Biometrika (2016), **99**, 1, *pp*. 1–17

Printed in Great Britain

Advance Access publication on 2016

Replicates in high dimensions, with applications to latent variable graphical models

BY KEAN MING TAN

Department of Operations Research and Financial Engineering, Princeton University, Princeton, New Jersey 08544, U.S.A.

kmtan@princeton.edu

YANG NING

Department of Statistical Science, Cornell University, Ithaca, New York 14853, U.S.A.

yn265@cornell.edu

DANIELA M. WITTEN

Department of Statistics, University of Washington, Seattle, Washington 98195, U.S.A.

dwitten@uw.edu

AND HAN LIU

Department of Operations Research and Financial Engineering, Princeton University, Princeton, New Jersey 08544, U.S.A.

hanliu@princeton.edu

SUMMARY

In classical statistics, much thought has been put into experimental design and data collection. In the high-dimensional setting, however, experimental design has been less of a focus. In this paper, we stress the importance of collecting multiple replicates for each subject in this setting. We consider learning the structure of a graphical model with latent variables, under the assumption that these variables take a constant value across replicates within each subject. By collecting multiple replicates for each subject, we are able to estimate the conditional dependence relationships among the observed variables given the latent variables. To test the null hypothesis of conditional independence between two observed variables, we propose a pairwise decorrelated score test. Theoretical guarantees are established for parameter estimation and for this test. We show that our proposal is able to estimate latent variable graphical models more accurately than some existing proposals, and apply the proposed method to a brain imaging dataset.

Some key words: Experimental design; Nuisance parameter; Pairwise decorrelated score test; Semiparametric exponential family graphical model.

1. INTRODUCTION

Experimental design and data collection have been the subjects of extensive research (Box et al., 2005; Montgomery, 2008). For instance, randomised clinical trials are conducted to determine the treatment effect of a new drug, and sample size calculations are performed to determine the smallest number of patients needed to give sufficient power to detect the treatment effect. In

contrast, in the high-dimensional setting, statisticians are usually not involved in experimental design and data collection. Given a cost constraint, investigators often try to obtain the largest possible number of subjects; that is, replicates are typically not collected for each subject. In this paper, we show that collecting replicates aids when learning an undirected graphical model with latent variables.

In an undirected graphical model, each node represents a random variable, and an edge connecting a pair of nodes indicates that the two variables are conditionally dependent, given all the other variables. The Gaussian graphical model has been studied extensively (Meinshausen & Bühlmann, 2006; Yuan & Lin, 2007; Friedman et al., 2008; Rothman et al., 2008; Peng et al., 2009; Ravikumar et al., 2011; Cai et al., 2011; Sun & Zhang, 2013). Other authors have considered extensions to the case in which each node-conditional distribution belongs to a univariate exponential family (Ravikumar et al., 2010; Yang et al., 2015; Lee & Hastie, 2015; Chen et al., 2015). Others have considered estimating conditional dependence relationships using semiparametric or nonparametric approaches (Liu et al., 2009, 2012; Fellinghauer et al., 2013; Voorman et al., 2014).

However, in many scientific studies, we observe only a subset of the relevant variables. For instance, in the context of a gene expression study, certain patients may have undiagnosed disease or some unknown risk factors. If the heterogeneity among patients is ignored, then the estimated conditional relationships among the genes may be distorted. This is made apparent in recent work on Gaussian graphical modelling in the presence of latent variables (Chandrasekaran et al., 2012), which showed that after marginalizing over the latent variables, the conditional independence graph corresponding to the observed variables may be dense.

In this paper, we propose an estimator and develop theory for the semiparametric exponential family graphical model with latent variables. This work builds upon an unpublished 2014 technical report by Yang et al. (arXiv:1412.8697), in which the semiparametric exponential family graphical model was introduced. We assume that these variables are constant across replicates within a given subject and that we have at least two replicates per subject. We exploit the replicates in order to construct a nuisance-free loss function that does not depend on the latent variables. In addition, we propose a pairwise decorrelated score test of the null hypothesis that two variables are conditionally independent, given all the other variables.

2. A MODEL FOR LATENT VARIABLE GRAPHICAL MODELS

2·1. *Review of the semiparametric exponential family graphical model*

We provide a brief review of the semiparametric exponential family graphical model proposed in Yang et al. (arXiv:1412:8697). Let X be a p-dimensional random vector and let $X_{-j} = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p)^{\mathrm{T}}$. The p-dimensional random vector X follows the semiparametric exponential family graphical model if, for any node j, the conditional density of X_j given X_{-j} satisfies

$$p(x_j \mid x_{-j}) = \exp\left\{x_j \beta_{j,-j}^{\mathrm{T}} x_{-j} + f_j(x_j) - A_j(\beta_j, f_j)\right\}, \tag{1}$$

where $\beta_{j,-j}$ encodes the conditional dependence relationships between the jth node and the other nodes, $f_j(\cdot)$ is an unknown function, and $A_j(\cdot)$ is the log-partition function. Because $f_j(\cdot)$ is unknown, obtaining the maximum likelihood estimator of (1) may be infeasible. To estimate $\beta_{j,-j}$, we can instead construct a loss function that does not depend on $f_j(\cdot)$.

Let X_i and x_i be the random variables and data corresponding to the ith subject, respectively. Let $x_{\cdot j} = (x_{1j}, \ldots, x_{nj})^{\mathrm{T}}$, and let $x_j^{(\cdot)}$ and $z_{\cdot j}$ be the order and rank statistics of $x_{\cdot j}$, respec-

tively. For instance, if $x_{\cdot j} = (1, 5, 2)^{\mathrm{T}}$, then $x_j^{(\cdot)} = (1, 2, 5)^{\mathrm{T}}$ and $z_{\cdot j} = (1, 3, 2)^{\mathrm{T}}$. Furthermore, let $x_{\cdot,-j}$ denote an $n \times (p-1)$ matrix obtained by stacking the vectors $x_{\cdot k}$ for $k \neq j$. The joint conditional density of the jth variable given the others can be decomposed as

$$p(x_{\cdot j} \mid x_{\cdot,-j}, \beta_{j,-j}, f_j) = p\left\{z_{\cdot j} \mid x_{\cdot,-j}, x_j^{(\cdot)}, \beta_{j,-j}\right\} p\left\{x_j^{(\cdot)} \mid x_{\cdot,-j}, \beta_{j,-j}, f_j\right\},$$

the product of the conditional density of the rank statistics given the order statistics, and the density of the order statistics. The former does not depend on $f_j(\cdot)$: the key insight is that the rank statistics given the order statistics have no information about $f_j(\cdot)$. Rather than estimating $\beta_{j,-j}$ from the joint conditional density that involves the unknown function $f_j(\cdot)$, we can estimate $\beta_{j,-j}$ by maximising the conditional density of the rank statistics.

However, computing the conditional density of the rank statistics may be computationally prohibitive. Thus, we can consider the conditional density formed by a single pair of samples, and construct a nuisance-free likelihood function by multiplying the conditional densities of the $n(n-1)/2$ pairs of samples. This approach is also considered in Ning et al. (2016) in the context of semiparametric regression.

2·2. *Semiparametric exponential family graphical models with latent variables*

We generalise the semiparametric exponential family graphical model to accommodate latent variables. Let $X = (X_O^{\mathrm{T}}, X_H^{\mathrm{T}})^{\mathrm{T}}$ be a $(p+h)$-dimensional random vector, where $X_O \in \mathbb{R}^p$ and $X_H \in \mathbb{R}^h$ are the vectors of observed and latent random variables, respectively. We let $O = \{1, \ldots, p\}$ and $H = \{p+1, \ldots, p+h\}$ denote the index sets of the observed and latent random variables, respectively.

DEFINITION 1. *A $(p+h)$-dimensional random vector $X = (X_O^{\mathrm{T}}, X_H^{\mathrm{T}})^{\mathrm{T}}$ follows a semiparametric exponential family graphical model with latent variables, if for any node j, the conditional density of X_j given X_{-j} satisfies*

$$p(x_j \mid x_{-j}) = \exp\left\{x_j \beta_{j,-j}^{\mathrm{T}} x_{-j} + f_j(x_j) - A_j(\beta_j, f_j)\right\},$$

where $f_j(x_j)$ is some possibly unknown function and $A_j(\beta_j, f_j)$ is the log-partition function.

For any $j \in O$, we write $X_{O\setminus j} = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^{p-1}$ and $X_{-j} = (X_{O\setminus j}^{\mathrm{T}}, X_H^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{p+h-1}$. Let $\beta_{j,O\setminus j} = (\beta_{j1}, \ldots, \beta_{j,j-1}, \beta_{j,j+1}, \ldots, \beta_{jp})^{\mathrm{T}} \in \mathbb{R}^{p-1}$, $\beta_{j,H} = (\beta_{j,p+1}, \ldots, \beta_{j,p+h})^{\mathrm{T}} \in \mathbb{R}^h$, and $\beta_j = (\beta_{j,O\setminus j}^{\mathrm{T}}, \beta_{j,H}^{\mathrm{T}})^{\mathrm{T}}$. The model introduced in Definition 1 can be rewritten as

$$p(x_j \mid x_{-j}) = \exp\left\{x_j \beta_{j,O\setminus j}^{\mathrm{T}} x_{O\setminus j} + x_j \beta_{j,H}^{\mathrm{T}} x_H + f_j(x_j) - A_j(\beta_j, f_j)\right\}. \tag{2}$$

The parameters $\beta_{j,O\setminus j}$ and $\beta_{j,H}$ encode the conditional dependence relationships between the jth node and all the other observed and latent variables, respectively. In particular, $\beta_{jk} = 0$ if and only if the jth and kth nodes are conditionally independent, given all the other nodes.

In this paper, we assume that: $\beta_{jk} = \beta_{kj}$ and $\exp\{\sum_{j=1}^{p+h} \sum_{k\neq j} \beta_{jk} x_j x_k / 2 + \sum_{j=1}^{p+h} f_j(x_j)\}$ is integrable with respect to its measure. By an application of Proposition 1 in Chen et al. (2015), under these conditions, there exists a joint probability distribution for the model introduced in Definition 1 that takes the form

$$p(x) \propto \exp\left\{\frac{1}{2} \sum_{j=1}^{p+h} \sum_{k\neq j} \beta_{jk} x_j x_k + \sum_{j=1}^{p+h} f_j(x_j)\right\}. \tag{3}$$

We provide two special cases of (2), and consider them in Section 4.

Example 1. The Gaussian graphical model with latent variables: let $X = (X_O^{\mathrm{T}}, X_H^{\mathrm{T}})^{\mathrm{T}} \sim N(0, \Sigma)$, where $\Sigma \in \mathbb{R}^{(p+h)\times(p+h)}$ and let $\Theta = \Sigma^{-1}$. For $j \in O$, the conditional density of X_j given all the other variables is

$$p(x_j \mid x_{-j}) = \left(\frac{\Theta_{jj}}{2\pi}\right)^{1/2} \exp\left\{-x_j \Theta_{j,O\setminus j}^{\mathrm{T}} x_{O\setminus j} - x_j \Theta_{j,H}^{\mathrm{T}} x_H - \Theta_{jj} x_j^2/2 - \frac{\left(\Theta_{j,-j}^{\mathrm{T}} x_{-j}\right)^2}{2\Theta_{jj}}\right\}.$$

Comparing this to (2), we see that $\beta_{j,O\setminus j} = -\Theta_{j,O\setminus j}$, $\beta_{j,H} = -\Theta_{j,H}$, $f_j(x_j) = -\Theta_{jj} x_j^2/2$, and $A_j(\beta_j, f_j) = (\sum_{k\neq j} \Theta_{jk} x_k)^2/(2\Theta_{jj}) + \log(2\pi/\Theta_{jj})/2$.

Example 2. The Ising model with latent variables: let $X_j \in \{0, 1\}$ with joint density $p(x) \propto \exp(\sum_{j<k} \Theta_{jk} x_j x_k)$. For $j \in O$, the conditional density of X_j given the other variables is

$$p(x_j \mid x_{-j}) = \exp\left[x_j \Theta_{j,O\setminus j}^{\mathrm{T}} x_{O\setminus j} + x_j \Theta_{j,H}^{\mathrm{T}} x_H - \log\left\{1 + \exp\left(\Theta_{j,-j}^{\mathrm{T}} x_{-j}\right)\right\}\right].$$

Comparing this to (2), we see that $\beta_{j,O\setminus j} = \Theta_{j,O\setminus j}$, $\beta_{j,H} = \Theta_{j,H}$, $f_j(x_j) = 0$, and $A_j(\beta_j, f_j) = \log\{1 + \exp(\Theta_{j,-j}^{\mathrm{T}} x_{-j})\}$.

2·3. *From replicates to a nuisance-free loss function*

Recall that we are interested in estimating the conditional dependence relationships among the observed variables given the latent variables, $\beta_{j,O\setminus j}$. Due to the presence of the possibly unknown function $f_j(x_j)$ and the latent variables x_H in (2), it is not possible to directly maximise (2) with respect to $\beta_{j,O\setminus j}$. By collecting multiple replicates per subject, and assuming that the latent variables are constant across replicates for a given subject, we construct a loss function that does not depend on the latent variables and the unknown function.

Let R_i be the number of replicates for the ith subject. To simplify bookkeeping, we assume that $R_1 = \cdots = R_n = R$, though this assumption is not critical. Suppose that X_i^r, the random vector for the rth replicate for the ith subject is distributed as in (3), for $i = 1, \ldots, n$ and $r = 1, \ldots, R$. Throughout the paper, we assume that X_i^r and $X_{i'}^r$ are independent, while X_i^r and $X_i^{r'}$ may be dependent. Let $x_i^r = \{(x_{iO}^r)^{\mathrm{T}}, (x_{iH}^r)^{\mathrm{T}}\}^{\mathrm{T}}$ be the data corresponding to the rth replicate of the ith subject. We make two assumptions on the replicates.

Assumption 1. The latent variables are constant across replicates, that is, $x_{iH}^r = x_{iH}^{r'} = x_{iH}$ for all $1 \leq r' \leq r \leq R$.

Assumption 2. Given the latent variables, the R replicates are mutually independent. That is, $p(x_{iO}^1, \ldots, x_{iO}^R \mid x_{iH}) = \prod_{r=1}^R p(x_{iO}^r \mid x_{iH})$.

Assumptions 1 and 2 are plausible in many scientific settings. For instance, consider a gene expression study in which the expression levels of thousands of genes are measured for a number of subjects. Certain subjects may have unknown risk factors that might be associated with their gene expression levels. In this setting, the observed variables are the genes, and the latent variables represent unknown risk factors. Assumption 1 is satisfied if the disease status or the unknown risk factors do not change across time. Assumption 2 is likely to be satisfied, if the gene expression levels are measured in multiple independent clinical visits.

We now construct a nuisance-free loss function using an approach similar to the one outlined in Section 2·1, by exploiting the fact that R replicates are available for each subject. Under

Assumption 2, the joint conditional density for the ith subject for $j \in O$ is

$$p(x_{ij}^1, \ldots, x_{ij}^R \mid x_{i,-j}^1, \ldots, x_{i,-j}^R) = \prod_{r=1}^{R} p(x_{ij}^r \mid x_{i,-j}^r).$$

Estimating $\beta_{j,O\setminus j}$ by maximising the joint conditional density, which depends on both the unmeasured data x_{iH} and the possibly unknown function $f_j(\cdot)$, may not be feasible.

Let $x_{ij}^{(r,r')} = \{\min(x_{ij}^r, x_{ij}^{r'}), \max(x_{ij}^r, x_{ij}^{r'})\}$ be the order statistics of a pair of replicates for the ith subject. The joint conditional density for the pair of replicates is

$$\begin{aligned} &p\left(X_{ij}^r = x_{ij}^r, X_{ij}^{r'} = x_{ij}^{r'} \mid x_{i,-j}^r, x_{i,-j}^{r'}\right) \\ &= p\left\{X_{ij}^r = x_{ij}^r, X_{ij}^{r'} = x_{ij}^{r'} \mid x_{i,-j}^r, x_{i,-j}^{r'}, x_{ij}^{(r,r')}\right\} p\left\{x_{ij}^{(r,r')} \mid x_{i,-j}^r, x_{i,-j}^{r'}\right\}. \end{aligned} \tag{4}$$

The following proposition shows that the conditional density $p\{X_{ij}^r = x_{ij}^r, X_{ij}^{r'} = x_{ij}^{r'} \mid x_{i,-j}^r, x_{i,-j}^{r'}, x_{ij}^{(r,r')}\}$ does not depend on x_{iH}^r, $x_{iH}^{r'}$, or on the unknown function $f_j(\cdot)$.

PROPOSITION 1. *Under Assumptions* 1 *and* 2*, for* $j \in O$*,*

$$p\left\{X_{ij}^r = x_{ij}^r, X_{ij}^{r'} = x_{ij}^{r'} \mid x_{i,-j}^r, x_{i,-j}^{r'}, x_{ij}^{(r,r')}\right\} = \left\{1 + R_{ij}^{rr'}(\beta_{j,O\setminus j})\right\}^{-1}, \tag{5}$$

where $R_{ij}^{rr'}(\beta_{j,O\setminus j}) = \exp\{-(x_{ij}^r - x_{ij}^{r'})\beta_{j,O\setminus j}^{\mathrm{T}}(x_{i,O\setminus j}^r - x_{i,O\setminus j}^{r'})\}$.

In the absence of latent variables, similar results were established in the context of the semiparametric generalised linear model (Equation (3.4) in Ning et al., 2016) and the semiparametric exponential family graphical model (Equation (3.1) in Yang et al. (arXiv:1412.8697)), both of which applied the approach outlined in Section 2·1.

Remark 1. When Assumption 1 is violated, the conditional density (5) takes the form

$$\left[1 + \exp\left\{-(x_{ij}^r - x_{ij}^{r'})\beta_{j,O\setminus j}^{\mathrm{T}}(x_{i,O\setminus j}^r - x_{i,O\setminus j}^{r'}) - (x_{ij}^r - x_{ij}^{r'})\beta_{j,H}^{\mathrm{T}}(x_{iH}^r - x_{iH}^{r'})\right\}\right]^{-1},$$

which has an additional term $(x_{ij}^r - x_{ij}^{r'})\beta_{j,H}^{\mathrm{T}}(x_{iH}^r - x_{iH}^{r'})$ that depends on the latent variables. Provided that $|x_{iH}^r - x_{iH}^{r'}|$ is sufficiently small, this term is ignorable, and therefore it has negligible effect on the estimation of $\beta_{j,O\setminus j}$.

To obtain an estimate of $\beta_{j,O\setminus j}$, we ignore the term $p\{x_{ij}^{(r,r')} \mid x_{i,-j}^r, x_{i,-j}^{r'}\}$ in (4), and consider the product of joint conditional densities over all pairs of replicates across the n subjects,

$$\prod_{i=1}^{n} \prod_{1 \le r < r' \le R} p\left\{x_{ij}^r, x_{ij}^{r'} \mid x_{i,-j}^r, x_{i,-j}^{r'}, x_{ij}^{(r,r')}\right\}.$$

This leads to a nuisance-free loss function that does not depend on the latent variables or on the unknown function, i.e.,

$$\ell_j(\beta_{j,O\setminus j}) = \frac{1}{n}\sum_{i=1}^{n}\left[\binom{R}{2}^{-1} \sum_{1 \le r < r' \le R} \log\left\{1 + R_{ij}^{rr'}(\beta_{j,O\setminus j})\right\}\right]. \tag{6}$$

From now onwards, we let $\beta_{j,O\setminus j}^*$ be the true parameter values in (2) that encode the underlying conditional dependence relationships between the jth variable and the observed variables.

Similarly, we let f_j^* be the underlying function in (2). The following proposition justifies the use of the loss function (6) for estimating $\beta^*_{j,O\setminus j}$.

PROPOSITION 2. *For all $j \in O$, $E\{\nabla \ell_j(\beta^*_{j,O\setminus j})\} = 0$ and $\beta^*_{j,O\setminus j}$ is a global minimizer of $E\{\ell_j(\beta_{j,O\setminus j})\}$, where $E(\cdot)$ is the expectation under the true parameters $(\beta^*_{j,O\setminus j}, f_j^*)$.*

To encourage the estimated parameter to contain many zero elements, we solve

$$\underset{\beta_{j,O\setminus j}\in\mathbb{R}^{p-1}}{\text{minimize}} \ \left\{\ell_j(\beta_{j,O\setminus j}) + \lambda\|\beta_{j,O\setminus j}\|_1\right\}, \tag{7}$$

where λ is a non-negative tuning parameter that controls the sparsity of the estimate $\hat{\beta}_{j,O\setminus j}$. The loss function (6) can be interpreted as a logistic loss function with $x^r_{ij} - x^{r'}_{ij}$ as the outcome and $x^r_{i,O\setminus j} - x^{r'}_{i,O\setminus j}$ as the covariates. We create a pseudo-binary outcome $\tilde{x}^{rr'}_{ij} = \text{sign}(x^r_{ij} - x^{r'}_{ij})$ and pseudo covariates $\tilde{x}^{rr'}_{i,O\setminus j} = (x^r_{i,O\setminus j} - x^{r'}_{i,O\setminus j})|x^r_{ij} - x^{r'}_{ij}|$. We can then solve (7) using the R package glmnet for fitting an ℓ_1-penalised logistic regression to obtain an estimate of $\beta_{j,O\setminus j}$. When there are ties in the outcome, that is, $x^r_{ij} = x^{r'}_{ij}$, we ignore the pair of observations, since its contribution to the loss function (6) is free of the parameter of interest, $\beta_{j,O\setminus j}$.

2·4. *Pairwise decorrelated score test*

In this section, we consider testing a pre-specified component in $\beta^*_{j,O\setminus j}$ and $\beta^*_{k,O\setminus k}$, that is,

$$H_0 : \beta^*_{jk} = \beta^*_{kj} = 0 \qquad \text{versus} \qquad H_1 : \beta^*_{jk} = \beta^*_{kj} \neq 0, \tag{8}$$

for any $j, k \in O$, by treating the remaining parameters $\beta^*_{j,O\setminus\{j,k\}}$ and $\beta^*_{k,O\setminus\{j,k\}}$ as nuisance parameters. The classical score test is often used for this purpose in the low-dimensional setting. However, in the high-dimensional setting, the score test statistic is not asymptotically normal, because the number of nuisance parameters is large. We propose a pairwise decorrelated score test to test the null hypothesis given in (8). The test is constructed so that the effect of the nuisance parameters is asymptotically negligible. The decorrelated score test has been previously considered in Ning & Liu (2016), Ning et al. (2016), and Yang et al. (arXiv:1412.8697).

Let $\nabla \ell_j(\beta_{j,O\setminus j}) \in \mathbb{R}^{p-1}$ and $\nabla^2 \ell_j(\beta_{j,O\setminus j}) \in \mathbb{R}^{(p-1)\times(p-1)}$ be the gradient and the Hessian of the loss function $\ell_j(\beta_{j,O\setminus j})$ in (6), respectively. For $k \in O \setminus j$, we let

$$\nabla_k \ell_j(\beta_{j,O\setminus j}) = \frac{\partial \ell_j(\beta_{j,O\setminus j})}{\partial \beta_{jk}} \in \mathbb{R}, \qquad \nabla_{-k} \ell_j(\beta_{j,O\setminus j}) = \frac{\partial \ell_j(\beta_{j,O\setminus j})}{\partial \beta_{j,O\setminus\{j,k\}}} \in \mathbb{R}^{p-2}.$$

Similarly, for $k \in O \setminus j$, we let

$$\nabla^2_{k,-k} \ell_j(\beta_{j,O\setminus j}) = \frac{\partial^2 \ell_j(\beta_{j,O\setminus j})}{\partial \beta_{jk} \partial \beta_{j,O\setminus\{j,k\}}} \in \mathbb{R}^{p-2}, \qquad \nabla^2_{-k,-k} \ell_j(\beta_{j,O\setminus j}) = \frac{\partial^2 \ell_j(\beta_{j,O\setminus j})}{(\partial \beta_{j,O\setminus\{j,k\}})^2} \in \mathbb{R}^{(p-2)\times(p-2)}.$$

Define $H^j = E\{\nabla^2 \ell_j(\beta^*_{j,O\setminus j})\} \in \mathbb{R}^{(p-1)\times(p-1)}$, and for $k \in O \setminus j$, let

$$H^j_{k,-k} = E\left\{\nabla^2_{k,-k} \ell_j(\beta^*_{j,O\setminus j})\right\} \in \mathbb{R}^{p-2}, \qquad H^j_{-k,-k} = E\left\{\nabla^2_{-k,-k} \ell_j(\beta^*_{j,O\setminus j})\right\} \in \mathbb{R}^{(p-2)\times(p-2)}.$$

Let $(w^*_{jk})^{\mathrm{T}} = (H^j_{k,-k})^{\mathrm{T}}(H^j_{-k,-k})^{-1} \in \mathbb{R}^{p-2}$, and let $\beta_{j\vee k} = (\beta_{jk}, \beta^{\mathrm{T}}_{j,O\setminus\{j,k\}}, \beta^{\mathrm{T}}_{k,O\setminus\{j,k\}})^{\mathrm{T}} \in \mathbb{R}^{2p-3}$ denote the parameters associated with the loss functions for the jth and kth observed

variables. The pairwise decorrelated score function for β_{jk} is defined as

$$S_{jk}(\beta_{j\vee k}) = \nabla_k \ell_j(\beta_{j,O\setminus j}) + \nabla_j \ell_k(\beta_{k,O\setminus k}) - (w^*_{jk})^{\mathrm{T}} \nabla_{-k} \ell_j(\beta_{j,O\setminus j}) - (w^*_{kj})^{\mathrm{T}} \nabla_{-j} \ell_k(\beta_{k,O\setminus k}). \tag{9}$$

The last two terms in (9) are constructed so that the effect of the nuisance parameters on the score function is asymptotically negligible (Section 3·2 of Ning et al., 2016).

