'|j'}.
\]
$(\hat{S}_{jk}^* - \hat{S}^*_{jk})$ is a degenerate U-statistic of order 2 since $\mathbb{E}(\chi_{ii'} | X_i) = 0$. To obtain the tail probability of $\sqrt{n}(\hat{S}_{jk}^* - \hat{S}^*_{jk})$ we use the following concentration inequality for degenerate second-order U-statistics. This lemma is obtained from Houdre and Reynaud-Bouret (2003).

Lemma C.4. Let $T_1, \ldots, T_n$ be i.i.d. random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, $h(\cdot, \cdot): \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ satisfy (i) $h(x, y) = h(y, x)$ for all $(x, y)$, (ii) there exists $A > 0$ such that $|h(x, y)| \leq A$ for all $(x, y)$ and (iii) $\mathbb{E}[h(T_i, T_j) | T_i] = 0$. Let $U_n = \sum_{i<j} h(T_i, T_j)$, we define four constants $B, C, D$ and $E$ as
\[
B := \max \sqrt{n}\left\{\sup_t \mathbb{E}[h^2(T_1, T_2) | T_1 = t]\right\}^{1/2}, \quad C := \mathbb{E}\left(\sup_{j, t} \left\{\sum_{i=1}^{j-1} h(T_i, t)\right\}\right),
\]
\[
D := \sup \left\{\sum_{i<j} \mathbb{E}[h(T_i, T_j)a_i(T_i)b_j(T_j)] : \sum_{i=1}^{n-1} \mathbb{E}[a_i^2(T_i)] \leq 1, \sum_{i=1}^{n-1} \mathbb{E}[b_j^2(T_j)] \leq 1\right\},
\]
\[
E := \left\{\sum_{i<j} \mathbb{E}[h^2(T_i, T_j)]\right\}^{1/2}.
\]

Let $u > 0, \epsilon > 0$, we have the following inequality for the tail probability of the degenerate U-statistic $U_n$:

$$
\mathbb{P}[U_n \geq F_h(\epsilon, u)] \leq 3 \exp(-u). \tag{C.40}
$$

where $F(\epsilon, u)$ is a function of $\epsilon$ and $u$ which is given by

$$
F(\epsilon, u) = \frac{\kappa(\epsilon)}{3Au^2} + (\sqrt{2\kappa(\epsilon)} + 4/3)Bu^{3/2} + ((1 + \epsilon)/3C + 4D)u + (1 + \epsilon)E\sqrt{2u}
$$

with $\kappa(\epsilon) = 2.5 + 32/\epsilon$.

**Proof of Lemma C.4.** See the proof of Theorem 3.1 in Houdre and Reynaud-Bouret (2003). \qed

Let $\mathcal{E} := \{ \|X_i\|_\infty \leq \delta \log d, i = 1, \ldots, n \}$ be the event that the random variables $X_i$ are thresholded by $\delta \log d$, where $\delta$ is a positive number. On event $\mathcal{E}$, we also have $\mathbb{E}(h_{ij}) = 0$, thus $(S_{jk} - S_{jk}^*)$ is degenerate on event $\mathcal{E}$.

Moreover, by the definition of $h_{ii',v}$, on event $\mathcal{E}$, $\|X_{ii'}\|_\infty \leq C(\delta)\log^2 d$ where $C(\delta)$ is a constant depending $\delta$. Hölder's inequality implies that $|\Omega_{jk}X_{ii'}| \leq C(\delta)w_0\log^2 d$ for all $j < k$ and $i < i'$. If we denote $\Omega_{jk}X_{ii'}$ as $g_{jk}(X_i, X_{i'})$, then for any $y \in \mathbb{R}^d$, we have

$$
\mathbb{E}\{g_{jk}^2(X_i, X_{i'}) | \mathcal{E}, X_{i'} = y\} \leq C(\delta)^2w_0^2\log^2 d
$$

Moreover, for any functions $\{a_i(\cdot)\}_{i=1}^{n-1}$ and $\{b_{i'}(\cdot)\}_{i'=2}^{n}$ satisfying

$$
\sum_{i=1}^{n-1} \mathbb{E}\{a_i^2(X_i)\} \leq 1 \quad \text{and} \quad \sum_{i'=2}^{n} \mathbb{E}\{b_{i'}^2(X_j)\} \leq 1,
$$

taking conditional expectation implies

$$
\mathbb{E}_\mathcal{E}\{g_{jk}(X_i, X_{i'})a_i(X_i)b_{i'}(X_{i'})\} = \mathbb{E}_\mathcal{E}\{a_i(X_i)\}\mathbb{E}_\mathcal{E}\{g_{jk}(X_i, X_{i'})b_{i'}(X_{i'}) | X_i\},
$$

where we use $\mathbb{E}_\mathcal{E}$ to indicate that the expectation is taken conditioning on the event $\mathcal{E}$. By Cauchy-Schwarz inequality we have

$$
\mathbb{E}_\mathcal{E}\{g_{jk}(X_i, X_{i'})a_i(X_i)b_{i'}(X_{i'})\} \leq \mathbb{E}_\mathcal{E}^{1/2}\{a_i^2(X_i)\}\mathbb{E}_\mathcal{E}^{1/2}\{\mathbb{E}_\mathcal{E}\{g_{jk}^2(X_i, X_{i'})b_{i'}^2(X_{i'}) | X_i\}\}. \tag{C.41}
$$