The pairwise decorrelated score function (9) depends on the unknown quantities w^*_{jk} and w^*_{kj}. We estimate them using a Dantzig selector type estimator (Candès & Tao, 2007),

$$\hat{w}_{jk} = \underset{w\in\mathbb{R}^{p-2}}{\arg\min} \|w\|_1 \quad \text{subject to} \quad \left\| \nabla^2_{k,-k} \ell_j(0, \hat{\beta}_{j,O\setminus\{j,k\}}) - w^{\mathrm{T}} \nabla^2_{-k,-k} \ell_j(0, \hat{\beta}_{j,O\setminus\{j,k\}}) \right\|_\infty \le \lambda_w, \tag{10}$$

where $(0, \hat{\beta}_{j,O\setminus\{j,k\}})$ is an estimate of $\beta_{j,O\setminus j}$ obtained by solving (7) and replacing $\hat{\beta}_{jk}$ with zero, and λ_w is a non-negative tuning parameter. With some abuse of notation in (10), we use the notation $(0, \hat{\beta}_{j,O\setminus\{j,k\}})$ to indicate $(\hat{\beta}_{j1}, \ldots, \hat{\beta}_{j,k-1}, 0, \hat{\beta}_{j,k+1}, \ldots, \hat{\beta}_{jp})$.

The estimated pairwise decorrelated score function for testing $\beta^*_{jk} = 0$ is obtained by replacing $\beta_{j,O\setminus j}$, $\beta_{k,O\setminus k}$, w^*_{jk}, and w^*_{kj} in (9) with the estimated parameters $(0, \hat{\beta}_{j,O\setminus\{j,k\}})$, $(0, \hat{\beta}_{k,O\setminus\{j,k\}})$, $\hat{w}_{jk}$, and $\hat{w}_{kj}$, respectively, leading to

$$\hat{S}_{jk} = \nabla_k \ell_j(0, \hat{\beta}_{j,O\setminus\{j,k\}}) + \nabla_j \ell_k(0, \hat{\beta}_{k,O\setminus\{j,k\}}) - \hat{w}^{\mathrm{T}}_{jk} \nabla_{-k} \ell_j(0, \hat{\beta}_{j,O\setminus\{j,k\}}) - \hat{w}^{\mathrm{T}}_{kj} \nabla_{-j} \ell_k(0, \hat{\beta}_{k,O\setminus\{j,k\}}). \tag{11}$$

Let

$$\hat{\sigma}^2_{jk} = \hat{\Sigma}^{jk}_{jk,jk} - 2\hat{\Sigma}^{jk}_{jk,j\setminus k}\hat{w}_{jk} - 2\hat{\Sigma}^{jk}_{jk,k\setminus j}\hat{w}_{kj} + \hat{w}^{\mathrm{T}}_{jk}\hat{\Sigma}^{jk}_{j\setminus k,j\setminus k}\hat{w}_{jk} + \hat{w}^{\mathrm{T}}_{kj}\hat{\Sigma}^{jk}_{k\setminus j,k\setminus j}\hat{w}_{kj}, \tag{12}$$

where $\hat{\Sigma}^{jk}$ is to be defined in (16). For a given significance level $0 < \alpha < 1$, our proposed pairwise decorrelated score test takes the form

$$\psi_{jk}(\alpha) = \begin{cases} 1, & \left| n^{1/2} \hat{S}_{jk} / \hat{\sigma}_{jk} \right| > \Phi^{-1}(1-\alpha/2), \\ 0, & \text{otherwise}, \end{cases} \tag{13}$$

where $\Phi(x)$ is the standard normal cumulative distribution function. We will show in Section 3·3 that under the null hypothesis given in (8), the type I error of $\psi_{jk}(\alpha)$ converges to α. We summarise the overall procedure for conducting the pairwise decorrelated score test for (8) in Algorithm 1.

Algorithm 1. Pairwise decorrelated score test for testing $H_0 : \beta^*_{jk} = \beta^*_{kj} = 0$.

1. Obtain $\hat{\beta}_{j,O\setminus j}$ and $\hat{\beta}_{k,O\setminus k}$ by solving the optimization problem (7).
2. Obtain $\hat{w}_{jk}$ and $\hat{w}_{kj}$ from (10).
3. Calculate the estimated pairwise decorrelated score function $\hat{S}_{jk}$ as in (11).
4. Calculate $\hat{\sigma}^2_{jk}$ as defined in (12).
5. Reject the null hypothesis $H_0 : \beta^*_{jk} = \beta^*_{kj} = 0$ if $|n^{1/2}\hat{S}_{jk}/\hat{\sigma}_{jk}| > \Phi^{-1}(1-\alpha/2)$, where $0 < \alpha < 1$ is the given significance level.

3. THEORETICAL RESULTS

3·1. *Notation*

We use the Landau symbol $f(n) = \mathcal{O}\{g(n)\}$ to indicate the existence of a constant $C > 0$ such that $f(n) \leq Cg(n)$ for two sequences $f(n)$ and $g(n)$. We write $f(n) = \Omega\{g(n)\}$ to indicate $g(n) = \mathcal{O}\{f(n)\}$. In addition, we write $f(n) = o\{g(n)\}$ if $\lim_{n\to\infty} f(n)/g(n) \to 0$. We use the stochastic Landau symbol $f(n) = \mathcal{O}_{\mathbb{P}}\{g(n)\}$ to indicate that $f(n) = \mathcal{O}\{g(n)\}$ with high probability. For a vector $v = (v_1, \ldots, v_d)^{\mathrm{T}} \in \mathbb{R}^d$, we let $v^{\otimes 2}$ denote the outer product vv^{T}. For a symmetric matrix $M \in \mathbb{R}^{d\times d}$, we let $\|M\|_\infty = \max_{1\leq j,j'\leq d} |M_{jj'}|$. Also, let $\Lambda_{\min}(M)$ and $\Lambda_{\max}(M)$ denote the minimum and maximum eigenvalues of M, respectively.

3·2. *Parameter estimation*

We provide an upper bound on the estimation error of $\hat{\beta}_{j,O\setminus j}$ obtained from solving (7). We study the asymptotic regime in which both n and p are allowed to grow, with R and h fixed. Proofs are deferred to the Supplementary Material. We first state an assumption on the first moment of the random variables and the local smoothness of the log-partition function.

Assumption 3. Let β^*_j, f^*_j be the true parameters in (3), and define the univariate function $\bar{A}_j(\cdot) : \mathbb{R} \to \mathbb{R}$ as

$$\bar{A}_j(u) = \log \left[\int \exp \left\{ ux_j + \frac{1}{2} \sum_{j=1}^{p+h} \sum_{k\neq j} \beta^*_{jk} x_j x_k + \sum_{j=1}^{p+h} f^*_j(x_j) \right\} d\nu(x) \right].$$

For all $j \in O$, we assume the following: (i) $|E(X_j)| \leq \kappa_m$, and (ii) $\max_{u:|u|\leq 1} \bar{A}''_j(u) \leq \kappa_h$.

Assumption 3 allows us to control the tail behaviors of the random variables. The same assumption has been used in recent work on the mixed graphical model (Chen et al., 2015).

Let $\mathcal{S}_j = \{k : \beta^*_{jk} \neq 0, k \in O \setminus j\}$ be the support set of $\beta^*_{j,O\setminus j}$ and let $s_j = |\mathcal{S}_j|$ be the cardinality of the set $\mathcal{S}_j$. Let $s_{\max} = \max_{j\in O} s_j$. Let $\kappa^2_{\min}$, $\mathrm{RE}_{\min}$, and $\rho_{q,\min}$ be the compatibility factor, restricted eigenvalue, and weak cone invertibility factor. These depend on the minimal eigenvalues of the Hessian matrix of the loss function, and will be defined rigorously in the Supplementary Material. These quantities are commonly used to establish upper bounds for estimation error in the context of ℓ_1-penalised regression (Bickel et al., 2009; van de Geer & Bühlmann, 2009). We now establish upper bounds on the estimation error of $\hat{\beta}_{j,O\setminus j}$.

THEOREM 1. *Let $\lambda = C(\log^5 p/n)^{1/2}$ for some constant $C > 0$. For $j \in O$, assume the event*

$$\mathcal{A} = \left\{ \max_{1\leq i\leq n} \max_{1\leq r<r'\leq R} \left\| \left(x^r_{ij} - x^{r'}_{ij} \right) \left(x^r_{i,O\setminus j} - x^{r'}_{i,O\setminus j} \right) \right\|_\infty \leq M \right\}$$

and that $Ms_{\max}\lambda/\kappa^2_{\min} = o(1)$. Under Assumption 3*, there exists a constant $C' > 0$ such that*

$$\begin{aligned}
&\|\hat{\beta}_{j,O\setminus j} - \beta^*_{j,O\setminus j}\|_1 \leq C' s_{\max}\lambda/\kappa^2_{\min},\\
&\|\hat{\beta}_{j,O\setminus j} - \beta^*_{j,O\setminus j}\|_2 \leq C'(s_{\max})^{1/2}\lambda/\mathrm{RE}_{\min},\\
&\|\hat{\beta}_{j,O\setminus j} - \beta^*_{j,O\setminus j}\|_q \leq C'(s_{\max})^{1/q}\lambda/\rho_{q,\min}, \quad (q \geq 1),
\end{aligned}$$

with probability at least $1 - p^{-1}$.

Theorem 1 generalises Theorem 4.4 in Yang et al. (arXiv:1412.8697). Interestingly, we obtain the same rate of convergence even when latent variables are present. Our rate of convergence

does not depend on the number of latent variables h. Theorem 1 holds with high probability conditioned on the event $\mathcal{A}$. For binary or categorical variables, $\mathcal{A}$ holds with M constant. In the case of sub-exponential random variables, it can be shown that $\mathcal{A}$ holds with high probability, with $M = C \log^2 p$ for a sufficiently large constant C.

The upper bounds on the estimation error in Theorem 1 depend on the quantities $\kappa^2_{\min}$, $\mathrm{RE}_{\min}$, and $\rho_{q,\min}$. These conditions can be bounded below by a positive constant when the random variables follow a multivariate Gaussian distribution.

THEOREM 2. *Assume that* $s_{\max}(\log^9 p/n)^{1/2} = o(1)$. *Let*

$$\{(X^r_{iO})^{\mathrm{T}}, X^{\mathrm{T}}_{iH}\}^{\mathrm{T}} \sim N(0, \Sigma), \qquad \Sigma = \begin{pmatrix} \Sigma_{O,O} & \Sigma_{O,H} \\ \Sigma_{H,O} & \Sigma_{H,H} \end{pmatrix}.$$

Under Assumption 3*, for* n *sufficiently large, the quantities* $\kappa^2_{\min}$, $\mathrm{RE}_{\min}$, *and* $\rho_{q,\min}$ *are larger than* $C\Lambda_{\min}(\Sigma)$ *with probability at least* $1 - p^{-1}$ *for some constant* $C > 0$.

3·3. *Pairwise decorrelated score test*

In this section, we show that the type I error of the pairwise decorrelated score test in (13) converges to the desired significance level, under the null hypothesis $H_0 : \beta^*_{jk} = \beta^*_{kj} = 0$. We start by introducing some additional notation. Let

$$U^j_i(\beta^*_{j,O\setminus j}) = -\frac{2}{R(R-1)} \sum_{1 \le r < r' \le R} \frac{R^{rr'}_{ij}(\beta^*_{j,O\setminus j})(x^r_{ij} - x^{r'}_{ij})(x^r_{i,O\setminus j} - x^{r'}_{i,O\setminus j})}{1 + R^{rr'}_{ij}(\beta^*_{j,O\setminus j})} \in \mathbb{R}^{p-1}.$$

Furthermore, let $U^j_{ik}(\beta^*_{j,O\setminus j})$ be the element in $U^j_i(\beta^*_{j,O\setminus j})$ corresponding to the kth feature and let $U^j_{i,-k}(\beta^*_{j,O\setminus j}) \in \mathbb{R}^{p-2}$ be the vector obtained by removing the entry $U^j_{ik}(\beta^*_{j,O\setminus j})$ from $U^j_i(\beta^*_{j,O\setminus j})$. For $j, k \in O$, we let

$$g^{jk}_i(\beta^*_{j\vee k}) = \begin{Bmatrix} U^j_{ik}(\beta^*_{j,O\setminus j}) + U^k_{ij}(\beta^*_{k,O\setminus k}) \\ U^j_{i,-k}(\beta^*_{j,O\setminus j}) \\ U^k_{i,-j}(\beta^*_{k,O\setminus k}) \end{Bmatrix} \in \mathbb{R}^{2p-3} \tag{14}$$

and

$$\Sigma^{jk} = E\left[\left\{g^{jk}_i(\beta^*_{j\vee k})\right\}^{\otimes 2}\right] = \begin{Bmatrix} \Sigma^{jk}_{jk,jk} & \Sigma^{jk}_{jk,j\setminus k} & \Sigma^{jk}_{jk,k\setminus j} \\ (\Sigma^{jk}_{jk,j\setminus k})^{\mathrm{T}} & \Sigma^{jk}_{j\setminus k,j\setminus k} & \Sigma^{jk}_{j\setminus k,k\setminus j} \\ (\Sigma^{jk}_{jk,k\setminus j})^{\mathrm{T}} & (\Sigma^{jk}_{j\setminus k,k\setminus j})^{\mathrm{T}} & \Sigma^{jk}_{k\setminus j,k\setminus j} \end{Bmatrix} \in \mathbb{R}^{(2p-3)\times(2p-3)}. \tag{15}$$

The quantity Σ^{jk} can be estimated using

$$\hat{\Sigma}^{jk}\left(0, \hat{\beta}_{j,O\setminus\{j,k\}}, \hat{\beta}_{k,O\setminus\{j,k\}}\right) = \frac{1}{n}\sum_{i=1}^{n}\left\{g^{jk}_i(0, \hat{\beta}_{j,O\setminus\{j,k\}}, \hat{\beta}_{k,O\setminus\{j,k\}})\right\}^{\otimes 2}. \tag{16}$$

In what follows, we write $\hat{\Sigma}^{jk}$ to indicate $\hat{\Sigma}^{jk}(0, \hat{\beta}_{j,O\setminus\{j,k\}}, \hat{\beta}_{k,O\setminus\{j,k\}})$.

We now state several assumptions. Recall from (9) that the pairwise decorrelated score function depends on the quantity $(w^*_{jk})^{\mathrm{T}} = (H^j_{k,-k})^{\mathrm{T}}(H^j_{-k,-k})^{-1} \in \mathbb{R}^{p-2}$. The following assumption on the expected Hessian of the loss function guarantees that the pairwise decorrelated score function (9) is well-defined.

Assumption 4. Let $H^j = E\{\nabla^2 \ell_j(\beta^*_{j,O\setminus j})\}$. For all $j \in O$, assume that

$$0 < \Lambda^H_{\text{lower}} \leq \Lambda_{\min}(H^j) \leq \Lambda_{\max}(H^j) \leq \Lambda^H_{\text{upper}} < \infty.$$

The next assumption guarantees that $g_i^{jk}(\beta^*_{j\vee k})$ defined in (14) is not degenerate, in the sense that the variance of any linear combination of the elements of $g_i^{jk}(\beta^*_{j\vee k})$ is not equal to zero. It is needed to guarantee the existence of the asymptotic variance of the score function (9).

Assumption 5. For $j, k \in O$, assume that $\Lambda_{\min}(\Sigma^{jk}) \geq \Lambda^{\Sigma}_{\text{lower}} > 0$.

The following theorem establishes that under the null hypothesis $H_0 : \beta^*_{jk} = \beta^*_{kj} = 0$, the type I error of $\psi_{jk}(\alpha)$ in (13) converges to α, and the associated p-value is asymptotically uniformly distributed in the $[0, 1]$ interval.

THEOREM 3. *Let the pairwise decorrelated score test with significance level* $0 < \alpha < 1$, $\psi_{jk}(\alpha)$, *be as defined in* (13). *We reject the null hypothesis* $H_0 : \beta^*_{jk} = \beta^*_{kj} = 0$ *if* $\psi_{jk}(\alpha) = 1$. *The associated* p*-value is defined as* $\hat{p}_{jk} = 2\{1 - \Phi(|n^{1/2}\hat{S}_{jk}/\hat{\sigma}_{jk}|)\}$, *where*

$$\hat{\sigma}^2_{jk} = \hat{\Sigma}^{jk}_{jk,jk} - 2\hat{\Sigma}^{jk}_{jk,j\setminus k}\hat{w}_{jk} - 2\hat{\Sigma}^{jk}_{jk,k\setminus j}\hat{w}_{kj} + \hat{w}^{\mathrm{T}}_{jk}\hat{\Sigma}^{jk}_{j\setminus k,j\setminus k}\hat{w}_{jk} + \hat{w}^{\mathrm{T}}_{kj}\hat{\Sigma}^{jk}_{k\setminus j,k\setminus j}\hat{w}_{kj}.$$

Under Assumptions 3–5 *and scaling assumptions in Assumptions S*1–*S*2 *in the Supplementary Material,* $\lim_{n\to\infty} \text{pr}\{\psi_{jk}(\alpha) = 1 \mid H_0\} = \alpha$ *and* $\hat{p}_{jk}$ *converges to a uniform distribution on the interval* $[0, 1]$.

Results similar to Theorem 3 have been proven in the context of semiparametric regression and graphical models (Theorem 4.1 in Ning et al., 2016; Theorem 4.7 of Yang et al. (arXiv:1412.8697)).

4. SIMULATION STUDIES

4·1. *Overview and competing proposals*

Recall from Definition 1 that $\beta^*_{jk} \neq 0$ if and only if the jth and kth nodes are conditionally dependent, given all the other variables. To evaluate the performance across different methods, we define the true positive rate as the proportion of correctly identified non-zeros, and the false positive rate as the proportion of zeros that are incorrectly identified to be non-zeros. To examine the finite-sample performance of the pairwise decorrelated score test, we test the null hypothesis $H_0 : \beta^*_{jk} = 0$. The type I error and power are calculated as the proportion of falsely rejected H_0 and correctly rejected H_0, respectively.

Five approaches are compared in our simulation studies: our proposal; the low-rank plus sparse latent variable Gaussian graphical model (Chandrasekaran et al., 2012); the semiparametric exponential family graphical model in Yang et al. (arXiv:1412.8697); the graphical lasso (Friedman et al., 2008); and the neighbourhood selection procedure (Meinshausen & Bühlmann, 2006; Ravikumar et al., 2011). Our proposal, Meinshausen & Bühlmann (2006), Ravikumar et al. (2011), and the semiparametric exponential family graphical model yield asymmetric estimates of the edge set. To symmetrise the edge set, we consider both the intersection and union rules described in Meinshausen & Bühlmann (2006), and report the best results for the competing proposals. We report our results using only the union rule.

Since the competing methods cannot accommodate replicates, we apply them to all nR observations, treating the replicates as independent samples. Our proposal, Friedman et al. (2008), Meinshausen & Bühlmann (2006), Ravikumar et al. (2011), and the semiparametric exponential

family graphical model each involves one tuning parameter. We applied a fine grid of tuning parameter values to obtain the curves shown in Figs. 1–4. There are two tuning parameters for Chandrasekaran et al. (2012). We set the second tuning parameter to equal ten times the first, and consider a fine grid of the first. Similar results were obtained for different ratios of the two tuning parameters.

4·2. *Gaussian graphical models with latent variables*

Let $\Theta = \Sigma^{-1}$ be the inverse covariance matrix of a Gaussian distribution, so that from Example 1, $\beta^*_{jk} = -\Theta_{jk}$. We partition Θ and Σ into

$$\Theta = \begin{pmatrix} \Theta_{O,O} & \Theta_{O,H} \\ \Theta_{H,O} & \Theta_{H,H} \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{O,O} & \Sigma_{O,H} \\ \Sigma_{H,O} & \Sigma_{H,H} \end{pmatrix},$$

where $\Theta_{O,O}$, $\Theta_{O,H}$, and $\Theta_{H,H}$ encode the conditional dependence relationships among the observed variables, between the observed and latent variables, and among the latent variables. We construct $\Theta_{O,O}$ by randomly setting 10% of the off-diagonal entries to 0·3. For $\Theta_{O,H}$ and $\Theta_{H,H}$, we randomly set 80% of the off-diagonal entries to 0·3. To ensure the positive definiteness of Θ, we set $\Theta_{jj} = |\Lambda_{\min}(\Theta)| + 0{\cdot}2$ for $j = 1, \ldots, p + h$. Finally, we set $\Sigma = \Theta^{-1}$.

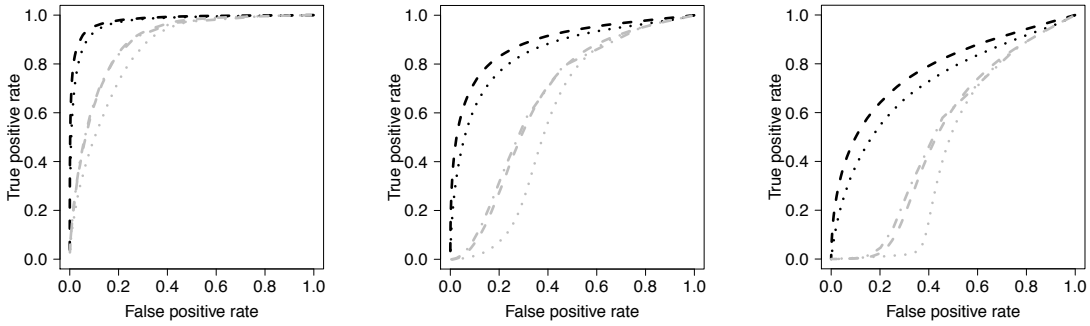
We first generate the latent variables x_{iH} for the n subjects from $N(0, \Sigma_{H,H})$. We then simulate the R replicates for each subject from the conditional distribution of the observed variables $N(\Sigma_{O,H}\Sigma^{-1}_{H,H}x_{iH}, \Sigma_{O,O} - \Sigma_{O,H}\Sigma^{-1}_{H,H}\Sigma_{H,O})$. The results for $n = 100$, $p = 100$, $h = \{2, 5, 10\}$, and $R = 10$, averaged over 100 datasets, are presented in Fig. 1.

In general, our proposal outperforms Friedman et al. (2008), Meinshausen & Bühlmann (2006), and the semiparametric exponential family graphical model, which do not model the latent variables. As shown in Fig. 1(a), our proposal has performance similar to Chandrasekaran et al. (2012), even though this is intended for the Gaussian setting which holds here, whereas our approach is semiparametric. As we increase the number of latent variables, the low-rank assumption of Chandrasekaran et al. (2012) is increasingly violated. Our proposal, which does not rely on the low-rank assumption, outperforms Chandrasekaran et al. (2012) when h is large.

Next, we investigate the role of the number of latent variables h and replicates R in the performance of our proposed method. We vary the ratio of R and h, while keeping $n = 100$ and $p = 100$ fixed. In addition, to study the tradeoff between n and R, we keep $p = 100$ and $h = 3$ fixed, and vary n and R with $nR = 600$. The results, averaged over 100 datasets, are shown in Fig. 2. From Figs. 2(a)–(b), we see that our proposal's performance improves as we increase the ratio R/h. From Fig. 2(c), we see that the performance of our method improves when $R > 2$. This suggests that for a fixed experimental budget, that is, keeping nR fixed, it may be beneficial to collect more than two replicates per sample.

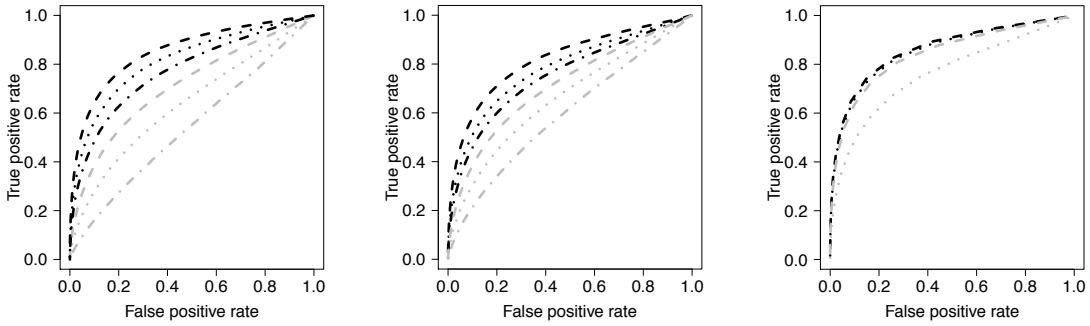
Our proposal relies on Assumption 1, which states that the latent variables are constant across replicates. We perform a sensitivity analysis by allowing the latent variables to vary across replicates within each subject. Let z^r_i be a h-vector with each element independently drawn from a uniform distribution $U[-\epsilon, \epsilon]$. We simulate the rth replicate for the ith observation from $N\{\Sigma_{O,H}\Sigma^{-1}_{H,H}(x_{iH} + z^r_i), \Sigma_{O,O} - \Sigma_{O,H}\Sigma^{-1}_{H,H}\Sigma_{H,O}\}$. We consider five values of $\epsilon = \{0, 1, 1{\cdot}5, 2, 2{\cdot}5\}$. Results averaged over 100 datasets are in Fig. 3, which shows that our proposal is robust to small perturbations of the latent variables.