By using Cauchy-Schwarz inequality again we have

$$
\mathbb{E}_\mathcal{E}^2\{g_{jk}(X_i, X_{i'})b_{i'}(X_{i'}) | X_i\} \leq \mathbb{E}_\mathcal{E}\{g_{jk}^2(X_i, X_{i'}) | X_i\}\mathbb{E}_\mathcal{E}\{b_{i'}^2(X_{i'})\}. \tag{C.42}
$$

By the law of total expectation,

$$
\mathbb{E}_\mathcal{E}\{\mathbb{E}_\mathcal{E}\{g_{jk}^2(X_i, X_{i'}) | X_i\}\} = \mathbb{E}_\mathcal{E}\{g_{jk}^2(X_i, X_{i'})\} \leq \mathbb{E}\{\Omega_{jk}^T X_{ii'} X_{ii'}^T \Omega_{jk}\} \leq 3w_0^2D. \tag{C.43}
$$

60
Combining (C.41), and (C.43) we have
\[
\mathbb{E}_* \left\{ g_{jk}(X_i, X_{i'}) a_i(X_i) b_{i'}(X_{i'}) \right\} \leq 3w_0^2 D \mathbb{E}_*^{1/2} \left\{ a_i^2(X_i) \right\} \mathbb{E}_*^{1/2} \left\{ b_{i'}^2(X_{i'}) \right\}.
\]
Hence we obtain
\[
\sum_{i < i'} \mathbb{E}_* \left\{ g_{jk}(X_i, X_{i'}) a_i(X_i) b_{i'}(X_{i'}) \right\} \leq 3w_0^2 D \sum_{i < i'} \left( \mathbb{E}_* \left\{ a_i^2(X_i) \right\} + \mathbb{E}_*^{1/2} \left\{ b_{i'}^2(X_{i'}) \right\} \right) \leq 1.5w_0^2 D n.
\]
Moreover, it holds that \( \sum_{i < i'} \mathbb{E} \left\{ g_{jk}^2(X_i, X_{i'}) \right\} = \sum_{i < i'} \mathbb{E} \left\{ \Omega_{jk}^T X_{ii'} X_{ii'}^T \Omega_{jk} \right\} \leq 1.5w_0^2 D n(n - 1). \)

By choosing \( H \) hence we obtain
\[
\text{Lemma C.4}\]
\[
\text{Proof of Lemma B.4.}
\]
\[
\text{C.2.4 Proof of Lemma B.4}
\]
\[
\text{Given a sequence of i.i.d. random variables } \xi_1, \xi_2, \ldots, \xi_n \sim N(0, 1), \text{ beside } \hat{S}_{jk}^* \text{ defined in (B.19), we write the Bootstrap counterpart of } \hat{S}_{jk} \text{ and } S_{jk}^* \text{ as }
\]
\[
\hat{S}_{jk}^* := \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{1}{n - 1} \sum_{i' \neq i} \hat{\Omega}_{jk}^{T} h_{ii'}(\hat{\beta}_{jvk}) (\xi_i + \xi_{i'}) \right\} \xi_i
\]
and
\[
S_{jk}^* := \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{1}{n - 1} \sum_{i' \neq i} \Omega_{jk}^{T} h_{ii'}(\beta_{jvk}) (\xi_i + \xi_{i'}) \right\} \xi_i.
\]
where $\Omega_{jk}$ and $\hat{\Omega}_{jk}$ are defined in (B.17) and we use $\hat{\beta}^T_{j\vee k}$ to denote $(0, \hat{\beta}^T_{j\vee k}, \hat{\beta}^T_{k\vee j})^T$. Then by definition $W = \sqrt{n} \max_{E^c} |\hat{S}_{jk}^B|/2$ and $W_0 = \sqrt{n} \max_{E^c} |\hat{S}_{jk}^*|/2$. The following maximal inequality from Chernozhukov et al. (2013) enable us to bound $\max_{E^c} |\hat{S}_{jk}^B - \hat{S}_{jk}^*|$.

**Lemma C.5 (Maximal inequality).** Let $Z_1, \ldots, Z_n$ be independent random vectors in $\mathbb{R}^d$ with $d \geq 2$. Define $M := \max_{i\in[n]} \max_{j\in[d]} |Z_{ij}|$ and $\sigma^2 := \max_{j\in[d]} \frac{1}{n} \sum_{i=1}^n E[Z_{ij}^2]$. Then

$$
\mathbb{E}\left\{ \max_{1 \leq j \leq p} \left| \frac{1}{n} \sum_{i=1}^n (Z_{ij} - \mathbb{E}[Z_{ij}]) \right| \right\} \lesssim \sigma \sqrt{\log d/n} + \sqrt{\mathbb{E}[M^2]} \log d/n.
$$

**Proof of Lemma C.5.** See the proof of Lemma A.1 in Chernozhukov et al. (2013) for the details. □

As a particular case of Lemma C.5, we consider the case where $Z_{ij} = W_{ij} \xi_i$ where $\{W_{ij}\}_{i\in[n], j\in[d]}$ are independent of the $n$ i.i.d. $N(0, 1)$ random variable $\xi_1, \ldots, \xi_n$. Then $M = \max_{i\in[n]} \max_{j\in[d]} |W_{ij}| \leq \max_{i\in[n], j\in[d]} |W_{ij}| \sqrt{\log n}$ with probability tending to one. By Lemma C.5, we have the following error bound with high probability:

$$
\mathbb{E}_\xi \left\{ \max_{E^c} |\hat{S}_{jk}^B - \hat{S}_{jk}^*| \right\} \lesssim \max_{E^c} \max_{i\in[n]} \Phi_{ij}^i (\log n)^{1/2} n^{-2} \log |E^c| + \max_{E^c} \Psi_{jk} n^{-2} \sqrt{\log |E^c|},
$$

where $\mathbb{E}_\xi$ means the expectation is taken with respect to $\xi_1, \ldots, \xi_n$, $\Phi_{jk}^i$ and $\Psi_{jk}$ are defined as

$$
\Phi_{jk}^i := \left\{ \sum_{i' \neq i} \left( \Omega_{jk}^{T} h_{i'i'}(\hat{\beta}_{j\wedge k}^T) - \Omega_{jk}^{T} h_{i'i'}(\hat{\beta}_{j\vee k}^T) \right) \right\},
$$

and

$$
\Psi_{jk} = \sum_{i=1}^n \left\{ \sum_{i' \neq i} \left( \Omega_{jk}^{T} h_{i'i'}(\hat{\beta}_{j\vee k}^T) - \Omega_{jk}^{T} h_{i'i'}(\hat{\beta}_{j\wedge k}^T) \right) \right\}^2.
$$

In what follows, we derive bounds for $\Phi_{jk}^i$ and $\Psi_{jk}^2$ respectively. For $\Phi_{jk}^i$, triangle inequality implies that

$$
\Phi_{jk}^i \leq \sum_{i' \neq i} \left( \Omega_{jk}^{T} h_{i'i'}(\hat{\beta}_{j\vee k}^T) - \Omega_{jk}^{T} h_{i'i'}(\hat{\beta}_{j\wedge k}^T) \right) \left| \sum_{i' \neq i} \left( \Omega_{jk}^{T} h_{i'i'}(\hat{\beta}_{j\vee k}^T) - \Omega_{jk}^{T} h_{i'i'}(\hat{\beta}_{j\wedge k}^T) \right) \right| + \sum_{i' \neq i} \left( \Omega_{jk} - \Omega_{jk} \right)^T h_{i'i'}(\hat{\beta}_{j\vee k}^T) \left| \sum_{i' \neq i} \left( \Omega_{jk} - \Omega_{jk} \right)^T h_{i'i'}(\hat{\beta}_{j\wedge k}^T) \right|.
$$

For term $I_1$, since there exist a constant $C > 0$ such that for all $\beta_{j\vee k} \in \mathbb{R}^{2d-3}$, $\|\nabla h_{i'i'}(\hat{\beta}_{j\vee k}^T)\|_\infty \leq C \log^4 d$ with probability tending to one, Hölder’s inequality and Mean-Value Theorem imply that there exists a $\beta_{j\vee k} \in \mathbb{R}^{2d-3}$ on the line segment between $\hat{\beta}_{j\vee k}$ and $\beta_{j\vee k}^*$ such that

$$
I_1 \leq \sum_{i' \neq i} \left| \sum_{i' \neq i} \left( \nabla h_{i'i'}(\beta_{j\vee k}^T) \right) \left| \sum_{i' \neq i} \left( \nabla h_{i'i'}(\hat{\beta}_{j\vee k}^T) \right) \right| \right| \lesssim w_0 s^* n \lambda \log^4 d,
$$

(C.44)
where we use the optimality of $\hat{\Omega}_{jk}$ and the estimation error rate of $\hat{\beta}_{jk}$. Inequality (C.44) holds for all $1 \leq j < k \leq d$ and $1 \leq i < i' \leq n$. For the second part $I_2$, Hölder’s inequality implies

$$I_2 \leq \sum_{i' \neq i} \left\| \hat{\Omega}_{jk} - \Omega_{jk} \right\|_1 \left\| h_{ii'}^j(\beta_{jk}^\ast) \right\|_\infty \lesssim s^* n \lambda_D \log^2 d. \quad (C.45)$$

For the last part $I_3$, Hölder’s inequality implies

$$I_3 \leq \sum_{i' \neq i} \left\| \Omega_{jk} \right\|_1 \left\| h_{ii'}^j(\beta_{jk}^\ast) - h_{ii'}^j(\beta_{jk}^\ast) \right\|_\infty \lesssim n w_0 \log^2 d. \quad (C.46)$$

Combining (C.44), (C.45) and (C.46) we can conclude that for

$$
\max_{E^c} \max_{i \in [n]} \Phi_{jk}(\log n)^{1/2} n^{-1/2} \log |E^c| \lesssim \left( w_0 + w_0 s^* \lambda \log^2 d + s^* \lambda_D \right) (\log n)^{1/2} \log^3 d/n \lesssim \left( w_0 + s^* \lambda_D \right) (\log n)^{1/2} \log^3 d/n \quad (C.47)
$$

with probability tending to one. The last inequality follows from $\lim_{n \to \infty} s^* \lambda \log^2 d = 0$.