We now perform the pairwise decorrelated score test described in Algorithm 1, in order to test the null hypothesis $H_0 : \beta^*_{jk} = 0$. The pairwise decorrelated score test involves two tuning parameters, λ in (7) and λ_w in (10). We select λ using 10-fold cross-validation, implemented in the R package glmnet. We use the R package fastclime to solve (10). We set $\lambda_w = 0{\cdot}06$, so that



(a) (b) (c)

Fig. 1: Results for the simulation study for the Gaussian graphical model with $n = 100$, $p = 100$, and $R = 10$. Panels (a), (b), and (c) correspond to $h = \{2, 5, 10\}$ latent variables, respectively. The different curves represent our proposal (long-dashed), Chandrasekaran et al. (2012) (dots), Meinshausen & Bühlmann (2006) (grey long-dashed), Friedman et al. (2008) (grey dots), and the semiparametric exponential family graphical model in Yang et al. (arXiv:1412.8697) (grey dot-dashed).



(a) (b) (c)

Fig. 2: Results of a simulation study investigating the relationship between h and R, and the tradeoff between n and R. Panels (a) and (b) are the results for $h = 8$ with $R = \{2, 4, 6, 8, 10, 12\}$, and $R = 6$ with $h = \{4, 5, 6, 8, 12, 24\}$, respectively, with $n = 100$ and $p = 100$. The curves represent different ratios R/h: 0·25 (grey dot-dashed), 0·5 (grey dots), 0·75 (grey long-dashed), 1 (dot-dashed), 1·25 (dots), and 1·5 (long-dashed). Panel (c) contains the results for $p = 100$ and $h = 3$, with different values of n and R such that $nR = 600$. The curves represent $n = 100$ and $R = 6$ (long-dashed), $n = 120$ and $R = 5$ (dots), $n = 150$ and $R = 4$ (grey long-dashed), and $n = 300$ and $R = 2$ (grey dots).

the estimates $\hat{w}_{jk}$ and $\hat{w}_{kj}$ contain a small number of non-zero entries. The results for $p = 100$, $R = 4$, and $h = 4$, over a range of sample sizes, are reported in Table 1. We see that the pairwise decorrelated score test is able to approximately control the type I error at level α =0·05.

4·3. *Ising model with latent variables*

We consider the Ising model with latent variables, as presented in Example 2. From Example 2, $\beta^*_{jk} = \Theta_{jk}$. We construct $\Theta_{O,O}$, $\Theta_{O,H}$, and $\Theta_{H,H}$ as in the previous section, but with nonzero

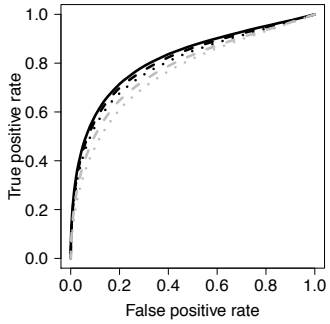


Fig. 3: Sensitivity analysis with $\mathrm{Unif}[-\epsilon, \epsilon]$ noise added to each replicate, with $\epsilon = \{1,\ 1{\cdot}5,\ 2,\ 2{\cdot}5\}$. Results are for $n = 100$, $p = 100$, $h = 4$, and $R = 6$. The curves correspond to $\epsilon = 0$ (solid), $\epsilon = 1$ (long-dashed), $\epsilon =1{\cdot}5$ (dots), $\epsilon = 2$ (grey long-dashed), and $\epsilon = 2{\cdot}5$ (grey dots).

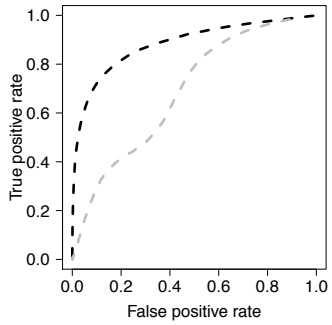


Fig. 4: Simulation results for the Ising model with $n = 100$, $p = 50$, $h = 5$, and $R = 10$, as described in Section 4·3. The curves represent our proposal (long-dashed) and the proposal of Ravikumar et al. (2011) (grey long-dashed).

entries drawn uniformly from $[-0{\cdot}5,-0{\cdot}25] \cup [0{\cdot}25,0{\cdot}5]$. Furthermore, we do not require Θ to be positive definite. To obtain samples from the joint density (3), we employ a Gibbs sampler as described in Section 4 of Guo et al. (2015). The results for $n = 100$, $p = 50$, $h = 5$, and $R = 10$, averaged over 100 data sets, are presented in Fig. 4. Our proposal outperforms that of Ravikumar et al. (2011), which does not model the latent variables.

As in Section 4·2, we perform the pairwise decorrelated score test of the null hypothesis $H_0 : \beta^*_{jk} = 0$. We set $\lambda_w = 0{\cdot}005$ in (10), so that the estimates $\hat{w}_{jk}$ and $\hat{w}_{kj}$ are sparse. The tuning parameter λ in (7) is again chosen by cross-validation. The type I error and power for $p = 50$, $R = 10$, and $h = 5$, over a range of sample sizes, are in Table 1. We see that the pairwise decorrelated score test is able to approximately control the type I error rate at level $\alpha = 0{\cdot}05$.

5. Application to ADHD-200 data

We applied our method to the ADHD-200 data (Biswal et al., 2010). The data consist of resting state functional magnetic resonance images on 197 subjects who have been diagnosed with attention deficit hyperactivity disorder, and 491 control subjects. The number of images for each subject ranges from 76 to 276. Covariates such as age, gender, site, and intelligence quotient

Table 1: Type I error and power of the pairwise decorrelated score test at the 5% significance level are calculated as the % of falsely rejected and correctly rejected null hypotheses, respectively, over 2000 data sets. Data were generated under the Gaussian graphical model with latent variables with $p = 100$, $R = 4$, and $h = 4$. Data were generated under the Ising model with latent variables with $p = 50$, $R = 10$, and $h = 5$

		$n = 50$	$n = 100$	$n = 200$	$n = 300$	$n = 400$
Gaussian	Type I error	9	7	6	5	5
	Power	18	27	45	59	70
Ising	Type I error	7	5	5	5	5
	Power	30	46	73	87	95

are also available. Similar to Power et al. (2011) and Qiu et al. (2016), we use 264 seed regions of interest to define the nodes in the graphical model.

We treat the images for each subject as replicates, and treat the covariates such as age and gender as latent variables. However, certain covariates such as age and gender serve as confounders that may alter the conditional dependence relationships among the variables. For instance, Qiu et al. (2016) showed that the brain networks at ages 7, 12, and 22 years are quite different. Biswal et al. (2010) showed that males and females have different brain connectivity networks. Thus, standard techniques for estimating graphical models that do not model the latent variables may yield inaccurate network estimates.

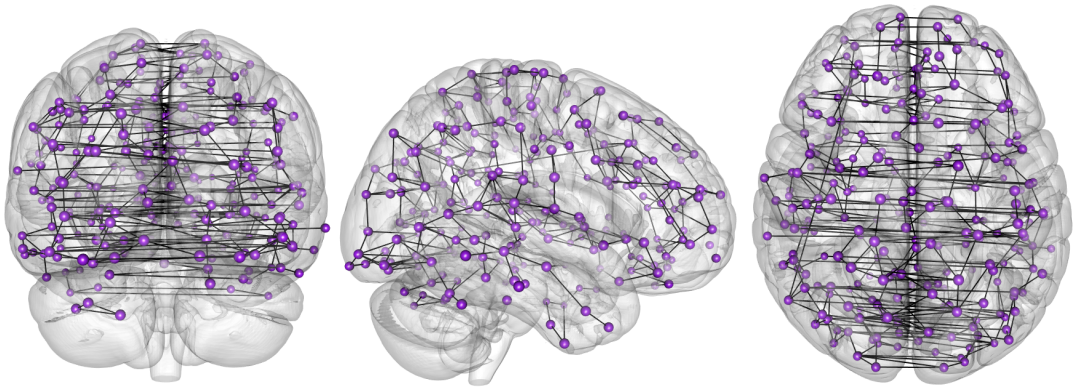
After removing subjects with missing values, we consider 465 control subjects in the dataset. For computational purposes, we choose $R = 10$ replicates randomly for each subject. Assumption 2 may not hold, since the replicate brain images for a given subject are very likely to be dependent. We standardize each seed region to have mean zero and standard deviation one for each subject. Our proposal (7) involves one tuning parameter λ. For visualization, we set $\lambda = 0{\cdot}2$ so that the estimated network is sparse, but in practice, λ can be chosen by cross-validation. We then symmetrise our estimates using the intersection rule described in Section 4. This yields an estimated network with 376 edges. Figs. 5(a)–(c) show coronal, sagittal, and transverse snapshots of the estimated brain connectivity network.

We compare our proposal to that of Friedman et al. (2008), which does not model the latent variables. We perform their proposal by treating the replicates as independent observations. For ease of comparison, the tuning parameter for Friedman et al. (2008) is chosen to yield 376 edges. The coronal, sagittal, and transverse snapshots of the estimated brain connectivity network from Friedman et al. (2008) are plotted in Figs. 5(d)–(f).

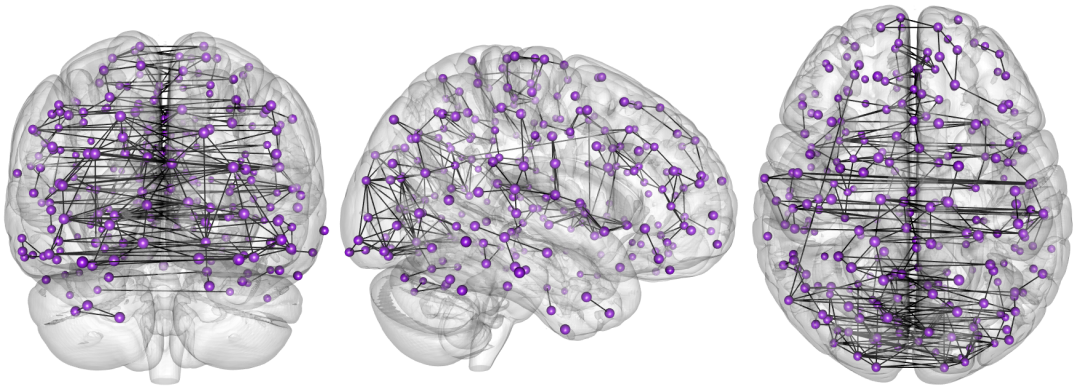
The two estimated networks are somewhat different. For instance, we see from Figs. 5(b) and 5(e) that the lower region of the brain connectivity network estimated by their proposal is more densely connected than that of our proposal. This might be a consequence of marginalizing over the latent variables, as discussed in Chandrasekaran et al. (2012). In contrast, edges in the network estimated by our proposal seem to be more spread throughout the network.

6. DISCUSSION

Our proposal can be generalised beyond estimating latent variable graphical models. For instance, in the context of regression, unmeasured confounders may remain constant across replicates. Without adjusting for these confounders, it can be shown that the estimated regression coefficients for the observed variables are biased. Using the ideas laid out in this paper, one can estimate the parameter of interest accurately by treating the confounders as nuisance parameters.



(a) Coronal, our proposal (b) Sagittal, our proposal (c) Transverse, our proposal



(d) Coronal, Friedman et al. (2008) (e) Sagittal, Friedman et al. (2008) (f) Transverse, Friedman et al. (2008)

Fig. 5: Coronal, sagittal, and transverse snapshots of the estimated brain connectivity networks resulting from our proposal and Friedman et al. (2008). Panels (a)–(c) and panels (d)–(f) contain the estimated networks from our proposal and Friedman et al. (2008), respectively.

Our model requires that the replicates are mutually conditionally independent given the latent variables; this is laid out in Assumption 2. In future work, it would be interesting to study whether that assumption can be relaxed.

An R package latentGraph will be made available on CRAN.

ACKNOWLEDGEMENT

We thank the editor, an associate editor, and two reviewers for helpful comments, Shizhe Chen for responding to our inquiries, and Huitong Qiu for providing us R code to plot Fig. 5.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes proofs of the theoretical results and the scaling assumptions used in Theorem 3.

REFERENCES

BICKEL, P. J., RITOV, Y. & TSYBAKOV, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37**, 1705–1732.

BISWAL, B. B., MENNES, M., ZUO, X.-N. et al. (2010). Toward discovery science of human brain function. *Proc. Natl. Acad. Sci.* **107**, 4734–4739.

BOX, G. E., HUNTER, J. S. & HUNTER, W. G. (2005). *Statistics for Experimenters: Design, Innovation, and Discovery*. New York: Wiley-Interscience, 2nd ed.

CAI, T. T., LIU, W. & LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Am. Statist. Assoc.* **106**, 594–607.

CANDÈS, E. J. & TAO, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* **35**, 2313–2351.

CHANDRASEKARAN, V., PARRILO, P. A. & WILLSKY, A. S. (2012). Latent variable graphical model selection via convex optimization. *Ann. Statist.* **40**, 1935–1967.

CHEN, S., WITTEN, D. M. & SHOJAIE, A. (2015). Selection and estimation for mixed graphical models. *Biometrika* **102**, 47–64.

FELLINGHAUER, B., BÜHLMANN, P., RYFFEL, M., VON RHEIN, M. & REINHARDT, J. D. (2013). Stable graphical model estimation with random forests for discrete, continuous, and mixed variables. *Comp. Statist. Data Anal.* **64**, 132–152.

FRIEDMAN, J. H., HASTIE, T. J. & TIBSHIRANI, R. J. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441.

GUO, J., CHENG, J., LEVINA, E., MICHAILIDIS, G. & ZHU, J. (2015). Estimating heterogenous graphical models for discrete data with an application to roll call voting. *Ann. Appl. Statist.* **9**, 821–848.

LEE, J. D. & HASTIE, T. J. (2015). Learning the structure of mixed graphical models. *J. Comp. Graph. Statist.* **25**, 230–253.

LIU, H., HAN, F., YUAN, M., LAFFERTY, J. D. & WASSERMAN, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.* **40**, 2293–2326.

LIU, H., LAFFERTY, J. D. & WASSERMAN, L. (2009). The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* **10**, 2295–2328.

MEINSHAUSEN, N. & BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34**, 1436–1462.

MONTGOMERY, D. C. (2008). *Design and Analysis of Experiments*. New York: John Wiley & Sons, 8th ed.

NING, Y. & LIU, H. (2016). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist.*, in press.

NING, Y., ZHAO, T. & LIU, H. (2016). A likelihood ratio framework for high dimensional semiparametric regression. *Ann. Statist.*, in press.

PENG, J., WANG, P., ZHOU, N. & ZHU, J. (2009). Partial correlation estimation by joint sparse regression models. *J. Am. Statist. Assoc.* **104**, 735–746.

POWER, J. D., COHEN, A. L., NELSON, S. M. et al. (2011). Functional network organization of the human brain. *Neuron* **72**, 665–678.

QIU, H., HAN, F., LIU, H. & CAFFO, B. (2016). Joint estimation of multiple graphical models from high dimensional time series. *J. R. Statist. Soc. B* **78**, 487–504.

RAVIKUMAR, P., WAINWRIGHT, M. J. & LAFFERTY, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.* **38**, 1287–1319.

RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. & YU, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Statist.* **5**, 935–980.

ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. & ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Statist.* **2**, 494–515.

SUN, T. & ZHANG, C.-H. (2013). Sparse matrix inversion with scaled lasso. *J. Mach. Learn. Res.* **14**, 3385–3418.

VAN DE GEER, S. A. & BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the lasso. *Electron. J. Statist.* **3**, 1360–1392.

VOORMAN, A., SHOJAIE, A. & WITTEN, D. M. (2014). Graph estimation with joint additive models. *Biometrika* **101**, 85–101.

YANG, E., RAVIKUMAR, P., ALLEN, G. I. & LIU, Z. (2015). Graphical models via univariate exponential family distributions. *J. Mach. Learn. Res.* **16**, 3813–3847.

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

[*Received M Y. Revised M Y*]

Biometrika (2016), **xx**, x, *pp*. 1–28
Printed in Great Britain

Supplementary material for 'Replicates in high dimensions, with applications to latent variable graphical models

BY KEAN MING TAN

Department of Operations Research and Financial Engineering, Princeton University, Princeton, New Jersey 08544, U.S.A.
kmtan@princeton.edu

YANG NING

Department of Statistical Science, Cornell University, Ithaca, New York 14853, U.S.A.
yn265@cornell.edu

DANIELA M. WITTEN

Department of Statistics, University of Washington, Seattle, Washington 98195, U.S.A.
dwitten@uw.edu

AND HAN LIU

Department of Operations Research and Financial Engineering, Princeton University, Princeton, New Jersey 08544, U.S.A.
hanliu@princeton.edu

S1. LIST OF NOTATION

In this section, we define some notation that will be used throughout this supplementary material. Recall from § 3·2 in the main paper that $\beta^*_{j,O\setminus j}$ is the vector of true parameter values of interest, $\mathcal{S}_j$ is the support set of $\beta^*_{j,O\setminus j}$, and $s_j = |\mathcal{S}_j|$. Let $s_{\max} = \max_{j\in O} s_j$. Let

$$R^{rr'}_{ij}(\beta_{j,O\setminus j}) = \exp\big\{-(x^r_{ij} - x^{r'}_{ij})\beta^{\mathrm{T}}_{j,O\setminus j}(x^r_{i,O\setminus j} - x^{r'}_{i,O\setminus j})\big\}.$$

The gradient of $\ell_j(\beta_{j,O\setminus j})$ is

$$\begin{aligned}\nabla\ell_j(\beta_{j,O\setminus j}) &= -\frac{1}{n}\sum_{i=1}^n\left\{\frac{2}{R(R-1)}\sum_{1\leqslant r<r'\leqslant R}\frac{R^{rr'}_{ij}(\beta_{j,O\setminus j})(x^r_{ij}-x^{r'}_{ij})(x^r_{i,O\setminus j}-x^{r'}_{i,O\setminus j})}{1+R^{rr'}_{ij}(\beta_{j,O\setminus j})}\right\}\\ &= \frac{1}{n}\sum_{i=1}^n U^j_i(\beta_{j,O\setminus j}) \in \mathbb{R}^{p-1}, \qquad \text{(S1)}\end{aligned}$$

where $U_1^j(\beta_{j,O\setminus j}), \ldots, U_n^j(\beta_{j,O\setminus j})$ are independent and identically distributed random variables. The Hessian of $\ell_j(\beta_{j,O\setminus j})$ is

$$\nabla^2 \ell_j(\beta_{j,O\setminus j}) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{2}{R(R-1)}\sum_{1\leqslant r<r'\leqslant R}\frac{R_{ij}^{rr'}(\beta_{j,O\setminus j})(x_{ij}^r - x_{ij}^{r'})^2(x_{i,O\setminus j}^r - x_{i,O\setminus j}^{r'})^{\otimes 2}}{\{1+R_{ij}^{rr'}(\beta_{j,O\setminus j})\}^2}\right] \in \mathbb{R}^{(p-1)\times(p-1)}.$$

For notational convenience, we write

$$h_{ij}^{rr'}(\beta_{j,O\setminus j}) = -\frac{R_{ij}^{rr'}(\beta_{j,O\setminus j})(x_{ij}^r - x_{ij}^{r'})(x_{i,O\setminus j}^r - x_{i,O\setminus j}^{r'})}{1+R_{ij}^{rr'}(\beta_{j,O\setminus j})} \in \mathbb{R}^{p-1} \tag{S2}$$

and, for $k \in O \setminus j$,

$$h_{ijk}^{rr'}(\beta_{j,O\setminus j}) = -\frac{R_{ij}^{rr'}(\beta_{j,O\setminus j})(x_{ij}^r - x_{ij}^{r'})(x_{ik}^r - x_{ik}^{r'})}{1+R_{ij}^{rr'}(\beta_{j,O\setminus j})} \in \mathbb{R}. \tag{S3}$$

Similarly, we let

$$T_{ij}^{rr'}(\beta_{j,O\setminus j}) = \frac{R_{ij}^{rr'}(\beta_{j,O\setminus j})(x_{ij}^r - x_{ij}^{r'})^2(x_{i,O\setminus j}^r - x_{i,O\setminus j}^{r'})^{\otimes 2}}{\{1+R_{ij}^{rr'}(\beta_{j,O\setminus j})\}^2} \in \mathbb{R}^{(p-1)\times(p-1)} \tag{S4}$$

and, for $k, l \in O \setminus j$,

$$T_{ijkl}^{rr'}(\beta_{j,O\setminus j}) = \frac{R_{ij}^{rr'}(\beta_{j,O\setminus j})(x_{ij}^r - x_{ij}^{r'})^2(x_{ik}^r - x_{ik}^{r'})(x_{il}^r - x_{il}^{r'})}{\{1+R_{ij}^{rr'}(\beta_{j,O\setminus j})\}^2} \in \mathbb{R}. \tag{S5}$$

Therefore, the gradient and Hessian of the loss function can also be written as

$$\nabla \ell_j(\beta_{j,O\setminus j}) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{2}{R(R-1)}\sum_{1\leqslant r<r'\leqslant R} h_{ij}^{rr'}(\beta_{j,O\setminus j})\right\} \in \mathbb{R}^{p-1} \tag{S6}$$

and

$$\nabla^2 \ell_j(\beta_{j,O\setminus j}) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{2}{R(R-1)}\sum_{1\leqslant r<r'\leqslant R} T_{ij}^{rr'}(\beta_{j,O\setminus j})\right\} \in \mathbb{R}^{(p-1)\times(p-1)}. \tag{S7}$$

S2. SCALING ASSUMPTIONS IN THEOREM 3

We state two assumptions on the scaling of n and p and on the magnitude of the regularization parameters λ and λ_w in (7) and (10), respectively. The following assumption is needed to show the asymptotic normality of (9).

Assumption S1. Let M be as defined in Theorem 1 and let $w_0 = \max_{j,k\in O}\|w_{jk}^*\|_1$. Furthermore, let $s'_{jk} = \|w_{jk}^*\|_0$ and $s'_{\max} = \max_{j,k\in O} s'_{jk}$. Assume that $Ms_{\max}\lambda/\kappa_{\min}^2 = o(1)$, $s'_{\max}\lambda_w = o(1)$,

$$\lambda_w = \Omega\left\{w_0\left(\frac{Ms_{\max}\lambda}{\kappa_{\min}^2} + \lambda\log^2 p\right)\right\}, \qquad \lim_{n\to\infty} n^{1/2}\left(\frac{s_{\max}\lambda_w\lambda}{\kappa_{\min}^2}\right) = 0$$

and

$$\lim_{n\to\infty} n^{1/2} s'_{\max} \lambda_w \lambda = 0.$$

Next, we state an additional scaling assumption to guarantee that $\hat{\sigma}_{jk}$ in (12) is a consistent estimator of its asymptotic variance.

Assumption S2. For any $j, k \in O$, assume that

$$(1 + w_0 + w_0^2)\left(\frac{s_{\max}\lambda \log^6 p}{\kappa_{\min}^2}\right) = o(1), \qquad w_0 s'_{\max}\lambda_w = o(1).$$

S3. Proof of the results in § 2·2

S3·1. *Proof of Proposition* 1

Under Assumption 1, we obtain

$$\begin{aligned}
&\mathrm{pr}\big\{X_{ij}^r = x_{ij}^r, X_{ij}^{r'} = x_{ij}^{r'} \mid x_{i,O\setminus j}^r, x_{i,O\setminus j}^{r'}, x_{iH}, x_{ij}^{(r,r')}\big\} \\
&= \mathrm{pr}(X_{ij}^r = x_{ij}^r \mid x_{i,O\setminus j}^r, x_{iH})\,\mathrm{pr}(X_{ij}^{r'} = x_{ij}^{r'} \mid x_{i,O\setminus j}^{r'}, x_{iH}) \Big/ \\
&\quad \Big\{\mathrm{pr}(X_{ij}^r = x_{ij}^r \mid x_{i,O\setminus j}^r, x_{iH})\,\mathrm{pr}(X_{ij}^{r'} = x_{ij}^{r'} \mid x_{i,O\setminus j}^{r'}, x_{iH}) \\
&\quad + \mathrm{pr}(X_{ij}^r = x_{ij}^{r'} \mid x_{i,O\setminus j}^r, x_{iH})\,\mathrm{pr}(X_{ij}^{r'} = x_{ij}^r \mid x_{i,O\setminus j}^{r'}, x_{iH})\Big\} \\
&= \left\{1 + \frac{\mathrm{pr}(X_{ij}^r = x_{ij}^{r'} \mid x_{i,O\setminus j}^r, x_{iH})\,\mathrm{pr}(X_{ij}^{r'} = x_{ij}^r \mid x_{i,O\setminus j}^{r'}, x_{iH})}{\mathrm{pr}(X_{ij}^r = x_{ij}^r \mid x_{i,O\setminus j}^r, x_{iH})\,\mathrm{pr}(X_{ij}^{r'} = x_{ij}^{r'} \mid x_{i,O\setminus j}^{r'}, x_{iH})}\right\}^{-1} \\
&= \left[1 + \exp\big\{-(x_{ij}^r - x_{ij}^{r'})\beta_{j,O\setminus j}^{\mathrm{T}}(x_{i,O\setminus j}^r - x_{i,O\setminus j}^{r'})\big\}\right]^{-1},
\end{aligned}$$

where the first equality follows from Assumption 2, that the replicates are mutually independent given the latent variables.