For term $\Psi_{jk}$, by inequality $(a + b + c)^2 \leq 3(a^2 + b^2 + c^2)$ we have $\Psi_{jk}^2 \leq 3(J_1 + J_2 + J_3)$ where $J_1$ and $J_2$ and $J_3$ are given by

$$J_1 = \sum_{i=1}^n \left\{ \sum_{i' \neq i} \hat{\Omega}_{jk}^T \left\{ h_{ii'}^j(\beta_{jk}^\ast) - h_{ii'}^j(\beta_{jk}^\ast) \right\} \right\}^2; J_2 = \sum_{i=1}^n \left\{ \sum_{i' \neq i} \left\{ \hat{\Omega}_{jk} - \Omega_{jk} \right\}^T h_{ii'}^j(\beta_{jk}^\ast) \right\}^2 \quad \text{and}
$$

$$J_3 = \sum_{i=1}^n \left\{ \sum_{i' \neq i} \Omega_{jk}^T \left\{ h_{ii'}^j(\beta_{jk}^\ast) - h_{ii'}^j(\beta_{jk}^\ast) \right\} \right\}^2.
$$

From the definition of $h_{ii'}^j(\beta_{jk}^\ast)$, we know that for any $\beta_{jk} \in \mathbb{R}^{2d-3}$, any $1 \leq i < i' \leq n$, $1 \leq j < k \leq d$, we have $\left\| \nabla h_{ii'}^j(\beta) \right\|_\infty \lesssim \log^2 d$ with probability tending to one. By Hölder’s inequality and Mean-Value Theorem, there exist a $\beta_{jk} \in \mathbb{R}^{2d-3}$ such that with high probability

$$\left| \hat{\Omega}_{jk}^T \left\{ h_{ii'}^j(\beta_{jk}^\ast) - h_{ii'}^j(\beta_{jk}^\ast) \right\} \right| \leq \left\| \hat{\Omega}_{jk} \right\|_1 \left\| \nabla h_{ii'}^j(\beta_{jk}^\ast) \right\|_\infty \left\| \beta_{jk} - \beta_{jk}^\ast \right\|_\infty \lesssim s^* \lambda w_0 \log^4 d.
$$

Therefore we can conclude that $J_1 \lesssim s^* n^3 \lambda^2 \log^2 d$ with high probability. For the second term $J_2$, by the definition of $\hat{\Sigma}(\beta)$ in (4.19), we can see that $J_2$ can be written as

$$J_2 = n(n-1)^2 \left( \hat{\Omega}_{jk} - \Omega_{jk} \right)^T \hat{\Sigma}^j(\beta^\ast)(\hat{\Omega}_{jk} - \Omega_{jk}). \quad (C.48)
$$

Using Hölder’s inequality twice and Lemma B.2, the right hand side of (C.48) can be bounded by

$$J_2 \leq n(n-1)^2 \left\| \hat{\Omega}_{jk} - \Omega_{jk} \right\|_1^2 \left\| \hat{\Sigma}^j(\beta^\ast) \right\|_\infty \lesssim s^* n^3 \lambda^2 D + \sqrt{\log^9 d/n}
$$

with probability tending to one. Finally, for the last term $J_3$, we can write $J_3$ as

$$J_3 = \sum_{i \in [n], i' \neq i} \Omega_{jk}^T N_{ii'i'} \Omega_{jk}, \quad (C.49)$$

63
where \( N_{ii'j}^{jk} = \{ h_{ii'}^{jk}(\beta^*) - h_{ii'}^{jk}(\beta^*) \} \{ h_{ii'}^{jk}(\beta^*) - h_{ii'}^{jk}(\beta^*) \}^T \). For notational simplicity, we use \( h_{ii'}, h_{ii'}^{jk} \) and \( N_{ii'j}^{jk} \) to denote \( h_{ii'}^{jk}(\beta^*) \), \( h_{ii'}^{jk}(\beta^*) \) and \( N_{ii'j}^{jk} \) respectively, then by definition of \( N_{ii'j}^{jk} \) we have for \( i \neq i' \neq i'' \)

\[
\mathbb{E}(N_{ii'j}^{jk}) = \mathbb{E}(h_{ii'}^{jk}h_{ii'}^{jk}) - \mathbb{E}(h_{ii'}^{jk}h_{ii'}^{jk}|i''(\beta^*)} = \mathbb{E}(h_{ii'}^{jk}h_{ii'}^{jk}|i(\beta^*)}^T = 0.
\]

Moreover, simple computation yields \( \mathbb{E}(N_{ii'j}^{jk}) = \Theta^{st} - \Sigma^{jk} \) where \( \Theta^{jk} = \mathbb{E}(h_{ii'}^{jk}(\beta^*)h_{ii'}^{jk}(\beta^*)^T) \). Similar to the proof of Lemma B.2, we can write \( J_3 \) as a sum of a third order U-statistic and a second order U-statistic:

\[
J_3 = 6 \sum_{i < i' < i''} \mathbb{E} \left[ \sum_{j<k} \Omega_{jk}^T N_{ii'j} \Omega_{jk} \right] + 2 \sum_{i<i'} \mathbb{E} \left[ \sum_{j<k} \Omega_{jk}^T N_{ii'j} \Omega_{jk} \right] \tag{C.50}
\]

Note that by Assumption 4.1, for any \( j, k \in [d] \) we have

\[
\mathbb{E} \left[ \frac{R_{ii'}^{jk}(\beta^*)}{1 + R_{ii'}^{jk}(\beta^*)} (X_{ij} - X_{i''j})(X_{ik} - X_{i''k}) \right] \leq \mathbb{E} \left[ |X_{ij} - X_{i''j}| \cdot |X_{ik} - X_{i''k}| \right] \leq |X_{ij} - X_{ik}| + \kappa |X_{ij}| + \kappa |X_{ik}| + \kappa v,
\]

we conclude that there exist two constants \( A_1 \) and \( A_2 \) such that for any entry \( [N_{ii'j}]_{ab,cd} \) of \( N_{ii'j} \),

\[
\mathbb{P} \left( [N_{ii'j}]_{ab,cd} > x \right) \leq A_1 \exp \left( A_2 x^{-1/4} \right).
\]