S3·2. *Proof of Proposition* 2

Recall from § 2·3 that $x_{ij}^{(r,r')} = \{\min(x_{ij}^r, x_{ij}^{r'}), \max(x_{ij}^r, x_{ij}^{r'})\}$ are the order statistics of a pair of replicates for the ith subject. Let $X_{ij}^{(r,r')} = \{\min(X_{ij}^r, X_{ij}^{r'}),\ \max(X_{ij}^r, X_{ij}^{r'})\}$. For notational convenience, we let $\mathcal{B}_{ij}^{rr'}$ denote the event $\{X_{i,-j}^r = x_{i,-j}^r,\ X_{i,-j}^{r'} = x_{i,-j}^{r'},\ X_{ij}^{(r,r')} = x_{ij}^{(r,r')}\}$. By Proposition 1, one can see that the conditional distribution of X_{ij}^r and $X_{ij}^{r'}$ given $\mathcal{B}_{ij}^{rr'}$ is binomial with

$$\mathrm{pr}\big(X_{ij}^r = x_{ij}^r, X_{ij}^{r'} = x_{ij}^{r'} \mid \mathcal{B}_{ij}^{rr'}\big) = \frac{1}{1 + \exp\big\{-(x_{ij}^r - x_{ij}^{r'})\beta_{j,O\setminus j}^{\mathrm{T}}(x_{i,O\setminus j}^r - x_{i,O\setminus j}^{r'})\big\}} \tag{S8}$$

and

$$\mathrm{pr}\big(X_{ij}^r = x_{ij}^{r'}, X_{ij}^{r'} = x_{ij}^r \mid \mathcal{B}_{ij}^{rr'}\big) = \frac{\exp\big\{-(x_{ij}^r - x_{ij}^{r'})\beta_{j,O\setminus j}^{\mathrm{T}}(x_{i,O\setminus j}^r - x_{i,O\setminus j}^{r'})\big\}}{1 + \exp\big\{-(x_{ij}^r - x_{ij}^{r'})\beta_{j,O\setminus j}^{\mathrm{T}}(x_{i,O\setminus j}^r - x_{i,O\setminus j}^{r'})\big\}}. \tag{S9}$$

Recall from (S2) that

$$h_{ij}^{rr'}(\beta_{j,O\setminus j}) = -\frac{R_{ij}^{rr'}(\beta_{j,O\setminus j})(x_{ij}^r - x_{ij}^{r'})(x_{i,O\setminus j}^r - x_{i,O\setminus j}^{r'})}{1 + R_{ij}^{rr'}(\beta_{j,O\setminus j})}$$
$$= -\frac{\exp\{-(x_{ij}^r - x_{ij}^{r'})\beta_{j,O\setminus j}^{\mathrm{T}}(x_{i,O\setminus j}^r - x_{i,O\setminus j}^{r'})\}(x_{ij}^r - x_{ij}^{r'})(x_{i,O\setminus j}^r - x_{i,O\setminus j}^{r'})}{1 + \exp\{-(x_{ij}^r - x_{ij}^{r'})\beta_{j,O\setminus j}^{\mathrm{T}}(x_{i,O\setminus j}^r - x_{i,O\setminus j}^{r'})\}}.$$

The conditional expectation of $h_{ij}^{rr'}(\beta_{j,O\setminus j}^*)$ given $\mathcal{B}_{ij}^{rr'}$ takes the form

$$\begin{aligned}
&E\{h_{ij}^{rr'}(\beta_{j,O\setminus j}^*) \mid \mathcal{B}_{ij}^{rr'}\} \\
&= E\left[-\frac{\exp\{-(X_{ij}^r - X_{ij}^{r'})\beta_{j,O\setminus j}^{\mathrm{T}}(x_{i,O\setminus j}^r - x_{i,O\setminus j}^{r'})\}(X_{ij}^r - X_{ij}^{r'})(x_{i,O\setminus j}^r - x_{i,O\setminus j}^{r'})}{1 + \exp\{-(X_{ij}^r - X_{ij}^{r'})\beta_{j,O\setminus j}^{\mathrm{T}}(x_{i,O\setminus j}^r - x_{i,O\setminus j}^{r'})\}} \,\middle|\, \mathcal{B}_{ij}^{rr'}\right] \\
&= -\frac{\exp\{-(x_{ij}^r - x_{ij}^{r'})\beta_{j,O\setminus j}^{\mathrm{T}}(x_{i,O\setminus j}^r - x_{i,O\setminus j}^{r'})\}(x_{ij}^r - x_{ij}^{r'})(x_{i,O\setminus j}^r - x_{i,O\setminus j}^{r'})}{1 + \exp\{-(x_{ij}^r - x_{ij}^{r'})\beta_{j,O\setminus j}^{\mathrm{T}}(x_{i,O\setminus j}^r - x_{i,O\setminus j}^{r'})\}} \mathrm{pr}(X_{ij}^r = x_{ij}^r, X_{ij}^{r'} = x_{ij}^{r'} \mid \mathcal{B}_{ij}^{rr'}) \\
&\quad - \frac{\exp\{-(x_{ij}^{r'} - x_{ij}^r)\beta_{j,O\setminus j}^{\mathrm{T}}(x_{i,O\setminus j}^r - x_{i,O\setminus j}^{r'})\}(x_{ij}^{r'} - x_{ij}^r)(x_{i,O\setminus j}^r - x_{i,O\setminus j}^{r'})}{1 + \exp\{-(x_{ij}^{r'} - x_{ij}^r)\beta_{j,O\setminus j}^{\mathrm{T}}(x_{i,O\setminus j}^r - x_{i,O\setminus j}^{r'})\}} \mathrm{pr}(X_{ij}^r = x_{ij}^{r'}, X_{ij}^{r'} = x_{ij}^r \mid \mathcal{B}_{ij}^{rr'}) \\
&= -\frac{\exp\{-(x_{ij}^r - x_{ij}^{r'})\beta_{j,O\setminus j}^{\mathrm{T}}(x_{i,O\setminus j}^r - x_{i,O\setminus j}^{r'})\}(x_{ij}^r - x_{ij}^{r'})(x_{i,O\setminus j}^r - x_{i,O\setminus j}^{r'})}{1 + \exp\{-(x_{ij}^r - x_{ij}^{r'})\beta_{j,O\setminus j}^{\mathrm{T}}(x_{i,O\setminus j}^r - x_{i,O\setminus j}^{r'})\}} \mathrm{pr}(X_{ij}^r = x_{ij}^r, X_{ij}^{r'} = x_{ij}^{r'} \mid \mathcal{B}_{ij}^{rr'}) \\
&\quad + \frac{(x_{ij}^r - x_{ij}^{r'})(x_{i,O\setminus j}^r - x_{i,O\setminus j}^{r'})}{1 + \exp\{-(x_{ij}^r - x_{ij}^{r'})\beta_{j,O\setminus j}^{\mathrm{T}}(x_{i,O\setminus j}^r - x_{i,O\setminus j}^{r'})\}} \mathrm{pr}(X_{ij}^r = x_{ij}^{r'}, X_{ij}^{r'} = x_{ij}^r \mid \mathcal{B}_{ij}^{rr'}).
\end{aligned}$$

Substituting (S8) and (S9) into the last expression yields $E\{h_{ij}^{rr'}(\beta_{j,O\setminus j}^*) \mid \mathcal{B}_{ij}^{rr'}\} = 0$. By the law of iterated expectation, we obtain $E\{h_{ij}^{rr'}(\beta_{j,O\setminus j}^*)\} = 0$.

To show $E\{\nabla \ell_j(\beta_{j,O\setminus j}^*)\} = 0$, we simply recall from (S6) that

$$\nabla \ell_j(\beta_{j,O\setminus j}^*) = \frac{1}{n}\sum_{i=1}^n \left\{ \frac{2}{R(R-1)} \sum_{r<r'} h_{ij}^{rr'}(\beta_{j,O\setminus j}^*) \right\}$$

and use the fact that $E\{h_{ij}^{rr'}(\beta_{j,O\setminus j}^*)\} = 0$.

Next, we show that $\beta_{j,O\setminus j}^*$ is a global minimizer of $E\{\ell_j(\beta_{j,O\setminus j})\}$. Each component of $h_{ij}^{rr'}(\beta_{j,O\setminus j})$ is dominated by the corresponding one of $|(x_{ij}^r - x_{ij}^{r'})(x_{i,O\setminus j}^r - x_{i,O\setminus j}^{r'})|$, and the latter is integrable by Proposition S1 in § S4·1. Applying the dominated convergence theorem, we can interchange the order of integration and differentiation. Thus, we obtain $\nabla E\{\ell_j(\beta_{j,O\setminus j})\}|_{\beta_{j,O\setminus j}=\beta_{j,O\setminus j}^*} = E\{\nabla \ell_j(\beta_{j,O\setminus j}^*)\} = 0$. Hence $\beta_{j,O\setminus j}^*$ is a stationary point of $E\{\ell_j(\beta_{j,O\setminus j})\}$. Following a similar argument, we have

$$\nabla^2 E\{\ell_j(\beta_{j,O\setminus j})\} = E\left[\frac{R_{ij}^{rr'}(\beta_{j,O\setminus j})(X_{ij}^r - X_{ij}^{r'})^2(X_{i,O\setminus j}^r - X_{i,O\setminus j}^{r'})^{\otimes 2}}{\{1 + R_{ij}^{rr'}(\beta_{j,O\setminus j})\}^2}\right] \succeq 0.$$

Therefore $E\{\ell_j(\beta_{j,O\setminus j})\}$ is convex. It follows that $\beta_{j,O\setminus j}^*$ is a global minimizer of $E\{\ell_j(\beta_{j,O\setminus j})\}$.

S4. PROOF OF THEOREMS 1 AND 2

S4·1. *Some technical lemmas*

Assumption 3 allows us to control the tail behaviour of the random variables. Given Assumption 3, we have the following proposition on the tail probability of the random variables.

PROPOSITION S1. *Under Assumption* 3*, for any* $t > 0$ *and* $j \in O$ *we have*

$$\mathrm{pr}\,(|X_j| \geqslant t) \leqslant c_1 \exp(-t),$$

where $c_1 = 2\exp(\kappa_m + \kappa_h/2)$*. Moreover, for two replicates of* X_j $(j \in O)$*, which we denote by* X_j^r *and* $X_j^{r'}$*, we have*

$$\mathrm{pr}\big(|X_j^r - X_j^{r'}| \geqslant t\big) \leqslant 2c_1 \exp(-t/2).$$

Also,

$$E(X_j^4) = \int \mathrm{pr}(X_j^4 \geqslant t)\,\mathrm{d}t \leqslant \int c_1 \exp(-t^{1/4})\,\mathrm{d}t = 24\,c_1$$

and

$$E(X_j^8) = \int \mathrm{pr}(X_j^8 \geqslant t)\,\mathrm{d}t \leqslant \int c_1 \exp(-t^{1/8})\,\mathrm{d}t = 40\,320\,c_1.$$

The proof of Proposition S1 involves the standard Chernoff bounding technique (see Proposition 3 in Yang et al., 2015). We now present a collection of lemmas that will be used to prove Theorems 1 and 2. The proofs of Lemmas S1–S5 are provided in § S6.

Recall from (S1) that the gradient of the loss function can be written as the average of independent and identically distributed random variables $\nabla \ell_j(\beta^*_{j,O\setminus j}) = \sum_{i=1}^n U_i^j(\beta^*_{j,O\setminus j})/n$, and that

$$U_{ik}^j(\beta^*_{j,O\setminus j}) = \frac{2}{R(R-1)} \sum_{1\leqslant r<r'\leqslant R} h_{ijk}^{rr'}(\beta^*_{j,O\setminus j}) \in \mathbb{R}. \tag{S10}$$

The following lemma shows that $U_i^j(\beta^*_{j,O\setminus j})$ is a random variable with general exponential tail.

LEMMA S1. *For* $j \in O$ *and* $k \in O \setminus j$*, we have*

$$\mathrm{pr}\big\{U_{ik}^j(\beta^*_{j,O\setminus j}) \geqslant t\big\} \leqslant 4c_1 R^2 \exp(-0{\cdot}5t^{1/2}),$$

where c_1 *is the constant in Assumption* 3*.*

The following lemma establishes an upper bound for the gradient of the loss function $\|\nabla \ell_j(\beta^*_{j,O\setminus j})\|_\infty$ with high probability.

LEMMA S2. *Under Assumption* 3*, let*

$$\lambda = \frac{(\xi+1)K}{\xi-1}\left(\frac{\log^5 p}{n}\right)^{1/2},$$

where $K > 0$ *and* $\xi > 1$*. Then, for* $j \in O$*,*

$$\Big\|\nabla \ell_j(\beta^*_{j,O\setminus j})\Big\|_\infty \leqslant \frac{(\xi-1)\lambda}{\xi+1}$$

with probability at least $1 - p^{-1}$*.*

Now, let $D(\hat{\beta}_{j,O\setminus j}, \beta_{j,O\setminus j}) = (\hat{\beta}_{j,O\setminus j} - \beta_{j,O\setminus j})^{\mathrm{T}}\{\nabla\ell_j(\hat{\beta}_{j,O\setminus j}) - \nabla\ell_j(\beta_{j,O\setminus j})\}$ be the symmetrized Bregman divergence of $\ell_j(\beta_{j,O\setminus j})$. The next lemma allows us to control the variability of the Bregman divergence and the Hessian matrix in a neighbourhood of $\beta_{j,O\setminus j}$.

LEMMA S3. *Let* $\Delta \in \mathbb{R}^{p-1}$ *and*

$$b = \max_{1\leqslant i\leqslant n} \max_{1\leqslant r<r'\leqslant R} \left|(x_{ij}^r - x_{ij}^{r'})\Delta^{\mathrm{T}}(x_{i,O\setminus j}^r - x_{i,O\setminus j}^{r'})\right|.$$

Then, for $j \in O$,

$$\exp(-b)\nabla^2\ell_j(\beta_{j,O\setminus j}) \leqslant \nabla^2\ell_j(\beta_{j,O\setminus j} + \Delta) \leqslant \exp(b)\nabla^2\ell_j(\beta_{j,O\setminus j})$$

and

$$\exp(-b)\Delta^{\mathrm{T}}\nabla^2\ell_j(\beta_{j,O\setminus j})\Delta \leqslant D(\beta_{j,O\setminus j} + \Delta, \beta_{j,O\setminus j}) \leqslant \exp(b)\Delta^{\mathrm{T}}\nabla^2\ell_j(\beta_{j,O\setminus j})\Delta.$$

We omit the proof, but the result follows from arguments similar to those in Ning et al. (2016).

Given Lemma S3, we have the following inequality.

LEMMA S4. *Let* $\hat{\Delta}_j = \hat{\beta}_{j,O\setminus j} - \beta^*_{j,O\setminus j}$. *Then, for* $j \in O$,

$$D\big(\hat{\beta}_{j,O\setminus j}, \beta^*_{j,O\setminus j}\big) + \big\{\lambda - \big\|\nabla\ell_j(\beta^*_{j,O\setminus j})\big\|_\infty\big\}\big\|\hat{\Delta}_{j,\mathcal{S}_j^{\mathrm{c}}}\big\|_1 \leqslant \big\{\lambda + \big\|\nabla\ell_j(\beta^*_{j,O\setminus j})\big\|_\infty\big\}\big\|\hat{\Delta}_{j,\mathcal{S}_j}\big\|_1.$$

Moreover, for any $\xi > 1$, $\|\hat{\Delta}_{j,\mathcal{S}_j^{\mathrm{c}}}\|_1 \leqslant \xi\|\hat{\Delta}_{j,\mathcal{S}_j}\|_1$, *provided that* $\|\nabla\ell_j(\beta^*_{j,O\setminus j})\|_\infty \leqslant \lambda(\xi - 1)/(\xi + 1)$.

Recall from § 2·4 that $H^j = E\{\nabla^2\ell_j(\beta^*_{j,O\setminus j})\}$. We next present a lemma on the deviation between the empirical Hessian matrix and the population Hessian matrix.

LEMMA S5. *Under Assumption* 3, *for* $j \in O$ *there exists a constant* $K_1 > 0$ *such that*

$$\left\|\nabla^2\ell_j(\beta^*_{j,O\setminus j}) - H^j\right\|_\infty \leqslant K_1\big(\log^9 p/n\big)^{1/2}$$

with probability at least $1 - p^{-1}$.

S4·2. *Proof of Theorem* 1

To obtain an upper bound for the estimation error, we state some conditions on the minimal eigenvalues of the Hessian matrix of the loss function. Given the constant $\xi > 1$, for each $j \in O$ we define the compatibility factor, restricted eigenvalue and weak cone invertibility factor as

$$\kappa^2\big\{\nabla^2\ell_j(\beta^*_{j,O\setminus j}), s_j\big\} = \min_v \left\{\frac{s_j v^{\mathrm{T}}\nabla^2\ell_j(\beta^*_{j,O\setminus j})v}{\|v_{S_j}\|_1^2} : v \in \mathbb{R}^{p-1}, v \neq 0, \|v_{S_j^{\mathrm{c}}}\|_1 \leqslant \xi\|v_{S_j}\|_1\right\}, \tag{S11}$$

$$\mathrm{RE}\big\{\nabla^2\ell_j(\beta^*_{j,O\setminus j}), s_j\big\} = \min_v \left\{\frac{v^{\mathrm{T}}\nabla^2\ell_j(\beta^*_{j,O\setminus j})v}{\|v\|_2^2} : v \in \mathbb{R}^{p-1}, v \neq 0, \|v_{S_j^{\mathrm{c}}}\|_1 \leqslant \xi\|v_{S_j}\|_1\right\} \tag{S12}$$

and, for $q \geqslant 1$,

$$\rho_q\big\{\nabla^2\ell_j(\beta^*_{j,O\setminus j}), s_j\big\} = \min_v \left\{\frac{s_j^{1/q} v^{\mathrm{T}}\nabla^2\ell_j(\beta^*_{j,O\setminus j})v}{\|v_{S_j}\|_1\|v\|_q} : v \in \mathbb{R}^{p-1}, v \neq 0, \|v_{S_j^{\mathrm{c}}}\|_1 \leqslant \xi\|v_{S_j}\|_1\right\}. \tag{S13}$$

These conditions have commonly been used to establish upper bounds for estimation error in the context of ℓ_1-penalized regression (Bickel et al., 2009; van de Geer & Bühlmann, 2009; Ye & Zhang, 2010). Define $\kappa^2_{\min} = \min_{j\in O}\, \kappa^2\{\nabla^2\ell_j(\beta^*_{j,O\setminus j}), s_j\}$, $\mathrm{RE}_{\min} = \min_{j\in O}\, \mathrm{RE}\{\nabla^2\ell_j(\beta^*_{j,O\setminus j}), s_j\}$ and $\rho_{q,\min} = \min_{j\in O}\, \rho_q\{\nabla^2\ell_j(\beta^*_{j,O\setminus j}), s_j\}$. We now give the proof of Theorem 1.

Proof of Theorem 1. The proof involves obtaining an upper bound on the gradient of the loss function. By Lemma S2 we have that $\|\nabla\ell_j(\beta^*_{j,O\setminus j})\|_\infty \leqslant \lambda(\xi-1)/(\xi+1)$ with probability at least $1-p^{-1}$. Throughout the proof, we conditioned on the event $\|\nabla\ell_j(\beta^*_{j,O\setminus j})\|_\infty \leqslant \lambda(\xi-1)/(\xi+1)$.

Let $\hat{\Delta}_j = \hat{\beta}_{j,O\setminus j} - \beta^*_{j,O\setminus j}$, $a = \hat{\Delta}_j/\|\hat{\Delta}_j\|_1$ and $k = \|\hat{\Delta}_j\|_1$. Recall from Lemma S3 that

$$b = \max_{1\leqslant i\leqslant n}\, \max_{1\leqslant r<r'\leqslant R} \left|(x^r_{ij} - x^{r'}_{ij})\hat{\Delta}_j^{\mathrm{T}}(x^r_{i,O\setminus j} - x^{r'}_{i,O\setminus j})\right|.$$

Consider the quantity $D(\hat{\beta}_{j,O\setminus j}, \beta^*_{j,O\setminus j})$. By Lemma S3,

$$D\big(\beta^*_{j,O\setminus j} + ka, \beta^*_{j,O\setminus j}\big) \geqslant \exp(-b)k^2 a^{\mathrm{T}}\nabla^2\ell_j(\beta^*_{j,O\setminus j})a \geqslant k^2\exp(-Mk)a^{\mathrm{T}}\nabla^2\ell_j(\beta^*_{j,O\setminus j})a, \tag{S14}$$

since

$$\begin{aligned} b &= \max_{1\leqslant i\leqslant n}\, \max_{1\leqslant r<r'\leqslant R} \left|(x^r_{ij} - x^{r'}_{ij})\hat{\Delta}_j^{\mathrm{T}}(x^r_{i,O\setminus j} - x^{r'}_{i,O\setminus j})\right| \\ &\leqslant \max_{1\leqslant i\leqslant n}\, \max_{1\leqslant r<r'\leqslant R} \left\|\big(x^r_{ij} - x^{r'}_{ij}\big)\big(x^r_{i,O\setminus j} - x^{r'}_{i,O\setminus j}\big)\right\|_\infty \|\hat{\Delta}_j\|_1 \\ &\leqslant Mk, \end{aligned}$$

where the last inequality holds conditioning on the event

$$\mathcal{A} = \left\{\max_{1\leqslant i\leqslant n}\, \max_{1\leqslant r<r'\leqslant R} \left\|\big(x^r_{ij} - x^{r'}_{ij}\big)\big(x^r_{i,O\setminus j} - x^{r'}_{i,O\setminus j}\big)\right\|_\infty \leqslant M\right\}.$$

By the definition of the compatibility factor, (S11), we have

$$a^{\mathrm{T}}\nabla^2\ell_j(\beta^*_{j,O\setminus j})a \geqslant \kappa^2\{\nabla^2\ell_j(\beta^*_{j,O\setminus j}), s_j\}\|a_{\mathcal{S}_j}\|_1^2/s_j.$$

Substituting this into (S14) gives

$$D\big(\beta^*_{j,O\setminus j} + ka, \beta^*_{j,O\setminus j}\big) \geqslant k^2\exp(-Mk)\kappa^2\{\nabla^2\ell_j(\beta^*_{j,O\setminus j}), s_j\}\|a_{\mathcal{S}_j}\|_1^2/s_j. \tag{S15}$$

Next, we derive an upper bound for $D(\beta^*_{j,O\setminus j} + ka, \beta^*_{j,O\setminus j})$. By Lemma S4, conditioning on the event $\|\nabla\ell_j(\beta^*_{j,O\setminus j})\|_\infty \leqslant \lambda(\xi-1)/(\xi+1)$, we have $\|\hat{\Delta}_{j,S_j^{\mathrm{c}}}\| \leqslant \xi\|\hat{\Delta}_{j,S_j}\|$. Hence,

$$\begin{aligned} D\big(\beta^*_{j,O\setminus j} + ka, \beta^*_{j,O\setminus j}\big) &\leqslant \{\lambda + \|\nabla\ell_j(\beta^*_{j,O\setminus j})\|_\infty\}\|\hat{\Delta}_{j,\mathcal{S}_j}\|_1 - \{\lambda - \|\nabla\ell_j(\beta^*_{j,O\setminus j})\|_\infty\}\|\hat{\Delta}_{j,\mathcal{S}_j^{\mathrm{c}}}\|_1 \\ &= k\{\lambda + \|\nabla\ell_j(\beta^*_{j,O\setminus j})\|_\infty\}\|a_{\mathcal{S}_j}\|_1 - k\{\lambda - \|\nabla\ell_j(\beta^*_{j,O\setminus j})\|_\infty\}\|a_{\mathcal{S}_j^{\mathrm{c}}}\|_1 \\ &\leqslant \frac{2k\xi\lambda}{\xi+1}\|a_{\mathcal{S}_j}\|_1 - \frac{2k\lambda}{\xi+1}\|a_{\mathcal{S}_j^{\mathrm{c}}}\|_1 + \frac{2k\lambda}{\xi+1}\|a_{\mathcal{S}_j}\|_1 - \frac{2k\lambda}{\xi+1}\|a_{\mathcal{S}_j}\|_1 \\ &= 2k\lambda\|a_{\mathcal{S}_j}\|_1 - \frac{2k\lambda}{\xi+1} \\ &\leqslant \frac{k\lambda(\xi+1)\|a_{\mathcal{S}_j}\|_1^2}{2}, \end{aligned} \tag{S16}$$

where the second equality follows from the fact that $\|a\|_1 = 1$ by construction, and the last inequality is obtained by using the fact that $a^2 + b^2 \geqslant 2ab$ for $a, b \in \mathbb{R}$.

Combining (S15) and (S16), we have

$$k \exp(-Mk) \leqslant \frac{\xi + 1}{2\kappa^2\{\nabla^2 \ell_j(\beta^*_{j,O\setminus j}), s_j\}} \lambda s_j.$$

Let $\tau = M(\xi+1)\lambda s_j / [2\kappa^2\{\nabla^2 \ell_j(\beta^*_{j,O\setminus j}), s_j\}]$. Then we have $Mk \exp(-Mk) \leqslant \tau$. Since η is the smallest solution of $z\exp(-z) = \tau$ by definition, and $z\exp(-z) - \tau$ is an increasing function of z for $z \leqslant 1$, this implies that $Mk \leqslant \eta$. Therefore

$$\|\hat{\beta}_{j,O\setminus j} - \beta^*_{j,O\setminus j}\|_1 = \|\hat{\Delta}_j\|_1 = k \leqslant \frac{\eta}{M} = \frac{\tau \exp(\eta)}{M} = \frac{(\xi+1)\exp(\eta)}{2\kappa^2\{\nabla^2 \ell_j(\beta^*_{j,O\setminus j}), s_j\}} \lambda s_j,$$

which implies $\|\hat{\beta}_{j,O\setminus j} - \beta^*_{j,O\setminus j}\|_1 \leqslant (\xi+1)\exp(\eta) s_{\max}\lambda / (2\kappa^2_{\min})$, where $\kappa^2_{\min} = \min_{j\in O} \kappa^2\{\nabla^2 \ell_j(\beta^*_{j,O\setminus j}), s_j\}$.

To prove $\|\hat{\beta}_{j,O\setminus j} - \beta^*_{j,O\setminus j}\|_2 \leqslant C'(s_{\max})^{1/2}\lambda / \mathrm{RE}_{\min}$, we recall that by the definition of the restricted eigenvalue, (S12), $a^{\mathrm{T}} \nabla^2 \ell_j(\beta^*_{j,O\setminus j}) a \geqslant \mathrm{RE}\{\nabla^2 \ell_j(\beta^*_{j,O\setminus j}), s_j\}\|a\|_2^2$. Thus, by (S14),

$$D\big(\beta^*_{j,O\setminus j} + ka, \beta^*_{j,O\setminus j}\big) \geqslant k^2 \exp(-Mk)\, \mathrm{RE}\{\nabla^2 \ell_j(\beta^*_{j,O\setminus j}), s_j\}\|a\|_2^2. \tag{S17}$$

Similar to (S16), by an application of Lemma S4 it can be shown that

$$D\big(\beta^*_{j,O\setminus j} + ka, \beta^*_{j,O\setminus j}\big) \leqslant \frac{2k\xi\lambda}{1+\xi}\|a_{\mathcal{S}_j}\|_1 \leqslant \frac{2k\xi\lambda}{1+\xi} s_j^{1/2}\|a_{\mathcal{S}_j}\|_2 \leqslant \frac{2k\xi\lambda}{1+\xi} s_j^{1/2}\|a\|_2. \tag{S18}$$

Upon combining (S17) and (S18), we have

$$\|\hat{\beta}_{j,O\setminus j} - \beta^*_{j,O\setminus j}\|_2 = k\|a\|_2 \leqslant \frac{2\xi\exp(\eta)}{(\xi+1)\,\mathrm{RE}\{\nabla^2 \ell_j(\beta^*_{j,O\setminus j}), s_j\}} s_j^{1/2}\lambda \leqslant \frac{2\xi\exp(\eta)}{(\xi+1)\,\mathrm{RE}_{\min}} s_{\max}^{1/2}\lambda,$$

where the first inequality holds because $Mk \leqslant \eta$ and $\mathrm{RE}_{\min} = \min_{j\in O} \mathrm{RE}\{\nabla^2 \ell_j(\beta^*_{j,O\setminus j}), s_j\}$.

To prove $\|\hat{\beta}_{j,O\setminus j} - \beta^*_{j,O\setminus j}\|_q \leqslant C'(s_{\max})^{1/q}\lambda/\rho_{q,\min}$, recall that by the definition of the weak cone invertibility factor, (S13), we have $a^{\mathrm{T}}\nabla^2 \ell_j(\beta^*_{j,O\setminus j})a \geqslant \rho_q\{\nabla^2 \ell_j(\beta^*_{j,O\setminus j}), s_j\}\|a_{\mathcal{S}_j}\|_1\|a\|_q / s_j^{1/q}$. Hence, by (S14),

$$D\big(\beta^*_{j,O\setminus j} + ka, \beta^*_{j,O\setminus j}\big) \geqslant k^2 \exp(-Mk)\rho_q\{\nabla^2 \ell_j(\beta^*_{j,O\setminus j}), s_j\}\|a_{\mathcal{S}_j}\|_1\|a\|_q / s_j^{1/q}. \tag{S19}$$

Moreover, from (S18) we have

$$D\big(\beta^*_{j,O\setminus j} + ka, \beta^*_{j,O\setminus j}\big) \leqslant \frac{2k\xi\lambda}{1+\xi}\|a_{\mathcal{S}_j}\|_1. \tag{S20}$$

Combining (S19) and (S20) yields

$$\big\|\hat{\beta}_{j,O\setminus j} - \beta^*_{j,O\setminus j}\big\|_q = k\|a\|_q \leqslant \frac{2\xi\exp(\eta)}{(\xi+1)\rho_q\{\nabla^2 \ell_j(\beta^*_{j,O\setminus j}), s_j\}} s_j^{1/q}\lambda \leqslant \frac{2\xi\exp(\eta)}{(\xi+1)\rho_{q,\min}} s_{\max}^{1/q}\lambda,$$

where the first inequality follows from the facts that $Mk \leqslant \eta$ and $\rho_{q,\min} = \min_{j\in O} \rho_q\{\nabla^2 \ell_j(\beta^*_{j,O\setminus j}), s_j\}$. □

S4·3. *Proof of Theorem* 2

Let $\mathcal{A} = (v \in \mathbb{R}^{p-1} : \|v_{S_j^c}\|_1 \leqslant \xi \|v_{S_j}\|_1)$. We first show that for any $v \in \mathcal{A}$, the quantity $v^{\mathrm{T}} E\{\nabla^2 \ell_j(\beta^*_{j,O\setminus j})\} v$ can be bounded below by a positive constant. Using the fact that the deviation between the empirical and population Hessian matrices is arbitrarily small when n is sufficiently large (see Lemma S5), we show that $v^{\mathrm{T}} \nabla^2 \ell_j(\beta^*_{j,O\setminus j}) v$ is also bounded below by a constant. Therefore the compatibility factor (S11), the restricted eigenvalue (S12), and the weak cone invertibility factor (S13) can all be bounded below by positive constants.

Lower bound for $v^{\mathrm{T}} E\{\nabla^2 \ell_j(\beta^*_{j,O\setminus j})\} v$. We show this via a truncation argument. For $\delta, S_1, S_2 > 0$ and $j \in O$, we define the events

$$
\begin{aligned}
F_{ij}^{rr'} &= \big(|x_{ij}^r| \leqslant \delta\big) \cap \big(|x_{ij}^{r'}| \leqslant \delta\big), \\
G_{ij}^r &= \big\{\big|(\beta^*_{j,O\setminus j})^{\mathrm{T}} x^r_{i,O\setminus j}\big| \leqslant S_1\big\}, \qquad G_{ij}^{r'} = \big\{\big|(\beta^*_{j,O\setminus j})^{\mathrm{T}} x^{r'}_{i,O\setminus j}\big| \leqslant S_1\big\}, \\
H_{ij}^r &= \big\{\big|(\beta^*_{j,-j})^{\mathrm{T}} x^r_{i,-j}\big| \leqslant S_2\big\}, \qquad H_{ij}^{r'} = \big\{\big|(\beta^*_{j,-j})^{\mathrm{T}} x^{r'}_{i,-j}\big| \leqslant S_2\big\}.
\end{aligned}
$$

Since $\exp(-z)/\{1+\exp(-z)\}^2$ is a decreasing function of z for $z \geqslant 0$, the quantity $R_{ij}^{rr'}(\beta^*_{j,O\setminus j})/\{1 + R_{ij}^{rr'}(\beta^*_{j,O\setminus j})\}^2$ can be bounded below by $C_1 = \exp(-4S_1\delta)/\{1+\exp(-4S_1\delta)\}^2$. Recall from (S4) and (S7) the definitions of $T_{ij}^{rr'}$ and $\nabla^2 \ell_j(\beta^*_{j,O\setminus j})$. We have

$$
\begin{aligned}
&\nabla^2 \ell_j(\beta^*_{j,O\setminus j}) \\
&\quad \geqslant \frac{1}{n}\sum_{i=1}^n \frac{2}{R(R-1)} \sum_{1\leqslant r<r'\leqslant R} T_{ij}^{rr'} I(F_{ij}^{rr'}) I(G_{ij}^r) I(G_{ij}^{r'}) \\
&\quad \geqslant \frac{1}{n}\sum_{i=1}^n \frac{2}{R(R-1)} \sum_{1\leqslant r<r'\leqslant R} C_1 (x_{ij}^r - x_{ij}^{r'})^2 (x^r_{i,O\setminus j} - x^{r'}_{i,O\setminus j})^{\otimes 2} I(F_{ij}^{rr'}) I(G_{ij}^r) I(G_{ij}^{r'}) \\
&\quad = \frac{1}{n}\sum_{i=1}^n W_{ij}.
\end{aligned}
$$

Let

$$
\Sigma = \begin{pmatrix} \Sigma_{O,O} & \Sigma_{O,H} \\ \Sigma_{H,O} & \Sigma_{H,H} \end{pmatrix}
$$

and let $\Theta = \Sigma^{-1}$. Assume that $X_{iH} \sim N(0, \Sigma_{H,H})$ and that $X_{iO}^r \mid X_{iH}$ are independent and identically distributed from $N(\Sigma_{O,H}\Sigma_{H,H}^{-1} X_{iH},\, \Sigma_{O,O} - \Sigma_{O,H}\Sigma_{H,H}^{-1}\Sigma_{H,O})$ for $r = 1, \ldots, R$. Then

$$
\begin{aligned}
\mathsf{p}(x_{ij}^r \mid x_{i,-j}^r, H_{ij}^r) &= \frac{\mathsf{p}(x_i^r \mid H_{ij}^r)}{\int \mathsf{p}(x_i^r \mid H_{ij}^r)\,\mathrm{d}x_{ij}} \\
&= \frac{\mathsf{p}(x_i^r)}{\mathsf{p}(H_{ij}^r)\int \mathsf{p}(x_i^r)/\mathsf{p}(H_{ij}^r)\,\mathrm{d}x_{ij}} \\
&= \mathsf{p}(x_{ij}^r \mid x_{i,-j}^r),
\end{aligned}
$$

where we have used the facts that $\mathsf{p}(x_i^r \mid H_{ij}^r) = \mathsf{p}(x_i^r)/\mathsf{p}(H_{ij}^r)$ and $\mathsf{p}(H_{ij}^r)$ is a constant. Recall from Example 1 that the conditional density of X_{ij}^r given $X_{i,-j}^r$ is

$$
\mathsf{p}(x_{ij}^r \mid x_{i,-j}^r) = (\Theta_{jj}/2\pi)^{1/2} \exp\left\{-\frac{\Theta_{jj}}{2}(x_{ij}^r)^2 - x_{ij}^r \Theta_{j,-j}^{\mathrm{T}} x_{i,-j}^r - \frac{1}{2\Theta_{jj}}\left(\Theta_{j,-j}^{\mathrm{T}} x_{i,-j}^r\right)^2\right\},
$$

where $\beta^*_{j,-j} = -\Theta_{j,-j}$. Since X^r_{ij} and $X^{r'}_{ij}$ are conditionally independent and identically normally distributed random variables, we can bound the following conditional expectation from below:

$$
\begin{aligned}
&E\big\{(X^r_{ij} - X^{r'}_{ij})^2 I(F^{rr'}_{ij}) \mid x^r_{i,-j}, x^{r'}_{i,-j}, H^r_{ij}, H^{r'}_{ij}\big\} \\
&= \Theta_{jj}/2\pi \int_{-\delta}^{\delta}\int_{-\delta}^{\delta} (x^r_{ij} - x^{r'}_{ij})^2 \exp\left\{-\frac{\Theta_{jj}}{2}(x^r_{ij})^2 - x^r_{ij}\Theta^{\mathrm{T}}_{j,-j}x^r_{i,-j} - \frac{1}{2\Theta_{jj}}\left(\Theta^{\mathrm{T}}_{j,-j}x^r_{i,-j}\right)^2\right\} \\
&\qquad\qquad \times \exp\left\{-\frac{\Theta_{jj}}{2}(x^{r'}_{ij})^2 - x^{r'}_{ij}\Theta^{\mathrm{T}}_{j,-j}x^{r'}_{i,-j} - \frac{1}{2\Theta_{jj}}\left(\Theta^{\mathrm{T}}_{j,-j}x^{r'}_{i,-j}\right)^2\right\} \mathrm{d}x^r_{ij}\,\mathrm{d}x^{r'}_{ij} \\
&\geqslant \Theta_{jj}/2\pi \int_{-\delta}^{\delta}\int_{-\delta}^{\delta} (x^r_{ij} - x^{r'}_{ij})^2 \\
&\qquad\qquad \times \exp\left[-\frac{\Theta_{jj}}{2}\left\{(x^r_{ij})^2 + (x^{r'}_{ij})^2\right\} - S_2(x^r_{ij} + x^{r'}_{ij}) - \frac{S_2^2}{2\Theta_{jj}}\right] \mathrm{d}x^r_{ij}\,\mathrm{d}x^{r'}_{ij},
\end{aligned}
$$

where the last inequality follows from the fact that conditioned on the events H^r_{ij} and $H^{r'}_{ij}$, $|(\beta^*_{j,-j})^{\mathrm{T}}x^r_{i,-j}| \leqslant S_2$. For notational convenience, we denote the last expression by C_2.

Therefore, by the law of iterated expectation, we obtain

$$
\begin{aligned}
v^{\mathrm{T}}E\big\{\nabla^2\ell_j(\beta^*_{j,O\setminus j})\big\}v &\geqslant v^{\mathrm{T}}E\left(\frac{1}{n}\sum_{i=1}^{n} W_{ij}\right)v \\
&= v^{\mathrm{T}}E(W_{ij})v \\
&= v^{\mathrm{T}}E\big\{E(W_{ij} \mid x^r_{i,-j}x^{r'}_{i,-j}, H^r_{ij}, H^{r'}_{ij}\ \forall\, r < r')\big\}v \\
&\geqslant C_1C_2E\left[\big\{(X^r_{i,O\setminus j} - X^{r'}_{i,O\setminus j})^{\mathrm{T}}v\big\}^2 I(G^r_{ij})I(G^{r'}_{ij})\right],
\end{aligned}
\tag{S21}
$$

where we have used the fact that the replicates are conditionally independent and identically distributed. We now establish a lower bound for (S21). By the Cauchy–Schwarz inequality, we have

$$
\begin{aligned}
&E\left[\big\{(X^r_{i,O\setminus j} - X^{r'}_{i,O\setminus j})^{\mathrm{T}}v\big\}^2\big\{1 - I(G^r_{ij})I(G^{r'}_{ij})\big\}\right] \\
&\quad\leqslant \left(E\left[\big\{(X^r_{i,O\setminus j} - X^{r'}_{i,O\setminus j})^{\mathrm{T}}v\big\}^4\right]\right)^{1/2}\left[\mathrm{pr}\big\{(G^r_{ij})^{\mathrm{c}} \cup (G^{r'}_{ij})^{\mathrm{c}}\big\}\right]^{1/2}.
\end{aligned}
\tag{S22}
$$

Recall that $X^r_{iO} \mid X_{iH}$ are independent and identically distributed from $N(\Sigma_{O,H}\Sigma^{-1}_{H,H}X_{iH},\, \Sigma_{O,O} - \Sigma_{O,H}\Sigma^{-1}_{H,H}\Sigma_{H,O})$. For notational convenience, we write $\Sigma' = \Sigma_{O,O} - \Sigma_{O,H}\Sigma^{-1}_{H,H}\Sigma_{H,O}$. Since $X^r_{i,O\setminus j} \mid X_{iH} \sim N(\Sigma_{O\setminus j,H}\Sigma^{-1}_{H,H}X_{iH},\, \Sigma'_{O\setminus j,O\setminus j})$, we have that $(X^r_{i,O\setminus j} - X^{r'}_{i,O\setminus j})^{\mathrm{T}}v \mid X_{iH} \sim N(0, 2v^{\mathrm{T}}\Sigma'_{O\setminus j,O\setminus j}v)$. Therefore, the kurtosis of a normal distribution is

$$
\begin{aligned}
E\left[\big\{(X^r_{i,O\setminus j} - X^{r'}_{i,O\setminus j})^{\mathrm{T}}v\big\}^4\right] &= E\left(E\left[\big\{(X^r_{i,O\setminus j} - X^{r'}_{i,O\setminus j})^{\mathrm{T}}v\big\}^4 \mid X_{iH}\right]\right) \\
&= 3(2v^{\mathrm{T}}\Sigma'_{O\setminus j,O\setminus j}v)^2 \\
&\leqslant 12\|v\|_2^4\,\Lambda^2_{\max}(\Sigma),
\end{aligned}
$$

where $\Lambda_{\max}(\Sigma)$ is the largest eigenvalue of Σ. The last inequality is obtained from the fact that $\Lambda_{\max}(\Sigma') \leqslant \Lambda_{\max}(\Sigma)$.

Recall that $X^r_{i,O\backslash j} \sim N(0, \Sigma_{O\backslash j, O\backslash j})$. Therefore $(\beta^*_{j,O\backslash j})^{\mathrm{T}} X^r_{i,O\backslash j} \sim N(0, \sigma_1^2)$, where $\sigma_1^2 = (\beta^*_{j,O\backslash j})^{\mathrm{T}} \Sigma_{O\backslash j, O\backslash j} \beta^*_{j,O\backslash j}$. By the Gaussian tail inequality in Lemma S13, we have

$$\mathrm{pr}\big\{(G^r_{ij})^{\mathrm{c}} \cup (G^{r'}_{ij})^{\mathrm{c}}\big\} \leqslant 2\,\mathrm{pr}\{(G^r_{ij})^{\mathrm{c}}\} = 2\,\mathrm{pr}\big\{\big|(\beta^*_{j,O\backslash j})^{\mathrm{T}} X^r_{i,O\backslash j}\big| > S_1\big\} \leqslant \frac{4\sigma_1}{S_1} \exp(-S_1^2/2\sigma_1^2).$$

We write the last expression as $C_3(S_1)$, indicating its dependence on S_1. Therefore, by (S22) and picking a sufficiently large S_1 such that $(12)^{1/2}\Lambda_{\max}(\Sigma)\{C_3(S_1)\}^{1/2} = \Lambda_{\min}(\Sigma)$, we obtain

$$\begin{aligned} &E\left[\big\{(X^r_{i,O\backslash j} - X^{r'}_{i,O\backslash j})^{\mathrm{T}} v\big\}^2 \big\{1 - I(G^r_{ij}) I(G^{r'}_{ij})\big\}\right] \\ &\quad \leqslant (12)^{1/2} \|v\|_2^2\, \Lambda_{\max}(\Sigma)\{C_3(S_1)\}^{1/2} = \|v\|_2^2\, \Lambda_{\min}(\Sigma). \end{aligned} \tag{S23}$$

In addition, by Hölder's inequality, we have

$$\begin{aligned} E\left[\big\{(X^r_{i,O\backslash j} - X^{r'}_{i,O\backslash j})^{\mathrm{T}} v\big\}^2\right] &= E\left(E\left[\big\{(X^r_{i,O\backslash j} - X^{r'}_{i,O\backslash j})^{\mathrm{T}} v\big\}^2 \;\middle|\; X_{iH}\right]\right) \\ &= 2v^{\mathrm{T}} \Sigma'_{O\backslash j, O\backslash j} v \\ &\geqslant 2\|v\|_2^2\, \Lambda_{\min}(\Sigma). \end{aligned} \tag{S24}$$

Combining (S23) and (S24), we obtain

$$2\|v\|_2^2 \Lambda_{\min}(\Sigma) \leqslant \|v\|_2^2 \Lambda_{\min}(\Sigma) + E\left[\big\{(X^r_{i,O\backslash j} - X^{r'}_{i,O\backslash j})^{\mathrm{T}} v\big\}^2 I(G^r_{ij}) I(G^{r'}_{ij})\right],$$

implying

$$E\left[\big\{(X^r_{i,O\backslash j} - X^{r'}_{i,O\backslash j})^{\mathrm{T}} v\big\}^2 I(G^r_{ij}) I(G^{r'}_{ij})\right] \geqslant \|v\|_2^2\, \Lambda_{\min}(\Sigma). \tag{S25}$$

By (S21) and (S25), we conclude that

$$v^{\mathrm{T}} E\{\nabla^2 \ell_j(\beta^*_{j,O\backslash j})\} v \geqslant C_1 C_2 \|v\|_2^2\, \Lambda_{\min}(\Sigma). \tag{S26}$$

Lower bound for $v^{\mathrm{T}} \nabla^2 \ell_j(\beta^*_{j,O\backslash j}) v$. Gaussian random variables satisfy Assumption 3 (see Yang et al., 2015). Also, recall that $H^j = E\{\nabla^2 \ell_j(\beta^*_{j,O\backslash j})\}$. Therefore, by Lemma S5,

$$\left\|\nabla^2 \ell_j(\beta^*_{j,O\backslash j}) - H^j\right\|_\infty \leqslant K_1 (\log^9 p/n)^{1/2}$$

with probability $1 - p^{-1}$. Let $\Gamma = H^j - \nabla^2 \ell_j(\beta^*_{j,O\backslash j})$. By Hölder's inequality,

$$v^{\mathrm{T}} H^j v - v^{\mathrm{T}} \nabla^2 \ell_j(\beta^*_{j,O\backslash j}) v \leqslant v^{\mathrm{T}} \Gamma v \leqslant \|v\|_1^2 \|\Gamma\|_\infty \leqslant \|v\|_1^2 K_1 (\log^9 p/n)^{1/2}. \tag{S27}$$

Note that $\|v_{\mathcal{S}_j}\|_1 \leqslant s_j^{1/2} \|v_{\mathcal{S}_j}\|_2 \leqslant s_j^{1/2} \|v\|_2$ and that for any $v \in \mathcal{A}$, $\|v_{\mathcal{S}_j^{\mathrm{c}}}\|_1 \leqslant \xi \|v_{\mathcal{S}_j}\|_1$. By combining (S26) and (S27) and using the above facts, we obtain

$$\begin{aligned} C_1 C_2 \|v\|_2^2 \Lambda_{\min}(\Sigma) &\leqslant \|v\|_1^2 K_1 (\log^9 p/n)^{1/2} + v^{\mathrm{T}} \nabla^2 \ell_j(\beta^*_{j,O\backslash j}) v \\ &\leqslant (1+\xi)^2 \|v_{\mathcal{S}_j}\|_1^2 K_1 (\log^9 p/n)^{1/2} + v^{\mathrm{T}} \nabla^2 \ell_j(\beta^*_{j,O\backslash j}) v \\ &\leqslant (1+\xi)^2 \|v\|_2^2 K_1 s_j (\log^9 p/n)^{1/2} + v^{\mathrm{T}} \nabla^2 \ell_j(\beta^*_{j,O\backslash j}) v. \end{aligned}$$

By the assumption $\lim_{n\to\infty} s_{\max} (\log^9 p/n)^{1/2} = 0$, for sufficiently large n we have

$$(1+\xi)^2 K_1 s_j (\log^9 p/n)^{1/2} \leqslant \frac{1}{2} C_1 C_2\, \Lambda_{\min}(\Sigma).$$

Hence

$$\frac{v^{\mathrm{T}}\nabla^2\ell_j(\beta^*_{j,O\setminus j})v}{\|v\|_2^2} \geqslant \frac{1}{2}C_1C_2\,\Lambda_{\min}(\Sigma). \tag{S28}$$

Compatibility factor. By the definition of the compatibility factor, (S11), together with (S28), we have

$$\begin{aligned}\kappa^2\big\{\nabla^2\ell_j(\beta^*_{j,O\setminus j}), s_j\big\} &= \min_{v\in\mathcal{A}} \frac{s_j v^{\mathrm{T}}\nabla^2\ell_j(\beta^*_{j,O\setminus j})v}{\|v_{\mathcal{S}_j}\|_1^2} \\ &\geqslant \min_{v\in\mathcal{A}} \frac{v^{\mathrm{T}}\nabla^2\ell_j(\beta^*_{j,O\setminus j})v}{\|v\|_2^2} \geqslant \frac{1}{2}C_1C_2\,\Lambda_{\min}(\Sigma),\end{aligned}$$

where we have used the fact that $\|v_{\mathcal{S}_j}\|_1 \leqslant s_j^{1/2}\|v_{\mathcal{S}_j}\|_2 \leqslant s_j^{1/2}\|v\|_2$.

Restricted eigenvalue. By the definition of the restricted eigenvalue, (S12), together with (S28), we have

$$\textsc{re}\big\{\nabla^2\ell_j(\beta^*_{j,O\setminus j}), s_j\big\} = \min_{v\in\mathcal{A}} \frac{v^{\mathrm{T}}\nabla^2\ell_j(\beta^*_{j,O\setminus j})v}{\|v\|_2^2} \geqslant \frac{1}{2}C_1C_2\,\Lambda_{\min}(\Sigma).$$

Weak cone invertibility factor. By the definition of the weak cone invertibility factor, (S13), together with (S28), we have

$$\begin{aligned}\rho_q\big\{\nabla^2\ell_j(\beta^*_{j,O\setminus j}), s_j\big\} &= \min_{v\in\mathcal{A}} \frac{s_j^{1/q} v^{\mathrm{T}}\nabla^2\ell_j(\beta^*_{j,O\setminus j})v}{\|v_{\mathcal{S}_j}\|_1\|v\|_q} \\ &\geqslant \min_{v\in\mathcal{A}} \frac{s_j v^{\mathrm{T}}\nabla^2\ell_j(\beta^*_{j,O\setminus j})v}{\|v_{\mathcal{S}_j}\|_1^2} \geqslant \frac{1}{2}C_1C_2\,\Lambda_{\min}(\Sigma),\end{aligned}$$

where the first inequality follows from Hölder's inequality, i.e., $\|v_{\mathcal{S}_j}\|_1 \leqslant s_j^{1-1/q}\|v\|_q$.