We denote \( J_1 := 6 \binom{n}{3}^{-1} \sum_{ii'j} N_{ii'j} \) and \( J_2 := 2 \binom{n}{2}^{-1} \sum_{ii'j} N_{ii'j} \), by (C.50) we have \( J_{31} = \binom{n}{3} \Omega_{jk}^T J_1 \Omega_{jk} \) and \( J_{32} = \binom{n}{2} \Omega_{jk}^T J_2 \Omega_{jk} \). Moreover, the expectation of \( J_1 \) and \( J_2 \) are given by \( \mathbb{E} \{ J_1 \} = 0 \) and \( \mathbb{E} \{ J_2 \} = 2 \Theta^{jk} - 2 \Sigma^{jk} \) respectively. Applying Lemma A.1 we obtain

\[
\mathbb{P} \left( [J_1]_{ab,cd} > x \right) \lesssim n^3 \exp (-C n^{1/6} x^{2/9}) \quad \text{and} \quad \mathbb{P} \left( [J_2 - \mathbb{E} J_2]_{ab,cd} > x \right) \lesssim n^3 \exp (-C n^{1/6} x^{2/9}).
\]

Taking a union bound we obtain that

\[
\| J_1 \|_\infty \leq C_1 \sqrt{\log d/n} \quad \text{and} \quad \| J_2 - \mathbb{E} J_2 \|_\infty \leq C_2 \sqrt{\log d/n} \tag{C.51}
\]

with probability tending to one, where \( C_1 \) and \( C_2 \) are generic constants. By Hölder’s inequality and (C.51) we obtain

\[
J_{31} \leq n^3 \| \Omega_{jk} \|_1 \| J_1 \|_\infty \leq C_1 n^3 \omega_0 \sqrt{\log d/n} \quad \text{and} \quad J_{32} \leq n^2 \| \Omega_{jk} \|_1 \| J_2 \|_\infty \leq n^2 \omega_0 (C_2 \sqrt{\log d/n} + 4D)
\]

with probability tending to one. Note that we use the condition that \( \| \Sigma^{jk} \|_\infty \leq D \) and \( \| \Theta^{jk} \|_\infty \leq D \) in the inequality above. Therefore we obtain that

\[
J_3 \leq \omega_0^2 n^3 \sqrt{\log d/n}.
\]
Therefore combining the bonds on \( J_1, J_2 \) and \( J_3 \) we have

\[
\Psi_{jk}^2 \lesssim n^3 \left( s^2 \lambda^2 w_0^2 \log^8 d + s^2 \lambda^2 D + w_0^2 \sqrt{\log^9 d/n} \right).
\]

Recall that \( \lambda \asymp (\log^n d/n)^{1/2} \) and \( \lambda_D \gtrsim w_0 s^2 \lambda \log^2 d \), Assumptions 4.13 and 4.18 implies that \( \Psi_{jk}^2 \lesssim n^3 (\log^n d/n)^{1/2} \). Therefore we have

\[
\max_{E^c} \Psi_{jk} n^{-2} \sqrt{\log |E^c|} \lesssim n^{-3/4} \log^{11/4} d.
\]

Thus by (C.47) and (C.52) we obtain

\[
E \{ \max_{E^c} \sqrt{n} | \hat{S}_{jk}^B - \hat{S}_{jk}^B \{ \} \} \lesssim (\log^{11} d/n)^{1/4} \log^{4} d \log^{1/2} n/n \lesssim (\log^{11} d/n)^{1/4},
\]

where the last inequality follows from \( \log^4 d \sqrt{\log n/n} \) is relatively negligible.

\[\Box\]

### C.3 Proof of Technical Lemmas

Finally, we prove the technical lemmas. These Lemmas are standard for high-dimensional linear regression, but proving them for our logistic-type loss function needs nontrivial extensions.

#### C.3.1 Proof of Lemma C.1

**Proof of Lemma C.1.** Let \( I \) and \( J \) be two index sets with \( I \cap J = \emptyset \), \( |I| \leq s, |J| \leq k \), for any \( \mathbf{u} \in \mathbb{R}^d \) with \( ||\mathbf{u} - \mathbf{u}_0||_2 \leq \rho \) and any \( \mathbf{v}, \mathbf{w} \in \mathbb{R}^d \), let \( \theta = \mathbf{v}_I + \alpha \mathbf{w}_J \) with some \( \alpha \in \mathbb{R} \), then by definition, \( ||\theta||_0 \leq s + k \). For notational simplicity, we denote \( s \)-sparse eigenvalues \( \rho_+(M(\cdot), \mathbf{u}_0; s, r) \) and \( \rho_-(M(\cdot), \mathbf{u}_0; s, r) \) as \( \rho_+(s) \) and \( \rho_-(s) \) respectively. By definition, we have

\[
\rho-(s+k)||\theta||_2^2 \leq \theta^T M(\mathbf{u}) \theta = \mathbf{v}_I^T M(\mathbf{u}) \mathbf{v}_I + 2\alpha \mathbf{v}_I^T M(\mathbf{u}) \mathbf{w}_J + \alpha^2 \mathbf{w}_J^T M(\mathbf{u}) \mathbf{w}_J.
\]

Since \( ||\theta||_2^2 = ||\mathbf{v}_I||_2^2 + \alpha^2 ||\mathbf{w}_J||_2^2 \). Rearranging the terms in (C.53) we have

\[
(A_3 - \rho-(s+k)||\mathbf{w}_J||_2^2)\alpha^2 + 2A_2\alpha + (A_1 - \rho-(s+k)||\mathbf{v}_I||_2^2) \geq 0 \quad \text{for all } \alpha \in \mathbb{R}.
\]