S5. PROOF OF THEOREM 3

We start by presenting some lemmas that will be used to prove Theorem 3. The proofs of these lemmas are provided in § S7. Recall from (9) that the pairwise decorrelated score function for β_{jk} is defined as

$$S_{jk}(\beta_{j\vee k}) = \nabla_k\ell_j(\beta_{j,O\setminus j}) + \nabla_j\ell_k(\beta_{k,O\setminus k}) - (w^*_{jk})^{\mathrm{T}}\nabla_{-k}\ell_j(\beta_{j,O\setminus j}) - (w^*_{kj})^{\mathrm{T}}\nabla_{-j}\ell_k(\beta_{k,O\setminus k}).$$

The estimated pairwise decorrelated score function defined in (11) is

$$\begin{aligned}\hat{S}_{jk} = {}& \nabla_k\ell_j(0, \hat{\beta}_{j,O\setminus\{j,k\}}) + \nabla_j\ell_k(0, \hat{\beta}_{k,O\setminus\{j,k\}}) \\ & - \hat{w}_{jk}^{\mathrm{T}}\nabla_{-k}\ell_j(0, \hat{\beta}_{j,O\setminus\{j,k\}}) - \hat{w}_{kj}^{\mathrm{T}}\nabla_{-j}\ell_k(0, \hat{\beta}_{k,O\setminus\{j,k\}}).\end{aligned}$$

The following lemma establishes the asymptotic normality of $S_{jk}(\beta_{j\vee k})$.

LEMMA S6. *Under Assumptions* 3–5*, for $j, k \in O$ we have that $n^{1/2}S_{jk}(\beta^*_{j\vee k})$ converges in distribution to $N(0, \sigma^2_{jk})$, where*

$$\sigma^2_{jk} = \Sigma^{jk}_{jk,jk} - 2\Sigma^{jk}_{jk,j\setminus k}w^*_{jk} - 2\Sigma^{jk}_{jk,k\setminus j}w^*_{kj} + (w^*_{jk})^{\mathrm{T}}\Sigma^{jk}_{j\setminus k,j\setminus k}w^*_{jk} + (w^*_{kj})^{\mathrm{T}}\Sigma^{jk}_{k\setminus j,k\setminus j}w^*_{kj}.$$

Lemma S6 shows that the pairwise decorrelated score function converges to a univariate Gaussian distribution. To derive the asymptotic distribution of $\hat{S}_{jk}$, we show that $n^{1/2}\{\hat{S}_{jk} - S_{jk}(\beta^*_{j\vee k})\} = o_{\mathbb{P}}(1)$. We then use Lemma S6 to establish the asymptotic normality of $\hat{S}_{jk}$. To this end, we need some additional assumptions on the scaling of n and p and on the magnitude of the regularization parameters λ and λ_w in (7) and (10), respectively, as stated in Assumption S1. We also need the following technical lemmas.

LEMMA S7. *Under Assumptions* 3 *and* 4*, for $j \in O$ we have*

$$\left\|\nabla^2 \ell_j(\hat{\beta}_{j,O\setminus j}) - H^j\right\|_\infty = \mathcal{O}_{\mathbb{P}}\left\{\frac{M s_{\max}\lambda}{\kappa^2_{\min}} + \left(\frac{\log^9 p}{n}\right)^{1/2}\right\},$$

where M is a constant from Theorem 1.

LEMMA S8. *Let $w_0 = \max_{j,k\in O} \|w^*_{jk}\|_1$, and let*

$$\lambda_w \geqslant C\left[w_0\left\{\frac{M s_{\max}\lambda}{\kappa^2_{\min}} + \left(\frac{\log^9 p}{n}\right)^{1/2}\right\}\right],$$

where M is as defined in Theorem 1 *and $C > 0$ is a sufficiently large constant. Under Assumptions* 3 *and* 4*, for $j, k \in O$,*

$$\left\|\nabla^2_{k,-k}\ell_j(\hat{\beta}_{j,O\setminus j}) - (w^*_{jk})^{\mathrm{T}}\nabla^2_{-k,-k}\ell_j(\hat{\beta}_{j,O\setminus j})\right\|_\infty \leqslant \lambda_w$$

with probability converging to 1.

LEMMA S9. *Let $s'_{jk} = \|w^*_{jk}\|_0$ and let $s'_{\max} = \max_{j,k\in O}\ s'_{jk}$. Under Assumptions* 3*,* 4 *and* S1*, $\|\hat{w}_{jk} - w^*_{jk}\|_1 = \mathcal{O}_{\mathbb{P}}(s'_{\max}\lambda_w)$ holds for all $j, k \in O$.*

With Lemmas S6–S9, we establish the asymptotic normality of the estimated pairwise decorrelated score function (11).

LEMMA S10. *Under Assumptions* 3–5 *and* S1 *and under the null hypothesis $H_0 : \beta^*_{jk} = \beta^*_{kj} = 0$ for $j, k \in O$, we have that $n^{1/2}\hat{S}_{jk} = n^{1/2}S_{jk}(\beta^*_{j\vee k}) + o_{\mathbb{P}}(1)$ converges in distribution to $N(0, \sigma^2_{jk})$, where*

$$\sigma^2_{jk} = \Sigma^{jk}_{jk,jk} - 2\Sigma^{jk}_{jk,j\setminus k}w^*_{jk} - 2\Sigma^{jk}_{jk,k\setminus j}w^*_{kj} + (w^*_{jk})^{\mathrm{T}}\Sigma^{jk}_{j\setminus k,j\setminus k}w^*_{jk} + (w^*_{kj})^{\mathrm{T}}\Sigma^{jk}_{k\setminus j,k\setminus j}w^*_{kj}. \tag{S29}$$

However, σ^2_{jk} depends on the unknown quantity Σ^{jk}. Recall from (16) that we estimate Σ^{jk} as

$$\hat{\Sigma}^{jk}(0, \hat{\beta}_{j,O\setminus\{j,k\}}, \hat{\beta}_{k,O\setminus\{j,k\}}) = \frac{1}{n}\sum_{i=1}^n \left\{g^{jk}_i(0, \hat{\beta}_{j,O\setminus\{j,k\}}, \hat{\beta}_{k,O\setminus\{j,k\}})\right\}^{\otimes 2},$$

and that we write $\hat{\Sigma}^{jk}$ to indicate $\hat{\Sigma}^{jk}(0, \hat{\beta}_{j,O\setminus\{j,k\}}, \hat{\beta}_{k,O\setminus\{j,k\}})$. The next lemma asserts that $\hat{\Sigma}^{jk}$ is a consistent estimator of Σ^{jk}.

LEMMA S11. *Let* $\hat{\beta}'_{j\vee k} = (0, \hat{\beta}^{\mathrm{T}}_{j,O\setminus\{j,k\}}, \hat{\beta}^{\mathrm{T}}_{k,O\setminus\{j,k\}})^{\mathrm{T}}$. *Under Assumptions* 3, 4 *and* S1 *and the null hypothesis* $H_0 : \beta^*_{jk} = \beta^*_{kj} = 0$,

$$\hat{\Sigma}^{jk}(\hat{\beta}'_{j\vee k}) = \frac{1}{n}\sum_{i=1}^{n}\{g_i^{jk}(\hat{\beta}'_{j\vee k})\}^{\otimes 2}$$

is a consistent estimator of $\Sigma^{jk} = E[\{g_i^{jk}(\beta^*_{j\vee k})\}^{\otimes 2}]$ *as defined in* (15). *In particular,*

$$\left\|\hat{\Sigma}^{jk}(\hat{\beta}'_{j\vee k}) - \Sigma^{jk}\right\|_\infty = \mathcal{O}_{\mathbb{P}}\left\{\frac{s_{\max}}{\kappa^2_{\min}}\lambda\log^6 p + \left(\frac{\log^9 p}{n}\right)^{1/2}\right\}.$$

With Lemma S11, we establish that $\hat{\sigma}^2_{jk}$ in (S29) is a consistent estimator of σ^2_{jk}.

LEMMA S12. *Let*

$$\hat{\sigma}^2_{jk} = \hat{\Sigma}^{jk}_{jk,jk} - 2\hat{\Sigma}^{jk}_{jk,j\setminus k}\hat{w}_{jk} - 2\hat{\Sigma}^{jk}_{jk,k\setminus j}\hat{w}_{kj} + \hat{w}^{\mathrm{T}}_{jk}\hat{\Sigma}^{jk}_{j\setminus k,j\setminus k}\hat{w}_{jk} + \hat{w}^{\mathrm{T}}_{kj}\hat{\Sigma}^{jk}_{k\setminus j,k\setminus j}\hat{w}_{kj}.$$

Under Assumptions 3–5, S1 *and* S2, $|\hat{\sigma}^2_{jk} - \sigma^2_{jk}| = o_{\mathbb{P}}(1)$.

Proof of Theorem 3. By Lemma S12, $\hat{\sigma}^2_{jk} = \sigma^2_{jk} + o_{\mathbb{P}}(1)$. The results follow from an application of Slutsky's theorem to (S29) in Lemma S10. □

S6. PROOF OF THE LEMMAS IN § S4

S6·1. *Proof of Lemma* S1

Proof. Recall from (S3) that

$$h^{rr'}_{ijk}(\beta^*_{j,O\setminus j}) = -\frac{R^{rr'}_{ij}(\beta^*_{j,O\setminus j})(x^r_{ij} - x^{r'}_{ij})(x^r_{ik} - x^{r'}_{ik})}{1 + R^{rr'}_{ij}(\beta^*_{j,O\setminus j})}.$$

By Proposition S1, for any $t > 0$ we have $\mathrm{pr}(|X^r_{ij} - X^{r'}_{ij}| \geqslant t) \leqslant 2c_1\exp(-t/2)$. Note that $R^{rr'}_{ij}(\beta^*_{j,O\setminus j}) > 0$ and therefore $R^{rr'}_{ij}(\beta^*_{j,O\setminus j})/\{1 + R^{rr'}_{ij}(\beta^*_{j,O\setminus j})\} < 1$. Hence, by the union bound, we have that for any $t > 0$,

$$\begin{aligned}\mathrm{pr}\left\{h^{rr'}_{ijk}(\beta^*_{j,O\setminus j}) \geqslant t\right\} &\leqslant \mathrm{pr}\left\{|(X^r_{ij} - X^{r'}_{ij})(X^r_{ik} - X^{r'}_{ik})| \geqslant t\right\}\\ &\leqslant \mathrm{pr}\left(|X^r_{ij} - X^{r'}_{ij}| \geqslant t^{1/2}\right) + \mathrm{pr}\left(|X^r_{ik} - X^{r'}_{ik}| \geqslant t^{1/2}\right)\\ &\leqslant 4c_1\exp(-t^{1/2}/2).\end{aligned}$$

Then, by another application of the union bound,

$$\begin{aligned}\mathrm{pr}\left\{\frac{2}{R(R-1)}\sum_{1\leqslant r<r'\leqslant R} h^{rr'}_{ijk}(\beta^*_{j,O\setminus j}) \geqslant t\right\} &= \mathrm{pr}\left\{\sum_{1\leqslant r<r'\leqslant R} h^{rr'}_{ijk}(\beta^*_{j,O\setminus j}) \geqslant tR(R-1)/2\right\}\\ &\leqslant \sum_{1\leqslant r<r'\leqslant R}\mathrm{pr}\left\{h^{rr'}_{ijk}(\beta^*_{j,O\setminus j}) \geqslant t\right\}\\ &\leqslant 4R^2c_1\exp(-t^{1/2}/2).\end{aligned}$$

S6·2. *Proof of Lemma* S2

From (S1) and Lemma S1, we see that the gradient of the loss function is the average of independent random variables with general exponential tail. To prove Lemma S2, we use a generalization of Bernstein's inequality for independent random variables with general exponential tail given in Lemma S14.

Recall from (S10) that $U_{ik}^{j}(\beta_{j,O\setminus j}^{*}) = 2\sum_{r<r'} h_{ijk}^{rr'}(\beta_{j,O\setminus j}^{*})/\{R(R-1)\}$. By an application of Proposition 2, $E\{U_{ik}^{j}(\beta_{j,O\setminus j}^{*})\} = 0$. In addition, by Lemma S1,

$$\mathrm{pr}\{U_{ik}^{j}(\beta_{j,O\setminus j}^{*}) \geqslant t\} \leqslant 4c_1R^2\exp(-0{\cdot}5t^{1/2}).$$

Since $U_{1k}^{j}(\beta_{j,O\setminus j}^{*}), \ldots, U_{nk}^{j}(\beta_{j,O\setminus j}^{*})$ are independent and identically distributed random variables with mean zero, by an application of Lemma S14 with $L_1 = 4R^2c_1$, $L_2 = 1/2$ and $q = 1/2$, we have

$$\begin{aligned}\mathrm{pr}\left\{\left|\nabla_k\ell_j(\beta_{j,O\setminus j}^{*})\right| \geqslant t\right\} &= \mathrm{pr}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} U_{ik}^{j}(\beta_{j,O\setminus j}^{*})\right| \geqslant t\right\} \\ &\leqslant 4\exp\left(-\frac{1}{8}n^{1/5}t^{2/5}\right) + 16nR^2c_1\exp\{-n^{1/5}t^{2/5}/(2^{3/2})\}.\end{aligned}$$

Therefore, by the union bound, we obtain

$$\mathrm{pr}\{\|\nabla\ell_j(\beta_{j,O\setminus j}^{*})\|_\infty \geqslant t\} \leqslant 4p\exp\left(-\frac{1}{8}n^{1/5}t^{2/5}\right) + 16npR^2c_1\exp\{-n^{1/5}t^{2/5}/(2^{3/2})\}.$$

Note that Lemma S14 holds only if $t \geqslant (8E[\{U_{ik}^{j}(\beta_{j,O\setminus j}^{*})\}^2]/n)^{1/2}$ for $k \in O \setminus j$. By Proposition S1, $E[\{U_{ik}^{j}(\beta_{j,O\setminus j}^{*})\}^2]$ is bounded since $E(X_j^4)$ is bounded. We take $t = K(\log^5 p/n)^{1/2}$ for sufficiently large $K > 0$ such that the inequality $t \geqslant (8E[\{U_{ik}^{j}(\beta_{j,O\setminus j}^{*})\}^2]/n)^{1/2}$ holds for all $k \in O \setminus j$. Then, for sufficiently large K, we have

$$\mathrm{pr}\left\{\left\|\nabla\ell_j(\beta_{j,O\setminus j}^{*})\right\|_\infty \geqslant K\left(\frac{\log^5 p}{n}\right)^{1/2}\right\} \leqslant p^{-1}.$$

We conclude that $\|\nabla\ell_j(\beta_{j,O\setminus j}^{*})\|_\infty \leqslant K(\log^5 p/n)^{1/2}$ with probability at least $1 - p^{-1}$.

S6·3. *Proof of Lemma* S4

Recall that $\mathcal{S}_j$ is the support set of $\beta_{j,O\setminus j}^{*}$ and that

$$D\left(\hat{\beta}_{j,O\setminus j}, \beta_{j,O\setminus j}\right) = \left(\hat{\beta}_{j,O\setminus j} - \beta_{j,O\setminus j}\right)^{\mathrm{T}}\left\{\nabla\ell_j(\hat{\beta}_{j,O\setminus j}) - \nabla\ell_j(\beta_{j,O\setminus j})\right\}$$

is the symmetrized Bregman divergence of $\ell_j(\beta_{j,O\setminus j})$. Also, recall that $\hat{\Delta}_j = \hat{\beta}_{j,O\setminus j} - \beta_{j,O\setminus j}^{*}$. Observe that

$$\begin{aligned}&D\left(\hat{\beta}_{j,O\setminus j}, \beta_{j,O\setminus j}^{*}\right) \\ &\quad= \hat{\Delta}_j^{\mathrm{T}}\left\{\nabla\ell_j(\beta_{j,O\setminus j}^{*} + \hat{\Delta}_j) - \nabla\ell_j(\beta_{j,O\setminus j}^{*})\right\} \\ &\quad= \sum_{k\in\mathcal{S}_j^{\mathrm{c}}} \hat{\beta}_{jk}\nabla_k\ell_j(\beta_{j,O\setminus j}^{*} + \hat{\Delta}_j) + \sum_{k\in\mathcal{S}_j} \hat{\Delta}_{jk}\nabla_k\ell_j(\beta_{j,O\setminus j}^{*} + \hat{\Delta}_j) - \hat{\Delta}_j^{\mathrm{T}}\nabla\ell_j(\beta_{j,O\setminus j}^{*}).\end{aligned}$$

By the Karush–Kuhn–Tucker conditions of (7), $\hat{\beta}_{j,O\setminus j}$ is a solution to (7) if and only if

$$\begin{cases} \nabla_k \ell_j(\hat{\beta}_{j,O\setminus j}) = -\lambda \operatorname{sign}(\hat{\beta}_{jk}) & \text{if } \hat{\beta}_{jk} \neq 0, \\ \big|\nabla_k \ell_j(\hat{\beta}_{j,O\setminus j})\big| \leqslant \lambda & \text{if } \hat{\beta}_{jk} = 0. \end{cases}$$

Note that $k \in \mathcal{S}_j^{\mathrm{c}}$ does not imply $\hat{\beta}_{jk} = 0$, since $\mathcal{S}_j$ is the support set of $\beta^*_{j,O\setminus j}$. Nonetheless, by the Karush–Kuhn–Tucker conditions, $\sum_{k\in\mathcal{S}_j^{\mathrm{c}}} \hat{\beta}_{jk} \nabla_k \ell_j(\beta^*_{j,O\setminus j} + \hat{\Delta}_j) = -\lambda \sum_{k\in\mathcal{S}_j^{\mathrm{c}}} \hat{\beta}_{jk} \operatorname{sign}(\hat{\beta}_{jk})$ and $\sum_{k\in\mathcal{S}_j} \hat{\Delta}_{jk} \nabla_k \ell_j(\beta^*_{j,O\setminus j} + \hat{\Delta}_j) \leqslant \lambda \sum_{k\in\mathcal{S}_j} |\hat{\Delta}_{jk}|$. Therefore

$$\begin{aligned} D\left(\hat{\beta}_{j,O\setminus j}, \beta^*_{j,O\setminus j}\right) &\leqslant -\lambda \sum_{k\in\mathcal{S}_j^{\mathrm{c}}} \hat{\beta}_{jk} \operatorname{sign}(\hat{\beta}_{jk}) + \lambda \sum_{k\in\mathcal{S}_j} |\hat{\Delta}_{jk}| - \hat{\Delta}_j^{\mathrm{T}} \nabla \ell_j(\beta^*_{j,O\setminus j}) \\ &\leqslant -\lambda \|\hat{\Delta}_{j,\mathcal{S}_j^{\mathrm{c}}}\|_1 + \lambda \|\hat{\Delta}_{j,\mathcal{S}_j}\|_1 + \|\hat{\Delta}_j\|_1 \|\nabla \ell_j(\beta^*_{j,O\setminus j})\|_\infty \\ &\leqslant -\lambda \|\hat{\Delta}_{j,\mathcal{S}_j^{\mathrm{c}}}\|_1 + \lambda \|\hat{\Delta}_{j,\mathcal{S}_j}\|_1 + (\|\hat{\Delta}_{j,\mathcal{S}_j}\|_1 + \|\hat{\Delta}_{j,\mathcal{S}_j^{\mathrm{c}}}\|_1) \|\nabla \ell_j(\beta^*_{j,O\setminus j})\|_\infty \\ &= \big\{\lambda + \|\nabla \ell_j(\beta^*_{j,O\setminus j})\|_\infty\big\} \|\hat{\Delta}_{j,S_j}\|_1 - \big\{\lambda - \|\nabla \ell_j(\beta^*_{j,O\setminus j})\|_\infty\big\} \|\hat{\Delta}_{j,S_j^{\mathrm{c}}}\|_1. \end{aligned} \tag{S30}$$

The inequality is obtained by rearranging the terms in the last expression. To show that $\|\hat{\Delta}_{j,\mathcal{S}_j^{\mathrm{c}}}\|_1 \leqslant \xi \|\hat{\Delta}_{j,\mathcal{S}_j}\|_1$, we use the fact that $D(\hat{\beta}_{j,O\setminus j}, \beta^*_{j,O\setminus j}) \geqslant 0$ since the loss function $\ell_j(\cdot)$ is a convex function. The inequality is obtained by substituting $\|\nabla \ell_j(\beta^*_{j,O\setminus j})\|_\infty \leqslant \lambda(\xi - 1)/(\xi + 1)$ into (S30) and rearranging the terms.

S6·4. *Proof of Lemma* S5

Recall the definitions of $T_{ij}^{rr'}(\beta^*_{j,O\setminus j})$, $T_{ijkl}^{rr'}(\beta^*_{j,O\setminus j})$ and $\nabla^2 \ell_j(\beta^*_{j,O\setminus j})$ from (S4), (S5) and (S7), respectively. We first show that $\|E\{T_{ij}^{rr'}(\beta^*_{j,O\setminus j})\}\|_\infty$ is bounded. By Proposition S1, for a sufficiently large constant C,

$$\big\|E\{T_{ij}^{rr'}(\beta^*_{j,O\setminus j})\}\big\|_\infty \leqslant C \max_{i,j,r} E|X_{ij}^r|^4 \leqslant 24 C c_1.$$

Hence, by the union bound, for any $t \geqslant 2\|E\{T_{ij}^{rr'}(\beta^*_{j,O\setminus j})\}\|_\infty$,

$$\begin{aligned} &\operatorname{pr}\Big[\Big|T_{ijkl}^{rr'}(\beta^*_{j,O\setminus j}) - E\{T_{ijkl}^{rr'}(\beta^*_{j,O\setminus j})\}\Big| \geqslant t\Big] \\ &\quad\leqslant \operatorname{pr}\Big[\Big|T_{ijkl}^{rr'}(\beta^*_{j,O\setminus j})\Big| \geqslant t - \Big|E\{T_{ijkl}^{rr'}(\beta^*_{j,O\setminus j})\}\Big|\Big] \\ &\quad\leqslant \operatorname{pr}\Big\{\Big|T_{ijkl}^{rr'}(\beta^*_{j,O\setminus j})\Big| \geqslant t/2\Big\} \\ &\quad\leqslant \operatorname{pr}\Big\{\Big|(X_{ij}^r - X_{ij}^{r'})^2 (X_{ik}^r - X_{ik}^{r'})(X_{il}^r - X_{il}^{r'})\Big| \geqslant t/2\Big\} \\ &\quad\leqslant \operatorname{pr}\Big\{\Big|X_{ij}^r - X_{ij}^{r'}\Big| \geqslant (t/2)^{1/4}\Big\} + \operatorname{pr}\Big\{\Big|X_{ik}^r - X_{ik}^{r'}\Big| \geqslant (t/2)^{1/4}\Big\} \\ &\qquad + \operatorname{pr}\Big\{\Big|X_{il}^r - X_{il}^{r'}\Big| \geqslant (t/2)^{1/4}\Big\} \\ &\quad\leqslant 6c_1 \exp\big(-t^{1/4} 2^{-5/4}\big), \end{aligned}$$

where the last inequality follows from Proposition S1.

Choosing $C_H = \max[6c_1, \exp\{(1.5c_1C)^{1/4}\}]$, we have that for any $t > 0$,

$$\operatorname{pr}\Big[\Big|T^{rr'}_{ijkl}(\beta^*_{j,O\setminus j}) - E\{T^{rr'}_{ijkl}(\beta^*_{j,O\setminus j})\}\Big| \geqslant t\Big] \leqslant C_H \exp\big(-t^{1/4}2^{-5/4}\big).$$

Let

$$V_{ikl} = \frac{2}{R(R-1)} \sum_{1\leqslant r<r'\leqslant R} \Big[T^{rr'}_{ijkl}(\beta^*_{j,O\setminus j}) - E\{T^{rr'}_{ijkl}(\beta^*_{j,O\setminus j})\}\Big].$$

Then, by the union bound,

$$\begin{aligned}\operatorname{pr}(V_{ikl} \geqslant t) &\leqslant \sum_{1\leqslant r<r'\leqslant R} \operatorname{pr}\Big[\Big|T^{rr'}_{ijkl}(\beta^*_{j,O\setminus j}) - E\{T^{rr'}_{ijkl}(\beta^*_{j,O\setminus j})\}\Big| \geqslant t\Big] \\ &\leqslant C_H R^2 \exp\big(-t^{1/4}2^{-5/4}\big).\end{aligned}$$

By the definition of V_{ikl}, we have $E(V_{ikl}) = 0$. Since $V_{1kl}, \ldots, V_{nkl}$ are independent and identically distributed random variables with mean zero, by an application of Lemma S14 with $L_1 = C_H R^2$, $L_2 = 2^{-5/4}$ and $q = 1/4$, we have that for any $t \geqslant \{E(V^2_{ijk})/n\}^{1/2}$,

$$\begin{aligned}\operatorname{pr}\Big\{\Big|\nabla^2_{kl}\ell_j(\beta^*_{j,O\setminus j}) - H^j_{kl}\Big| \geqslant t\Big\} &= \operatorname{pr}\left(\left|\frac{1}{n}\sum_{i=1}^n V_{ikl}\right| \geqslant t\right) \\ &\leqslant 4\exp\left(-\frac{1}{8}n^{1/9}t^{2/9}\right) + 4nC_HR^2\exp\big\{-n^{1/9}t^{2/9}/(2^{3/2})\big\}.\end{aligned}$$

Therefore, by the union bound, we obtain

$$\begin{aligned}&\operatorname{pr}\Big\{\Big\|\nabla^2\ell_j(\beta^*_{j,O\setminus j}) - H^j\Big\|_\infty \geqslant t\Big\} \\ &\quad\leqslant 4p^2\exp\left(-\frac{1}{8}n^{1/9}t^{2/9}\right) + 4np^2C_HR^2\exp\big\{-n^{1/9}t^{2/9}/(2^{3/2})\big\}.\end{aligned} \tag{S31}$$

Note that (S31) holds only if $t \geqslant \{E(V^2_{ijk})/n\}^{1/2}$. It can be verified that $E(V^2_{ijk})$ is bounded since $E(X_j^8)$ is bounded by Proposition S1. Therefore, taking $t = K_1(\log^9 p/n)^{1/2}$ for sufficiently large K_1, we obtain

$$\operatorname{pr}\Big(\Big\|\nabla^2\ell_j(\beta^*_{j,O\setminus j}) - H^j\Big\|_\infty \geqslant t\Big) \leqslant p^{-1}.$$

We conclude that $\|\nabla^2\ell_j(\beta^*_{j,O\setminus j}) - H^j\|_\infty \leqslant K_1(\log^9 p/n)^{1/2}$ with probability at least $1 - p^{-1}$.

S7. Proof of the lemmas in § S5

S7·1. *Proof of Lemma* S6

Recall from § 2·4 that the parameters associated with the jth and kth nodes are $\beta_{j\vee k} = (\beta_{jk}, \beta^{\mathrm{T}}_{j,O\setminus\{j,k\}}, \beta^{\mathrm{T}}_{k,O\setminus\{j,k\}})^{\mathrm{T}} \in \mathbb{R}^{2p-3}$. Let $L_{jk}(\beta_{j\vee k}) = \ell_j(\beta_{j,O\setminus j}) + \ell_k(\beta_{k,O\setminus k})$. The gradient of $L_{jk}(\beta_{j\vee k})$ evaluated at $\beta_{j\vee k}$ is

$$\begin{aligned}\nabla_{jk}L_{jk}(\beta_{j\vee k}) &= \frac{\partial L_{jk}(\beta_{j\vee k})}{\partial\beta_{jk}} = \nabla_k\ell_j(\beta_{j,O\setminus j}) + \nabla_j\ell_k(\beta_{k,O\setminus k}) \in \mathbb{R}, \\ \nabla_{j,-k}L_{jk}(\beta_{j\vee k}) &= \frac{\partial L_{jk}(\beta_{j\vee k})}{\partial\beta_{j,O\setminus\{j,k\}}} = \nabla_{-k}\ell_j(\beta_{j,O\setminus j}) \in \mathbb{R}^{p-2},\end{aligned} \tag{S32}$$

$$\nabla_{k,-j} L_{jk}(\beta_{j\vee k}) = \frac{\partial L_{jk}(\beta_{j\vee k})}{\partial \beta_{k,O\setminus\{j,k\}}} = \nabla_{-j}\ell_k(\beta_{k,O\setminus k}) \in \mathbb{R}^{p-2}.$$

From (S32) and the definition of $S_{jk}(\beta_{j\vee k})$ in (9), we see that $S_{jk}(\beta_{j\vee k})$ is a linear transformation of $\nabla L_{jk}(\beta_{j\vee k}) \in \mathbb{R}^{2p-3}$. Let $b = \{1, (-w^*_{jk})^{\mathrm{T}}, (-w^*_{kj})^{\mathrm{T}}\}^{\mathrm{T}} \in \mathbb{R}^{2p-3}$. Then it can be verified that

$$n^{1/2} S_{jk}(\beta^*_{j\vee k}) = n^{1/2} b^{\mathrm{T}} \nabla L_{jk}(\beta^*_{j\vee k}) = \frac{1}{n^{1/2}} \sum_{i=1}^{n} b^{\mathrm{T}} g_i^{jk}(\beta^*_{j\vee k}),$$

where $g_i^{jk}(\beta^*_{j\vee k})$ is as defined in (14). By Proposition 2, $g_1^{jk}(\beta^*_{j\vee k}), \ldots, g_n^{jk}(\beta^*_{j\vee k})$ are independent and identically distributed random variables with mean zero and $\mathrm{var}\{g_i^{jk}(\beta^*_{j\vee k})\} = \Sigma^{jk}$, where Σ^{jk} is as defined in (15). By an application of the central limit theorem, we have that $n^{1/2} S_{jk}(\beta^*_{j\vee k})$ converges in distribution to $N(0, \sigma^2_{jk})$, where

$$\sigma^2_{jk} = \Sigma^{jk}_{jk,jk} - 2\Sigma^{jk}_{jk,j\setminus k} w^*_{jk} - 2\Sigma^{jk}_{jk,k\setminus j} w^*_{kj} + (w^*_{jk})^{\mathrm{T}} \Sigma^{jk}_{j\setminus k, j\setminus k} w^*_{jk} + (w^*_{kj})^{\mathrm{T}} \Sigma^{jk}_{k\setminus j, k\setminus j} w^*_{kj}.$$

S7·2. *Proof of Lemma* S7

Recall from § 2·4 that $H^j = E\{\nabla^2 \ell_j(\beta^*_{j,O\setminus j})\}$. By the triangle inequality,

$$\begin{aligned}\left\|\nabla^2 \ell_j(\hat\beta_{j,O\setminus j}) - H^j\right\|_\infty &\leqslant \left\|\nabla^2 \ell_j(\hat\beta_{j,O\setminus j}) - \nabla^2 \ell_j(\beta^*_{j,O\setminus j})\right\|_\infty + \left\|\nabla^2 \ell_j(\beta^*_{j,O\setminus j}) - H^j\right\|_\infty \\ &= I_1 + I_2.\end{aligned}$$

By Lemma S5, we have $I_2 = \mathcal{O}_{\mathbb{P}}\{(\log^9 p/n)^{1/2}\} = o_{\mathbb{P}}(1)$. By Lemma S3,

$$\nabla^2 \ell_j(\hat\beta_{j,O\setminus j}) - \nabla^2 \ell_j(\beta^*_{j,O\setminus j}) \leqslant \{\exp(b) - 1\} \nabla^2 \ell_j(\beta^*_{j,O\setminus j}),$$

where

$$\begin{aligned} b &= \max_{1\leqslant i\leqslant n} \max_{1\leqslant r<r'\leqslant R} \left|(x^r_{ij} - x^{r'}_{ij})(\hat\beta_{j,O\setminus j} - \beta^*_{j,O\setminus j})^{\mathrm{T}}(x^r_{i,O\setminus j} - x^{r'}_{i,O\setminus j})\right| \\ &\leqslant M\|\hat\beta_{j,O\setminus j} - \beta^*_{j,O\setminus j}\|_1,\end{aligned}$$

with M as defined in Theorem 1. By Assumption 4, $\|H^j\|_\infty = \mathcal{O}(1)$. Therefore, we obtain

$$\begin{aligned} I_1 &\leqslant |\exp(b) - 1| \|\nabla^2 \ell_j(\beta^*_{j,O\setminus j})\|_\infty \\ &\leqslant |\exp(b) - 1| \left(I_2 + \|H^j\|_\infty\right) \\ &= \mathcal{O}_{\mathbb{P}}(|b|)\{o_{\mathbb{P}}(1) + \mathcal{O}(1)\} \\ &= \mathcal{O}_{\mathbb{P}}(M s_{\max} \lambda/\kappa^2_{\min}),\end{aligned}$$

where the first inequality follows from Hölder's inequality, the second inequality follows from an application of the triangle inequality, and the last equality is obtained from an application of Theorem 1.

S7·3. *Proof of Lemma* S8

Recall from § 2·4 that $(w^*_{jk})^{\mathrm{T}} = (H^j_{k,-k})^{\mathrm{T}} (H^j_{-k,-k})^{-1}$. By the triangle inequality and the definition of w^*_{jk}, we obtain

$$\left\|\nabla^2_{k,-k} \ell_j(\hat\beta_{j,O\setminus j}) - (w^*_{jk})^{\mathrm{T}} \nabla^2_{-k,-k} \ell_j(\hat\beta_{j,O\setminus j})\right\|_\infty$$

$$\begin{aligned}
&\leqslant \Big\|\nabla^2_{k,-k}\ell_j(\hat\beta_{j,O\setminus j}) - H^j_{k,-k}\Big\|_\infty + \Big\|(w^*_{jk})^{\mathrm{T}}\nabla^2_{-k,-k}\ell_j(\hat\beta_{j,O\setminus j}) - H^j_{k,-k}\Big\|_\infty \\
&= \Big\|\nabla^2_{k,-k}\ell_j(\hat\beta_{j,O\setminus j}) - H^j_{k,-k}\Big\|_\infty + \Big\|(w^*_{jk})^{\mathrm{T}}\big\{\nabla^2_{-k,-k}\ell_j(\hat\beta_{j,O\setminus j}) - H^j_{-k,-k}\big\}\Big\|_\infty \\
&= I_1 + I_2.
\end{aligned}$$

By Lemma S7, we have $I_1 = \mathcal{O}_{\mathbb{P}}\{Ms_{\max}\lambda/\kappa^2_{\min} + (\log^9 p/n)^{1/2}\}$. Similarly, by Hölder's inequality and Lemma S7, we have $I_2 = \mathcal{O}_{\mathbb{P}}[w_0\{Ms_{\max}\lambda/\kappa^2_{\min} + (\log^9 p/n)^{1/2}\}]$. Therefore,

$$\begin{aligned}
&\Big\|\nabla^2_{k,-k}\ell_j(\hat\beta_{j,O\setminus j}) - (w^*_{jk})^{\mathrm{T}}\nabla^2_{-k,-k}\ell_j(\hat\beta_{j,O\setminus j})\Big\|_\infty \\
&\qquad = \mathcal{O}_{\mathbb{P}}\Big[w_0\big\{Ms_{\max}\lambda/\kappa^2_{\min} + (\log^9 p/n)^{1/2}\big\}\Big].
\end{aligned}$$

Picking $\lambda_w \geqslant Cw_0\{Ms_{\max}\lambda/\kappa^2_{\min} + (\log^9 p/n)^{1/2}\}$ for some sufficiently large C, we have

$$\Big\|\nabla^2_{k,-k}\ell_j(\hat\beta_{j,O\setminus j}) - (w^*_{jk})^{\mathrm{T}}\nabla^2_{-k,-k}\ell_j(\hat\beta_{j,O\setminus j})\Big\|_\infty \leqslant \lambda_w.$$

S7·4. *Proof of Lemma* S9

By Lemma S8, for $\lambda_w \geqslant Cw_0\{Ms_{\max}\lambda/\kappa^2_{\min} + (\log^9 p/n)^{1/2}\}$, w^*_{jk} is in the feasible region of the Dantzig selector problem (10); that is,

$$\Big\|\nabla^2_{k,-k}\ell_j(0,\hat\beta_{j,O\setminus\{j,k\}}) - (w^*_{jk})^{\mathrm{T}}\nabla^2_{-k,-k}\ell_j(0,\hat\beta_{j,O\setminus\{j,k\}})\Big\|_\infty \leqslant \lambda_w \tag{S33}$$

with high probability. For notational convenience, we let $\hat\Delta = \hat w_{jk} - w^*_{jk}$. By the triangle inequality,

$$\begin{aligned}
\Big\|\hat\Delta^{\mathrm{T}}\nabla^2_{-k,-k}\ell_j(0,\hat\beta_{j,O\setminus\{j,k\}})\Big\|_\infty &\leqslant \Big\|\nabla^2_{k,-k}\ell_j(0,\hat\beta_{j,O\setminus\{j,k\}}) - \hat w^{\mathrm{T}}_{jk}\nabla^2_{-k,-k}\ell_j(0,\hat\beta_{j,O\setminus\{j,k\}})\Big\|_\infty \\
&\quad + \Big\|\nabla^2_{k,-k}\ell_j(0,\hat\beta_{j,O\setminus\{j,k\}}) - (w^*_{jk})^{\mathrm{T}}\nabla^2_{-k,-k}\ell_j(0,\hat\beta_{j,O\setminus\{j,k\}})\Big\|_\infty \\
&\leqslant 2\lambda_w,
\end{aligned} \tag{S34}$$

where the last inequality follows from (S33) and the constraint in (10). By Hölder's inequality and (S34),

$$\hat\Delta^{\mathrm{T}}\nabla^2_{-k,-k}\ell_j(0,\hat\beta_{j,O\setminus\{j,k\}})\hat\Delta \leqslant \|\hat\Delta\|_1\|\hat\Delta^{\mathrm{T}}\nabla^2_{-k,-k}\ell_j(0,\hat\beta_{j,O\setminus\{j,k\}})\|_\infty \leqslant 2\lambda_w\|\hat\Delta\|_1. \tag{S35}$$

Let $\mathcal{B}_{jk}$ be the support set of w^*_{jk}, i.e., $\mathcal{B}_{jk} = \{l : (w^*_{jk})_l \neq 0\}$. Also, let $s'_{jk} = |\mathcal{B}_{jk}|$ be the cardinality of $\mathcal{B}_{jk}$. By the definition of the Dantzig selector, (10), $\|\hat w_{jk}\|_1 \leqslant \|w^*_{jk}\|_1$, implying

$$\sum_{l\in\mathcal{B}_{jk}} |(w^*_{jk})_l| \geqslant \sum_{l\in\mathcal{B}_{jk}} |(\hat w_{jk})_l| + \sum_{l\in\mathcal{B}^{\mathrm{c}}_{jk}} |(\hat w_{jk})_l|. \tag{S36}$$

By the triangle inequality,

$$\sum_{l\in\mathcal{B}_{jk}} |(\hat w_{jk})_l - (w^*_{jk})_l| \geqslant \sum_{l\in\mathcal{B}_{jk}} |(w^*_{jk})_l| - \sum_{l\in\mathcal{B}_{jk}} |(\hat w_{jk})_l|. \tag{S37}$$

Upon adding (S36) and (S37) and rearranging the terms, we obtain $\|\hat\Delta_{\mathcal{B}^{\mathrm{c}}_{jk}}\|_1 \leqslant \|\hat\Delta_{\mathcal{B}_{jk}}\|_1$, which implies

$$\|\hat\Delta\|_1 \leqslant 2\|\hat\Delta_{\mathcal{B}_{jk}}\|_1. \tag{S38}$$

Substituting (S38) into (S35) gives

$$\hat{\Delta}^{\mathrm{T}}\nabla^2_{-k,-k}\ell_j(0,\hat{\beta}_{j,O\setminus\{j,k\}})\hat{\Delta} \leqslant 4\lambda_w\|\hat{\Delta}_{\mathcal{B}_{jk}}\|_1. \tag{S39}$$

We now derive a lower bound for $\hat{\Delta}^{\mathrm{T}}\nabla^2_{-k,-k}\ell_j(0,\hat{\beta}_{j,O\setminus\{j,k\}})\hat{\Delta}$. Let $\mathcal{A} = (v \in \mathbb{R}^{p-2} : \|v_{\mathcal{B}^{\mathrm{c}}_{jk}}\|_1 \leqslant \|v_{\mathcal{B}_{jk}}\|_1)$. We first show that for sufficiently large n, the quantity

$$\inf_{v\in\mathcal{A}} \frac{s'_{jk}\big\{v^{\mathrm{T}}\nabla^2_{-k,-k}\ell_j(0,\hat{\beta}_{j,O\setminus\{j,k\}})v\big\}}{\|v_{\mathcal{B}_{jk}}\|_1^2}$$

can be bounded below by a positive constant. By an application of Lemma S3, we obtain

$$\frac{s'_{jk}\big\{v^{\mathrm{T}}\nabla^2_{-k,-k}\ell_j(0,\hat{\beta}_{j,O\setminus\{j,k\}})v\big\}}{\|v_{\mathcal{B}_{jk}}\|_1^2} \geqslant \frac{s'_{jk}\big\{v^{\mathrm{T}}\nabla^2_{-k,-k}\ell_j(0,\beta^*_{j,O\setminus\{j,k\}})v\big\}}{\|v_{\mathcal{B}_{jk}}\|_1^2}\exp(-b), \tag{S40}$$

where

$$b = \max_{1\leqslant i\leqslant n}\max_{1\leqslant r<r'\leqslant R}\left|(x^r_{ij}-x^{r'}_{ij})(0,\hat{\beta}_{j,O\setminus\{j,k\}}-\beta^*_{j,O\setminus\{j,k\}})^{\mathrm{T}}(x^r_{i,O\setminus j}-x^{r'}_{i,O\setminus j})\right|.$$

By Theorem 1 and Assumption S1,

$$b \leqslant M\big\|\hat{\beta}_{j,O\setminus\{j,k\}}-\beta^*_{j,O\setminus\{j,k\}}\big\|_1 \leqslant \frac{M\exp(\eta)(\xi+1)}{2\kappa^2_{\min}}s_{\max}\lambda \leqslant \log 2 \tag{S41}$$

for sufficiently large n.

By (S40), (S41) and Assumption 4 that $\Lambda_{\min}(H^j) \geqslant \Lambda^H_{\mathrm{lower}} > 0$, we have

$$\begin{aligned}
&\frac{s'_{jk}\big\{v^{\mathrm{T}}\nabla^2_{-k,-k}\ell_j(0,\hat{\beta}_{j,O\setminus\{j,k\}})v\big\}}{\|v_{\mathcal{B}_{jk}}\|_1^2}\\
&\quad\geqslant \frac{1}{2}\frac{s'_{jk}\big\{v^{\mathrm{T}}\nabla^2_{-k,-k}\ell_j(0,\beta^*_{j,O\setminus\{j,k\}})v\big\}}{\|v_{\mathcal{B}_{jk}}\|_1^2}\\
&\quad= \frac{s'_{jk}v^{\mathrm{T}}\big\{H^j_{-k,-k}-H^j_{-k,-k}+\nabla^2_{-k,-k}\ell_j(0,\beta^*_{j,O\setminus\{j,k\}})\big\}v}{2\|v_{\mathcal{B}_{jk}}\|_1^2}\\
&\quad\geqslant \frac{s'_{jk}\|v\|_2^2\Lambda^H_{\mathrm{lower}}-s'_{jk}\|H^j_{-k,-k}-\nabla^2_{-k,-k}\ell_j(0,\beta^*_{j,O\setminus\{j,k\}})\|_\infty\|v\|_1^2}{2\|v_{\mathcal{B}_{jk}}\|_1^2},
\end{aligned} \tag{S42}$$

where the last expression follows by an application of Hölder's inequality. Noting that $\|v_{\mathcal{B}_{jk}}\|_1 \leqslant (s'_{jk})^{1/2}\|v_{\mathcal{B}_{jk}}\|_2 \leqslant (s'_{jk})^{1/2}\|v\|_2$ and $\|v\|_1^2 \leqslant 4\|v_{\mathcal{B}_{jk}}\|_1^2$ for any $v \in \mathcal{A}$, we obtain

$$\begin{aligned}
&\frac{s'_{jk}\|v\|_2^2\Lambda^H_{\mathrm{lower}}-s'_{jk}\|H^j_{-k,-k}-\nabla^2_{-k,-k}\ell_j(0,\beta^*_{j,O\setminus\{j,k\}})\|_\infty\|v\|_1^2}{2\|v_{\mathcal{B}_{jk}}\|_1^2}\\
&\quad\geqslant \frac{1}{2}\Lambda^H_{\mathrm{lower}}-2s'_{jk}\|H^j_{-k,-k}-\nabla^2_{-k,-k}\ell_j(0,\beta^*_{j,O\setminus\{j,k\}})\|_\infty\\
&\quad\geqslant \frac{1}{2}\Lambda^H_{\mathrm{lower}}-2K_1s'_{jk}(\log^9 p/n)^{1/2}\\
&\quad= \frac{1}{2}\Lambda^H_{\mathrm{lower}}+o_{\mathbb{P}}(1),
\end{aligned} \tag{S43}$$

where the last inequality is obtained by applying Lemma S7 and the last equality follows from Assumption S1.

Upon setting $v = \hat{\Delta}$ and combining (S42) and (S43), we obtain

$$\hat{\Delta}^{\mathrm{T}} \nabla^2_{-k,-k} \ell_j(0, \hat{\beta}_{j,O\setminus\{j,k\}}) \hat{\Delta} \geqslant \frac{1}{2} \Lambda^H_{\mathrm{lower}} (s'_{jk})^{-1} \|\hat{\Delta}_{\mathcal{B}_{jk}}\|_1^2. \tag{S44}$$

Finally, combining (S39) and (S44) yields

$$\|\hat{\Delta}_{\mathcal{B}_{jk}}\|_1 \leqslant 8 s'_{jk} \lambda_w / \Lambda^H_{\mathrm{lower}},$$

which implies

$$\|\hat{w}_{jk} - w^*_{jk}\|_1 = \|\hat{\Delta}\|_1 \leqslant 2 \|\hat{\Delta}_{\mathcal{B}_{jk}}\|_1 \leqslant 16 s'_{jk} \lambda_w / \Lambda^H_{\mathrm{lower}} = \mathcal{O}_{\mathbb{P}}(s'_{\max} \lambda_w),$$

where the first inequality is obtained from (S38).

S7·5. *Proof of Lemma* S10

The proof consists of two parts. We first show that under the null hypothesis $H_0 : \beta^*_{jk} = \beta^*_{kj} = 0$ in (8), $n^{1/2} \hat{S}_{jk} = n^{1/2} S_{jk}(\beta^*_{j\vee k}) + o_{\mathbb{P}}(1)$. Then, by an application of Lemma S6, we show that $n^{1/2} \hat{S}_{jk}$ is asymptotically normal.

Recall the definitions of $S_{jk}(\beta^*_{j\vee k})$ and $\hat{S}_{jk}$ from (9) and (11), respectively. With some abuse of notation, throughout this proof we write $\hat{\beta}_{j,-k} = (0, \hat{\beta}_{j,O\setminus\{j,k\}})$ and $\beta^*_{j,-k} = (0, \beta^*_{j,O\setminus\{j,k\}})$. Under the null hypothesis $H_0 : \beta^*_{jk} = \beta^*_{kj} = 0$, we have that $\hat{S}_{jk} - S_{jk}(\beta^*_{j\vee k}) = I_{1j} + I_{2j} + I_{1k} + I_{2k}$, where I_{1j} and I_{2j} are defined as

$$\begin{aligned} I_{1j} &= \nabla_k \ell_j(\hat{\beta}_{j,-k}) - \nabla_k \ell_j(\beta^*_{j,-k}) - \hat{w}^{\mathrm{T}}_{jk} \big\{ \nabla_{-k} \ell_j(\hat{\beta}_{j,-k}) - \nabla_{-k} \ell_j(\beta^*_{j,-k}) \big\}, \\ I_{2j} &= (w^*_{jk} - \hat{w}_{jk})^{\mathrm{T}} \nabla_{-k} \ell_j(\beta^*_{j,-k}). \end{aligned}$$

The terms I_{1k} and I_{2k} are defined similarly by interchanging the subscripts j and k in I_{1j} and I_{2j}. The goal is to show that each of the four terms is $o_{\mathbb{P}}(n^{-1/2})$.

Upper bound for I_{1j}. Let $\hat{\Delta}_{j,-k} = \hat{\beta}_{j,-k} - \beta^*_{j,-k}$. By an application of the mean value theorem, there exists $\tilde{\beta}_{j,-k} \in \mathbb{R}^{p-1}$ on the line segment between $\hat{\beta}_{j,-k}$ and $\beta^*_{j,-k}$ such that

$$I_{1j} = \big\{ \nabla^2_{k,-k} \ell_j(\tilde{\beta}_{j,-k}) - \hat{w}^{\mathrm{T}}_{jk} \nabla^2_{-k,-k} \ell_j(\tilde{\beta}_{j,-k}) \big\} \hat{\Delta}_{j,-k}.$$

By the triangle inequality and Hölder's inequality, we have

$$\begin{aligned} |I_{1j}| &\leqslant \Big\| \nabla^2_{k,-k} \ell_j(\hat{\beta}_{j,-k}) - \hat{w}^{\mathrm{T}}_{jk} \nabla^2_{-k,-k} \ell_j(\hat{\beta}_{j,-k}) \Big\|_\infty \|\hat{\Delta}_{j,-k}\|_1 \\ &\quad + \Big\| \nabla^2_{k,-k} \ell_j(\hat{\beta}_{j,-k}) - \nabla^2_{k,-k} \ell_j(\tilde{\beta}_{j,-k}) \Big\|_\infty \|\hat{\Delta}_{j,-k}\|_1 \\ &\quad + \Big\| \hat{w}^{\mathrm{T}}_{jk} \big\{ \nabla^2_{-k,-k} \ell_j(\hat{\beta}_{j,-k}) - \nabla^2_{-k,-k} \ell_j(\tilde{\beta}_{j,-k}) \big\} \Big\|_\infty \|\hat{\Delta}_{j,-k}\|_1 \\ &= I_{1j1} + I_{1j2} + I_{1j3}. \end{aligned}$$

We now obtain an upper bound for each of the three terms separately.

By Theorem 1, the definition (10) of a Dantzig selector-type estimator, and Assumption S1, we have

$$I_{1j1} = \mathcal{O}_{\mathbb{P}} \left(\frac{s_{\max} \lambda \lambda_w}{\kappa^2_{\min}} \right) = o_{\mathbb{P}}(n^{-1/2}).$$

By the triangle inequality,

$$\begin{aligned} I_{1j2} &= \left\|\nabla^2_{k,-k}\ell_j(\hat{\beta}_{j,-k}) - \nabla^2_{k,-k}\ell_j(\tilde{\beta}_{j,-k})\right\|_\infty \|\hat{\Delta}_{j,-k}\|_1 \\ &\leqslant \left\|\nabla^2_{k,-k}\ell_j(\hat{\beta}_{j,-k}) - \nabla^2_{k,-k}\ell_j(\beta^*_{j,-k})\right\|_\infty \|\hat{\Delta}_{j,-k}\|_1 \\ &\quad + \left\|\nabla^2_{k,-k}\ell_j(\tilde{\beta}_{j,-k}) - \nabla^2_{k,-k}\ell_j(\beta^*_{j,-k})\right\|_\infty \|\hat{\Delta}_{j,-k}\|_1 \\ &= \mathcal{O}_{\mathbb{P}}\left(\frac{Ms_{\max}\lambda}{\kappa^2_{\min}}\right)\mathcal{O}_{\mathbb{P}}\left(\frac{s_{\max}\lambda}{\kappa^2_{\min}}\right), \end{aligned}$$

where the last equality follows from the proof of Lemma S7 and Theorem 1. We can write the last expression as $I_{ij2} = \mathcal{O}_{\mathbb{P}}(s_{\max}\lambda\lambda_w/\kappa^2_{\min}) = o_{\mathbb{P}}(n^{-1/2})$, since $\lambda_w = \Omega\{w_0(Ms_{\max}\lambda/\kappa^2_{\min} + \lambda\log^2 p)\}$ by Assumption S1.