Note that the left-hand side (C.54) is a univariate quadratic function in \( \alpha \), thus (C.54) implies that

\[
(A_1 - \rho-(s+k)||\mathbf{v}_I||_2^2) \geq A_2^2.
\]

Therefore by multiplying \( 4||\mathbf{v}_I||_2^2/(A_2^2||\mathbf{w}_J||_2^2) \) on to both sides of (C.55) we have

\[
\frac{4A_2^2||\mathbf{v}_I||_2^2}{A_1^2||\mathbf{w}_J||_2^2} \leq \frac{A_1||\mathbf{w}_J||_2^2}{A_1||\mathbf{v}_I||_2^2} \left( \frac{A_1 - \rho-(s+k)||\mathbf{v}_I||_2^2}{A_1} \right)
\]

\[
\frac{A_3 - \rho-(s+k)||\mathbf{w}_J||_2^2}{A_2^2}.
\]
By the inequality of arithmetic and geometric means, we have
\[
\frac{\rho_-(s+k)\|v_I\|_2^2}{A_1} \left( \frac{A_1 - \rho_-(s+k)\|v_I\|_2^2}{A_1} \right) \leq \frac{1}{4}.
\]
Then the right-hand side of (C.55) can be bounded by
\[
\frac{4A_2^2\|v_I\|_2^2}{A_1^2\|w_J\|_2^2} \leq \frac{A_3 - \rho_-(s+k)\|w_J\|_2^2}{\rho_-(s+k)\|w_J\|_2^2} \leq \frac{\rho_+(k)}{\rho_-(s+k)} - 1,
\]
where the last inequality follows from \(A_3 \leq \rho_+(k)\|w_J\|_2^2\). Note that by the relationship between \(\ell_2\) and \(\ell_\infty\) norm, we have \(\|w_J\|_2 \leq \sqrt{K}\|w_J\|_\infty\), which further implies that
\[
\frac{v_J^T M(u)w_J\|v_I\|_2^2}{v_J^T M(u)v_I\|w_J\|_\infty^2} \leq \sqrt{K}\frac{\|v_J^T M(u)w_J\|_2^2}{\|v_J^T M(u)v_I\|_2^2} \leq \sqrt{A_2}\frac{\|v_I\|_2^2}{\|w_J\|_2^2} \leq \frac{\sqrt{K}}{2}\sqrt{\rho_+(k)/\rho_-(s+k) - 1}.
\]
Taking a supremum over \(v, w \in \mathbb{R}^d\) finally yields Lemma C.1. \(\square\)

### C.3.2 Proof of Lemma C.2

**Proof of Lemma C.2.** For \(v = (v_1, \ldots, v_d)^T \in \mathbb{R}^d\), without loss of generality, we assume that \(F^c = [s_1]\) where \(s_1 = |F^c| \leq s\). 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 \(v\) such that \(|v_j| \geq |v_{j+1}|\) for all \(j\) greater than \(s_1\). 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 \(\|v_J\|_\infty \leq \|v_{J_{i-1}}\|_1/k\) when \(i \geq 2\) because by the definition of \(J_i\) we have \(\sum_{i \geq 2} \|v_{J_i}\|_\infty \leq \|v_F\|_1/k\). Note that by the definition of index sets \(I\) and \(J_i\), \(|J_i| \leq k\) and \(|I| = k + s_1 \leq k + s\). We denote the restricted correlation coefficients \(\pi(M, u_0; s, k)\) as \(\pi(s, k)\), then by the definition of \(\pi(s+k, k)\) we have
\[
|v_J^T M(u)v_J| \leq \pi(s+k, k)(v_J^T M(u)v_J)\|v_J\|_\infty/\|v_I\|_2.
\]
Thus we have the following upper bound for \(|v_I^T M(u)v_F|\):
\[
|v_I^T M(u)v_F| \leq \sum_{i \geq 2} |v_I^T M(u)v_{J_i}| \leq \pi(s+k, k)\|v_I\|_2^{-1} (v_I^T M(u)v_I) \sum_{i \geq 2} \|v_{J_i}\|_\infty
\]
\[
\leq \pi(s+k, k)\|v_I\|_2^{-1} (v_I^T M(u)v_I) \|v_F\|_1/k.
\] (C.57)

Because \(v^T M(u)v \geq v_I^T M(u)v_I + 2v_I^T M(u)v_F\), by (C.57) we have
\[
v^T M(u)v \geq v_I^T M(u)v_I - 2\pi(s+k, k)\|v_I\|_2^{-1} (v_I^T M(u)v_I) \|v_F\|_1/k
\]
\[
= (v_I^T M(u)v_I) \left( 1 - 2\pi(s+k, k)\|v_I\|_2^{-1} \|v_F\|_1/k \right).
\]
Thus we can bound the right-hand side of the last formula using the sparse eigenvalue condition
\[
v^T M(u)v \geq \rho_-(s+k) \left( 1 - 2\pi(s+k, k)\|v_I\|_2^{-1} \|v_F\|_1 \right) \|v_I\|_2^2.
\]
This concludes the proof of Lemma C.2. \(\square\)
C.3.3 Proof of Lemma C.3

Proof of Lemma C.3. Let $F(t) = L_j(\beta(t)) - L_j(\beta_1) - \langle \nabla L_j(\beta_1), \beta(t) - \beta_1 \rangle$, because the derivative of $L_j(\beta(t))$ with respect to $t$ is $\langle \nabla L_j(\beta(t)), \beta_2 - \beta_1 \rangle$ then the derivative of $F$ is

$$F'(t) = \langle \nabla L_j(\beta(t)) - \nabla L_j(\beta_1), \beta_2 - \beta_1 \rangle.$$ 