By the definition (10) of the Dantzig selector-type estimator, $\|\hat{w}_{jk}\|_1 \leqslant \|w^*_{jk}\|_1$. Therefore, by the triangle inequality and Hölder's inequality, we obtain

$$\begin{aligned} I_{1j3} &= \left\|\hat{w}^{\mathrm{T}}_{jk}\{\nabla^2_{-k,-k}\ell_j(\hat{\beta}_{j,-k}) - \nabla^2_{-k,-k}\ell_j(\tilde{\beta}_{j,-k})\}\right\|_\infty \|\hat{\Delta}_{j,-k}\|_1 \\ &\leqslant \|w^*_{jk}\|_1 \left\|\nabla^2_{-k,-k}\ell_j(\hat{\beta}_{j,-k}) - \nabla^2_{-k,-k}\ell_j(\tilde{\beta}_{j,-k})\right\|_\infty \|\hat{\Delta}_{j,-k}\|_1 \\ &= \|w^*_{jk}\|_1\mathcal{O}_{\mathbb{P}}\left(\frac{Ms_{\max}\lambda}{\kappa^2_{\min}}\right)\mathcal{O}_{\mathbb{P}}\left(\frac{s_{\max}\lambda}{\kappa^2_{\min}}\right) \\ &= \mathcal{O}_{\mathbb{P}}\left(\frac{s_{\max}\lambda\lambda_w}{\kappa^2_{\min}}\right) \\ &= o_{\mathbb{P}}(n^{-1/2}), \end{aligned}$$

where the equalities hold by Assumption S1.

Upper bound for I_{2j}. By Lemmas S2–S9 and an application of the Hölder's inequality, we obtain

$$|I_{2j}| \leqslant \|w^*_{jk} - \hat{w}_{jk}\|_1\|\nabla_{-k}\ell_j(\beta^*_{j,-k})\|_\infty = \mathcal{O}_{\mathbb{P}}(s'_{\max}\lambda_w)\mathcal{O}_{\mathbb{P}}(\lambda) = o_{\mathbb{P}}(n^{-1/2}),$$

where the last equality holds by Assumption S1.

Combining the upper bounds for I_{1j} and I_{2j}, we obtain $I_{1j} + I_{2j} = o_{\mathbb{P}}(n^{-1/2})$. Similarly, $I_{1k} + I_{2k} = o_{\mathbb{P}}(n^{-1/2})$. Thus, $n^{1/2}\{\hat{S}_{jk} - S_{jk}(\beta^*_{j\vee k})\} = o_{\mathbb{P}}(1)$. Asymptotic normality of $n^{1/2}\hat{S}_{jk}$ is established by an application of Lemma S6.

S7·6. *Proof of Lemma* S11

Recall from (14) and (15) the definitions of $g^{jk}_i(\beta_{j\vee k})$ and Σ^{jk}, respectively. Also, recall from (16) that

$$\hat{\Sigma}^{jk}(0, \hat{\beta}_{j,O\setminus\{j,k\}}, \hat{\beta}_{k,O\setminus\{j,k\}}) = \frac{1}{n}\sum_{i=1}^n\{g^{jk}_i(0, \hat{\beta}_{j,O\setminus\{j,k\}}, \hat{\beta}_{k,O\setminus\{j,k\}})\}^{\otimes 2}.$$

For notational convenience, we write $\hat{\beta}'_{j\vee k} = (0, \hat{\beta}_{j,O\setminus\{j,k\}}, \hat{\beta}_{k,O\setminus\{j,k\}})$.

By the triangle inequality, we obtain

$$\begin{aligned} \left\|\hat{\Sigma}^{jk}(\hat{\beta}'_{j\vee k}) - \Sigma^{jk}\right\|_\infty &\leqslant \left\|\hat{\Sigma}^{jk}(\beta^*_{j\vee k}) - \Sigma^{jk}\right\|_\infty + \left\|\hat{\Sigma}^{jk}(\hat{\beta}'_{j\vee k}) - \hat{\Sigma}^{jk}(\beta^*_{j\vee k})\right\|_\infty \\ &= I_1 + I_2. \end{aligned}$$

We now obtain an upper bound for each term separately.

Upper bound for I_1. For notational simplicity, throughout the proof we suppress the superscript jk and write g_i^* to indicate $g_i^{jk}(\beta^*_{j\vee k})$. We also write U^j_{ik} for $U^j_{ik}(\beta^*_{j,O\setminus j})$, where $U^j_{ik}(\beta^*_{j,O\setminus j})$ is as defined in (S10). Note that

$$\hat{\Sigma}^{jk}(\beta^*_{j\vee k}) - \Sigma^{jk} = \frac{1}{n}\sum_{i=1}^{n}\left[(g_i^*)^{\otimes 2} - E\{(g_i^*)^{\otimes 2}\}\right],$$

where $(g_i^*)^{\otimes 2}$ takes the form

$$(g_i^*)^{\otimes 2} = \begin{Bmatrix} (U^j_{ik}+U^k_{ij})^2 & (U^j_{ik}+U^k_{ij})(U^j_{i,-k})^{\mathrm{T}} & (U^j_{ik}+U^k_{ij})(U^k_{i,-j})^{\mathrm{T}} \\ U^j_{i,-k}(U^j_{ik}+U^k_{ij}) & (U^j_{i,-k})^{\otimes 2} & U^j_{i,-k}(U^k_{i,-j})^{\mathrm{T}} \\ U^k_{i,-j}(U^j_{ik}+U^k_{ij}) & U^k_{i,-j}(U^j_{i,-k})^{\mathrm{T}} & (U^k_{i,-j})^{\otimes 2} \end{Bmatrix} \in \mathbb{R}^{(2p-3)\times(2p-3)}. \tag{S45}$$

We see from (S45) that each term in $(g_i^*)^{\otimes 2} - E\{(g_i^*)^{\otimes 2}\}$ involves the quantity

$$U^j_{il}U^k_{im} - E(U^j_{il}U^k_{im}) \tag{S46}$$

for $j \neq l$, $k \neq m$, and $j, k, l, m \in O$.

We now show that (S46) is bounded with high probability. First, note that $\|E(U^j_i U^k_i)\|_\infty \leqslant C_u$ for some sufficiently large constant C_u, since $E(X_j^4)$ is bounded by Proposition S1. Hence, by the union bound, for any $t > 2\|E(U^j_i U^k_i)\|_\infty$ we have

$$\begin{aligned}\mathrm{pr}\left\{\left|U^j_{il}U^k_{im} - E(U^j_{il}U^k_{im})\right| \geqslant t\right\} &\leqslant \mathrm{pr}\left(|U^j_{il}U^k_{im}| \geqslant t/2\right) \\ &\leqslant \mathrm{pr}\{|U^j_{il}| \geqslant (t/2)^{1/2}\} + \mathrm{pr}\{|U^k_{im}| \geqslant (t/2)^{1/2}\} \\ &\leqslant 8R^2c_1\exp(-t^{1/4}2^{-5/4}),\end{aligned}$$

where the last inequality is obtained by an application of Lemma S1. Note that the above inequality holds only if $t > 2C_u$. Choosing $C_H = \max\{8R^2c_1, \exp(2^{-1}C_u^{1/4})\}$, we have

$$\mathrm{pr}\left\{\left|U^j_{il}U^k_{im} - E(U^j_{il}U^k_{im})\right| \geqslant t\right\} \leqslant C_H\exp(-t^{1/4}2^{-5/4})$$

for all $t > 0$. Therefore, for any $j \neq l$, $k \neq m$ and $j, k, l, m \in O$, we have

$$\begin{aligned}&\mathrm{pr}\left\{\left|\hat{\Sigma}^{jk}(\beta^*_{j\vee k}) - \Sigma^{jk}\right|_{jl,km} \geqslant t\right\} \\ &\leqslant 4\,\mathrm{pr}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\left\{U^j_{il}U^k_{im} - E(U^j_{il}U^k_{im})\right\}\right| \geqslant \frac{t}{4}\right] \\ &\leqslant 16\exp\left(-\frac{1}{8}4^{-2/9}n^{1/9}t^{2/9}\right) + 16nC_H\exp\left(-\frac{1}{8}4^{-2/9}n^{1/9}t^{2/9}\right),\end{aligned}$$

where the first inequality is obtained by an application of the union bound and the last inequality is obtained by an application of Lemma S14 with $L_1 = C_H$, $L_2 = 2^{-5/4}$ and $q = 1/4$. Hence, by the union bound, we obtain

$$\mathrm{pr}\left(\left\|\hat{\Sigma}^{jk}(\beta^*_{j\vee k}) - \Sigma^{jk}\right\|_\infty \geqslant t\right)$$

$$\leqslant 64p^2 \exp\left(-\frac{1}{8}4^{-2/9}n^{1/9}t^{2/9}\right) + 64np^2C_H \exp\left(-\frac{1}{8}4^{-2/9}n^{1/9}t^{2/9}\right).$$

The above inequality holds only if $t \geqslant [E\{U_{il}^j U_{im}^k - E(U_{il}^j U_{im}^k)\}^2/n]^{1/2}$. By Proposition S1, the numerator is bounded since $E(X_j^8)$ is bounded. Taking $t = K_2(\log^9 p/n)^{1/2}$ for sufficiently large $K_2 > 0$, we have $\mathrm{pr}\{\|\hat{\Sigma}^{jk}(\beta^*_{j\vee k}) - \Sigma^{jk}\|_\infty \geqslant t\} \leqslant p^{-1}$. We conclude that

$$I_1 = \left\|\hat{\Sigma}^{jk}(\beta^*_{j\vee k}) - \Sigma^{jk}\right\|_\infty = \mathcal{O}_{\mathbb{P}}\big\{(\log^9 p/n)^{1/2}\big\}.$$

Upper bound for I_2. For notational convenience, we let $G^i(\beta_{j\vee k}) = \{g_i^{jk}(\beta_{j\vee k})\}^{\otimes 2} \in \mathbb{R}^{(2p-3)\times(2p-3)}$. From (S45), we see that for any $(a,b),(c,d) \in \{(p,q) : p,q \in O, p \neq q\}$, $G^i_{ab,cd}(\beta_{j\vee k})$ is a linear combination of $U^j_{ik}(\beta_{j,O\setminus k})U^k_{ij}(\beta_{k,O\setminus j})$ and $\{U^j_{ik}(\beta_{j,O\setminus k})\}^2$. It can be shown that $\partial U^j_{ik}(\beta_{j,O\setminus k})U^k_{ij}(\beta_{k,O\setminus j})/\partial\beta_{jk}$ involves the term $(X^r_{ij} - X^{r'}_{ij})^3(X^r_{ik} - X^{r'}_{ik})^3$.

Since $\max_{i,j,r} X^r_{ij} = \mathcal{O}_{\mathbb{P}}(\log p)$ by Proposition S1, there exists a constant C_v such that

$$\|\nabla G^i_{ab,cd}(\beta_{j\vee k})\|_\infty \leqslant C_v \log^6 p$$

for all $j,k \in O$ with $j \neq k$ and for any $(a,b),(c,d) \in \{(p,q) : p,q \in O, p \neq q\}$, with high probability. By the mean value theorem, there exists a $\tilde{\beta}_{j\vee k}$ on the line segment between $\hat{\beta}_{j\vee k}$ and $\beta^*_{j\vee k}$ such that

$$\begin{aligned} G^i_{ab,cd}(\hat{\beta}_{j\vee k}) - G^i_{ab,cd}(\beta^*_{j\vee k}) &= \nabla G^i_{ab,cd}(\tilde{\beta}_{j\vee k})(\hat{\beta}_{j\vee k} - \beta^*_{j\vee k}) \\ &\leqslant \left\|\nabla G^i_{ab,cd}(\tilde{\beta}_{j\vee k})\right\|_\infty \left\|\hat{\beta}_{j\vee k} - \beta^*_{j\vee k}\right\|_1 \\ &= \mathcal{O}_{\mathbb{P}}\big(s_{\max}\lambda \log^6 p/\kappa^2_{\min}\big), \end{aligned}$$

where the last equality holds by an application of Theorem 1. Thus,

$$\|I_2\|_\infty = \mathcal{O}_{\mathbb{P}}(s_{\max}\lambda \log^6 p/\kappa^2_{\min}).$$

The result is obtained by combining the upper bounds for I_1 and I_2.

S7·7. *Proof of Lemma* S12

Recall from (15) and (16) the definitions of Σ^{jk} and $\hat{\Sigma}^{jk}$, respectively. Let

$$\sigma^2_{jk} = \Sigma^{jk}_{jk,jk} - 2\Sigma^{jk}_{jk,j\setminus k}w^*_{jk} - 2\Sigma^{jk}_{jk,k\setminus j}w^*_{kj} + (w^*_{jk})^{\mathrm{T}}\Sigma^{jk}_{j\setminus k,j\setminus k}w^*_{jk} + (w^*_{kj})^{\mathrm{T}}\Sigma^{jk}_{k\setminus j,k\setminus j}w^*_{kj}$$

and

$$\hat{\sigma}^2_{jk} = \hat{\Sigma}^{jk}_{jk,jk} - 2\hat{\Sigma}^{jk}_{jk,j\setminus k}\hat{w}_{jk} - 2\hat{\Sigma}^{jk}_{jk,k\setminus j}\hat{w}_{kj} + \hat{w}^{\mathrm{T}}_{jk}\hat{\Sigma}^{jk}_{j\setminus k,j\setminus k}\hat{w}_{jk} + \hat{w}^{\mathrm{T}}_{kj}\hat{\Sigma}^{jk}_{k\setminus j,k\setminus j}\hat{w}_{kj}.$$

To show that $|\hat{\sigma}^2_{jk} - \sigma^2_{jk}| = o_{\mathbb{P}}(1)$, we use the results from Lemmas S9 and S11 and the fact that $\|\Sigma^{jk}\|_\infty = \mathcal{O}(1)$ since $E(X_j^4)$ is bounded by Proposition S1. For notational simplicity, we suppress the superscripts in $\hat{\Sigma}^{jk}$ and Σ^{jk} in the following proof.

By the triangle inequality,

$$\begin{aligned} \left|\hat{\sigma}^2_{jk} - \sigma^2_{jk}\right| \leqslant{}& \left|\hat{\Sigma}_{jk,jk} - \Sigma_{jk,jk}\right| + 2\left|\hat{\Sigma}_{jk,j\setminus k}\hat{w}_{jk} - \Sigma_{jk,j\setminus k}w^*_{jk}\right| + 2\left|\hat{\Sigma}_{jk,k\setminus j}\hat{w}_{kj} - \Sigma_{jk,k\setminus j}w^*_{kj}\right| \\ &+ \left|\hat{w}^{\mathrm{T}}_{jk}\hat{\Sigma}_{j\setminus k,j\setminus k}\hat{w}_{jk} - (w^*_{jk})^{\mathrm{T}}\Sigma_{j\setminus k,j\setminus k}w^*_{jk}\right| \\ &+ \left|\hat{w}^{\mathrm{T}}_{kj}\hat{\Sigma}_{k\setminus j,k\setminus j}\hat{w}_{kj} - (w^*_{kj})^{\mathrm{T}}\Sigma_{k\setminus j,k\setminus j}w^*_{kj}\right| \end{aligned}$$

$$= I_1 + I_{2j} + I_{2k} + I_{3j} + I_{3k}.$$

Upper bound for I_1. By Lemma S11, we obtain

$$I_1 = \mathcal{O}_{\mathbb{P}}\left\{\frac{s_{\max}}{\kappa_{\min}^2}\lambda\log^6 p + \left(\frac{\log^9 p}{n}\right)^{1/2}\right\} = \mathcal{O}_{\mathbb{P}}\left(\frac{s_{\max}}{\kappa_{\min}^2}\lambda\log^6 p\right).$$

Upper bound for I_{2j}. By the triangle inequality,

$$\begin{aligned} I_{2j} &= \left|\hat{\Sigma}_{jk,j\backslash k}\hat{w}_{jk} - \Sigma_{jk,j\backslash k}w^*_{jk}\right| \\ &\leqslant \left|(\hat{\Sigma}_{jk,j\backslash k} - \Sigma_{jk,j\backslash k})(\hat{w}_{jk} - w^*_{jk})\right| + \left|\Sigma_{jk,j\backslash k}(\hat{w}_{jk} - w^*_{jk})\right| + \left|(\hat{\Sigma}_{jk,j\backslash k} - \Sigma_{jk,j\backslash k})w^*_{jk}\right| \\ &= I_{2j1} + I_{2j2} + I_{2j3}. \end{aligned}$$

By Hölder's inequality and Lemmas S9 and S11, we have

$$\begin{aligned} I_{2j1} &\leqslant \|\hat{w}_{jk} - w^*_{jk}\|_1 \|\hat{\Sigma}_{jk,j\backslash k} - \Sigma_{jk,j\backslash k}\|_\infty \\ &= \mathcal{O}_{\mathbb{P}}(s'_{\max}\lambda_w)\mathcal{O}_{\mathbb{P}}\big\{s_{\max}\lambda\log^6 p/\kappa_{\min}^2 + (\log^9 p/n)^{1/2}\big\} \\ &= \mathcal{O}_{\mathbb{P}}\big(s'_{\max}\lambda_w s_{\max}\lambda\log^6 p/\kappa_{\min}^2\big). \end{aligned}$$

By Hölder's inequality, Lemma S9, and the fact that $\|\Sigma\|_\infty = \mathcal{O}(1)$, we obtain

$$I_{2j2} \leqslant \|\Sigma_{jk,j\backslash k}\|_\infty \|\hat{w}_{jk} - w^*_{jk}\|_1 = \mathcal{O}_{\mathbb{P}}(s'_{\max}\lambda_w).$$

Recall that $w_0 = \max_{j,k\in O} \|w^*_{jk}\|_1$. By Hölder's inequality and Lemma S11,

$$I_{2j3} \leqslant \|\hat{\Sigma}_{jk,j\backslash k} - \Sigma_{jk,j\backslash k}\|_\infty \|w^*_{jk}\|_1 = \mathcal{O}_{\mathbb{P}}\big(w_0 s_{\max}\lambda\log^6 p/\kappa_{\min}^2\big).$$

Combining the upper bounds, we obtain

$$\begin{aligned} I_{2j} &= \mathcal{O}_{\mathbb{P}}\big\{(s'_{\max}\lambda_w + w_0)s_{\max}\lambda\log^6 p/\kappa_{\min}^2 + s'_{\max}\lambda_w\big\} \\ &= \mathcal{O}_{\mathbb{P}}\big[\{o(1) + w_0\}s_{\max}\lambda\log^6 p/\kappa_{\min}^2 + o(1)\big] \\ &= \mathcal{O}_{\mathbb{P}}\big(w_0 s_{\max}\lambda\log^6 p/\kappa_{\min}^2\big), \end{aligned}$$

where we have used the fact that $s'_{\max}\lambda_w = o(1)$ by Assumption S1.

Upper bound for I_3. By the triangle inequality,

$$\begin{aligned} I_{3j} &\leqslant \left|\hat{w}_{jk}^{\mathrm{T}}(\hat{\Sigma}_{j\backslash k,j\backslash k} - \Sigma_{j\backslash k,j\backslash k})\hat{w}_{jk}\right| + \left|\hat{w}_{jk}^{\mathrm{T}}\Sigma_{j\backslash k,j\backslash k}\hat{w}_{jk} - (w^*_{jk})^{\mathrm{T}}\Sigma_{j\backslash k,j\backslash k}w^*_{jk}\right| \\ &= I_{3j1} + I_{3j2}. \end{aligned}$$

By the definition of $\hat{w}_{jk}$ in (10), we have $\|\hat{w}_{jk}\|_1 \leqslant \|w^*_{jk}\|_1$. Therefore, by Hölder's inequality and Lemma S11,

$$\begin{aligned} I_{3j1} &\leqslant \|w^*_{jk}\|_1^2 \, \|\hat{\Sigma}_{j\backslash k,j\backslash k} - \Sigma_{j\backslash k,j\backslash k}\|_\infty \\ &= \mathcal{O}_{\mathbb{P}}\big[w_0^2\big\{s_{\max}\lambda\log^6 p/\kappa_{\min}^2 + (\log^9 p/n)^{1/2}\big\}\big] \\ &= \mathcal{O}_{\mathbb{P}}\big(w_0^2 s_{\max}\lambda\log^6 p/\kappa_{\min}^2\big). \end{aligned}$$

By the triangle inequality,

$$\begin{aligned}
I_{3j2} &= \left|\hat{w}_{jk}^{\mathrm{T}}\Sigma_{j\backslash k,j\backslash k}\hat{w}_{jk} - (w_{jk}^*)^{\mathrm{T}}\Sigma_{j\backslash k,j\backslash k}w_{jk}^*\right| \\
&= \left|(\hat{w}_{jk}-w_{jk}^*)^{\mathrm{T}}\Sigma_{j\backslash k,j\backslash k}(\hat{w}_{jk}-w_{jk}^*) + 2(w_{jk}^*)^{\mathrm{T}}\Sigma_{j\backslash k,j\backslash k}(\hat{w}_{jk}-w_{jk}^*)\right| \\
&\leqslant \left|(\hat{w}_{jk}-w_{jk}^*)^{\mathrm{T}}\Sigma_{j\backslash k,j\backslash k}(\hat{w}_{jk}-w_{jk}^*)\right| + 2\left|(w_{jk}^*)^{\mathrm{T}}\Sigma_{j\backslash k,j\backslash k}(\hat{w}_{jk}-w_{jk}^*)\right| \\
&\leqslant \|\hat{w}_{jk}-w_{jk}^*\|_1^2\|\Sigma_{j\backslash k,j\backslash k}\|_\infty + 2\|\Sigma_{j\backslash k,j\backslash k}w_{jk}^*\|_\infty\|\hat{w}_{jk}-w_{jk}^*\|_1 \\
&= \mathcal{O}_{\mathbb{P}}\{(s'_{\max}\lambda_w)^2\} + \mathcal{O}_{\mathbb{P}}\{(s'_{\max}\lambda_w)w_0\} \\
&= \mathcal{O}_{\mathbb{P}}\{(s'_{\max}\lambda_w)w_0\},
\end{aligned}$$

where the last two equalities are obtained by an application of Lemma S9, the fact that $\|\Sigma\|_\infty = \mathcal{O}(1)$, and Assumption S1. Similar upper bounds can be obtained for I_{2k} and I_{3k}.

Upon combining the upper bounds, we have

$$\begin{aligned}
|\hat{\sigma}_{jk}^2 - \sigma_{jk}^2| &\leqslant I_1 + I_{2j} + I_{2k} + I_{3j} + I_{3k} \\
&= \mathcal{O}_{\mathbb{P}}\big\{(1+w_0+w_0^2)s_{\max}\lambda\log^6 p/\kappa_{\min}^2\big\} + \mathcal{O}_{\mathbb{P}}\big(w_0 s'_{\max}\lambda_w\big) \\
&= o_{\mathbb{P}}(1),
\end{aligned}$$

where the last expression holds by Assumption S2.

S8. TAIL INEQUALITIES

In this section, we state some results that are frequently used in our proofs.

LEMMA S13 (GAUSSIAN TAIL INEQUALITY). *Let $X \sim N(0,\sigma^2)$. Then, for $x > 0$,*

$$\mathrm{pr}(|X| \geqslant x) \leqslant \frac{2\sigma}{x}\exp(-x^2/2\sigma^2).$$

LEMMA S14 (LEMMA H.3 IN NING & LIU, 2016). *Let $X_1,\ldots,X_n$ be independent and identically distributed random variables with $E(X_i)=0$. Let $\bar{X}_n = \sum_{i=1}^n X_i/n$. If there exist constants L_1, L_2 and q such that for $x>0$,*

$$\mathrm{pr}(|X_i| \geqslant x) \leqslant L_1\exp(-L_2x^q),$$

then for $x \geqslant \{8E(X_i^2)/n\}^{1/2}$,

$$\mathrm{pr}(|\bar{X}_n| \geqslant x) \leqslant 4\exp\left\{-\frac{1}{8}n^{q/(2+q)}x^{2q/(2+q)}\right\} + 4nL_1\exp\left\{-\frac{L_2n^{q/(2+q)}x^{2q/(2+q)}}{2^q}\right\}.$$

REFERENCES

BICKEL, P. J., RITOV, Y. & TSYBAKOV, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37**, 1705–32.

NING, Y. & LIU, H. (2016). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist.* in press.

NING, Y., ZHAO, T. & LIU, H. (2016). A likelihood ratio framework for high dimensional semiparametric regression. *Ann. Statist.* in press.

VAN DE GEER, S. A. & BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the lasso. *Electron. J. Statist.* **3**, 1360–92.

YANG, E., RAVIKUMAR, P., ALLEN, G. I. & LIU, Z. (2015). Graphical models via univariate exponential family distributions. *J. Mach. Learn. Res.* **16**, 3813–47.

YE, F. & ZHANG, C.-H. (2010). Rate minimaxity of the lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *J. Mach. Learn. Res.* **11**, 3519–40.