Therefore the Bregman divergence $D_j(\beta(t), \beta_1)$ can be written as

$$D_j(\beta(t), \beta_1) = \langle \nabla L_j(\beta(t)) - \nabla L_j(\beta_1), t(\beta_2 - \beta_1) \rangle = tF'(t).$$

By definition, it is easy to see that $F'(1) = D_j(\beta_2, \beta_1)$. To derive Lemma C.3, it suffices to show that $F(t)$ is convex, which implies that $F'(t)$ is non-decreasing and $D_j(\beta(t), \beta_1) = tF'(t) \leq tF'(1) = tD_j(\beta_2, \beta_1)$.

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

$$\langle \nabla L_j(\beta_1), \beta(t_1 x + t_2 y) - \beta_1 \rangle = t_1 \langle \nabla L_j(\beta_1), \beta(x) - \beta_1 \rangle + t_2 \langle \nabla L_j(\beta_1), \beta(y) - \beta_1 \rangle. \quad (C.58)$$

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

$$L_j(\beta(t_1 x + t_2 y)) \leq t_1 L_j(\beta(x)) + t_2 L_j(\beta(y)). \quad (C.59)$$

Adding (C.58) and (C.59) we obtain

$$F(t_1 x + t_2 y) \leq t_1 F(x) + t_2 F(y).$$

Therefore $F(t)$ is convex, thus we have $D_j(\beta(t), \beta_1) \leq t D_j(\beta_2, \beta_1)$.

References

Allen, G. I. and Liu, Z. (2012). A log-linear graphical model for inferring genetic networks from high-throughput sequencing data. In Proceedings of the 2012 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE Computer Society.

Anandkumar, A., Tan, V. Y., Huang, F., Willsky, A. S. et al. (2012). High-dimensional structure estimation in Ising models: Local separation criterion. The Annals of Statistics 40 1346–1375.

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

Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica 80 2369–2429.
Belloni, A., Chernozhukov, V. and Wei, Y. (2013). Honest confidence regions for a regression parameter in logistic regression with a large number of controls. arXiv preprint arXiv:1304.3969.

Bickel, P. J., Levina, E. et al. (2008). Covariance regularization by thresholding. The Annals of Statistics 36 2577–2604.

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

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

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. The Annals of Applied Statistics 5 232–253.

Cai, T., Liu, W. and Luo, X. (2011). A constrained ℓ1 minimization approach to sparse precision matrix estimation. Journal of the American Statistical Association 106 594–607.

Candés, E. J. and Tao, T. (2005). Decoding by linear programming. IEEE Transactions on Information Theory 51 4203–4215.

Chen, S., Witten, D. and Shojaie, A. (2013). Selection and estimation for mixed graphical models. arXiv preprint arXiv:1311.0085.

Cheng, J., Levina, E., Wang, P. and Zhu, J. (2012). Sparse Ising models with covariates. arXiv preprint arXiv:1209.6342.

Cheng, J., Levina, E. and Zhu, J. (2013). High-dimensional mixed graphical models. arXiv preprint arXiv:1304.2810.

Chernozhukov, V., Chetverikov, D., Kato, K. et al. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. The Annals of Statistics 41 2786–2819.

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

DasGupta, A. (2008). Asymptotic Theory of Statistics and Probability. Springer.

Dempster, A. P. (1972). Covariance selection. Biometrics 157–175.

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

Drton, M., Perlman, M. D. et al. (2007). Multiple testing and error control in Gaussian graphical model selection. Statistical Science 22 430–449.
Edwards, D. (2000). *Introduction to Graphical Modelling*. Springer.

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

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

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

Fellinghauer, B., Bühlmann, P., Ryffel, M., Von Rhein, M. and Reinhardt, J. D. (2013). Stable graphical model estimation with random forests for discrete, continuous, and mixed variables. *Computational Statistics & Data Analysis* **64** 132–152.

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

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Applications of the Lasso and grouped Lasso to the estimation of sparse graphical models. Tech. rep.

Guo, J., Levina, E., Michailidis, G. and Zhu, J. (2010). Joint structure estimation for categorical markov networks.

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

Guo, J., Levina, E., Michailidis, G. and Zhu, J. (2014). Graphical models for ordinal data. *Journal of Computational and Graphical Statistics* To appear.

Höfling, H. and Tibshirani, R. (2009). Estimation of sparse binary pairwise markov networks using pseudo-likelihoods. *The Journal of Machine Learning Research* **10** 883–906.

Houdre, C. and Reynaud-Bouret, P. (2003). Exponential inequalities, with constants, for u-statistics of order two. *Stochastic Inequalities and Applications* **56** 5569.

Jankova, J. and van de Geer, S. (2014). Confidence intervals for high-dimensional inverse covariance estimation. *arXiv preprint arXiv:1403.6752*.

Javanmard, A. and Montanari, A. (2013). Confidence intervals and hypothesis testing for high-dimensional regression. *arXiv preprint arXiv:1306.3171*.

Johnson, C. C., Jalali, A. and Ravikumar, P. D. (2012). High-dimensional sparse inverse covariance estimation using greedy methods. In *International Conference on Artificial Intelligence and Statistics*.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT press.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of statistics* **37** 4254.
Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press.

Lee, J. D. and Hastie, T. J. (2014). Learning the structure of mixed graphical models. *Journal of Computational and Graphical Statistics* To appear.

Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2013). Exact inference after model selection via the Lasso. *arXiv preprint arXiv:1311.6238*.

Lee, S.-I., Ganapathi, V. and Koller, D. (2006). Efficient structure learning of markov networks using $\ell_1$-regularization. *Advances in neural Information processing systems* 817–824.

Liu, H., Han, F., Yuan, M., Lafferty, J., Wasserman, L. et al. (2012). High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics* 40 2293–2326.

Liu, H., Lafferty, J. and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *The Journal of Machine Learning Research* 10 2295–2328.

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

Lockhart, R., Taylor, J., Tibshirani, R. J., Tibshirani, R. et al. (2014). A significance test for the Lasso. *The Annals of Statistics* 42 413–468.

Loh, P.-L. and Wainwright, M. J. (2013). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *arXiv preprint arXiv:1305.2436*.

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

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

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

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

Mohan, K., London, P., Fazel, M., Witten, D. and Lee, S.-I. (2014). Node-based learning of multiple Gaussian graphical models. *The Journal of Machine Learning Research* 15 445–488.

Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science* 27 538–557.

Ning, Y. and Liu, H. (2014). Sparc: Optimal estimation and asymptotic inference under semiparametric sparsity. *ArXiv e-prints*.

Peng, J., Wang, P., Zhou, N. and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association* 104 735–746.
Raskutti, G., Wainwright, M. J. and Yu, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *The Journal of Machine Learning Research* **11** 2241–2259.

Ravikumar, P., Wainwright, M. J., Lafferty, J. D. et al. (2010). High-dimensional Ising model selection using $\ell_1$-regularized logistic regression. *The Annals of Statistics* **38** 1287–1319.

Ravikumar, P., Wainwright, M. J., Raskutti, G., Yu, B. et al. (2011). High-dimensional covariance estimation by minimizing $\ell_1$-penalized log-determinant divergence. *Electronic Journal of Statistics* **5** 935–980.

Ren, Z., Sun, T., Zhang, C.-H. and Zhou, H. H. (2013). Asymptotic normality and optimalities in estimation of large Gaussian graphical model. *Annals of Statistics* To appear.

Rothman, A. J., Bickel, P. J., Levina, E. and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2** 494–515.

Schmidt, M., Niculescu-Mizil, A. and Murphy, K. (2007). Learning graphical model structure using $\ell_1$-regularization paths. In *Proceedings of the 22nd national conference on Artificial intelligence-Volume 2*. AAAI Press.

Shah, R. D. and Samworth, R. J. (2013). Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75** 55–80.

Shen, X., Pan, W. and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association* **107** 223–232.

Sun, T. and Zhang, C.-H. (2013). Sparse matrix inversion with scaled Lasso. *The Journal of Machine Learning Research* **14** 3385–3418.

Tan, K. M., London, P., Mohan, K., Lee, S.-I., Fazel, M. and Witten, D. (2014). Learning Graphical Models With Hubs. *ArXiv e-prints*.

Taylor, J., Lockhart, R., Tibshirani, R. J. and Tibshirani, R. (2014). Post-selection adaptive inference for least angle regression and the Lasso. *arXiv preprint arXiv:1401.3889*.

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

Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J. and Vlahavas, I. (2011). Mulan: A java library for multi-label learning. *The Journal of Machine Learning Research* **12** 2411–2414.

Turnbull, D., Barrington, L., Torres, D. and Lanckriet, G. (2008). Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio, Speech, and Language Processing* **16** 467–476.

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

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

71
Van de Geer, S. A. (2008). High-dimensional generalized linear models and the Lasso. *The Annals of Statistics* **36** 614–645.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*, vol. 3. Cambridge university press.

Voorman, A., Shojaie, A. and Witten, D. (2014a). Graph estimation with joint additive models. *Biometrika* **101** 85–101.

Voorman, A., Shojaie, A. and Witten, D. (2014b). Inference in high dimensions with the penalized score test. *arXiv preprint arXiv:1401.2678*.

Wang, Z., Liu, H. and Zhang, T. (2014). Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *The Annals of Statistics* **42** 2164–2201.

Wasserman, L., Kolar, M., Rinaldo, A. et al. (2014). Berry-esseen bounds for estimating undirected graphs. *Electronic Journal of Statistics* **8** 1188–1224.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Annals of statistics* **37** 2178–2201.

Whittaker, J. (2009). *Graphical Models in Applied Multivariate Statistics*. Wiley Publishing.

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

Xue, L., Zou, H., Cai, T. et al. (2012a). Nonconcave penalized composite conditional likelihood estimation of sparse Ising models. *The Annals of Statistics* **40** 1403–1429.

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

Yang, E., Ravikumar, P., Allen, G. I., Baker, Y., Wan, Y.-W. and Liu, Z. (2014). A General Framework for Mixed Graphical Models. *ArXiv e-prints*.

Yang, E., Ravikumar, P., Allen, G. I. and Liu, Z. (2013a). On graphical models via univariate exponential family distributions. *arXiv preprint arXiv:1301.4183*.

Yang, E., Ravikumar, P., Allen, G. I. and Liu, Z. (2013b). On poisson graphical models. *Advances in Neural Information Processing Systems* 1718–1726.

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

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35.

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

Zhang, C.-H. et al. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38** 894–942.
Zhang, T. (2010). Analysis of multi-stage convex relaxation for sparse regularization. *The Journal of Machine Learning Research* **11** 1081–1107.

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

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *The Journal of Machine Learning Research* **7** 2541–2563.

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