Joint Structure Estimation for Categorical Markov Networks

Jian Guo, Elizaveta Levina, George Michailidis and Ji Zhu
Department of Statistics, The University of Michigan, Ann Arbor

April 5, 2010

Abstract

We consider the problem of identifying and estimating non-zero parameters in the Markov model for binary variables. We approximate the full likelihood by a pseudo-likelihood function and propose a joint ℓ_1-penalized logistic regression method, which imposes overall sparsity on the parameters. We show that the proposed method leads to consistent parameter estimation and model selection under high-dimensional asymptotics, and we develop an efficient local quadratic approximation algorithm for computing the estimator. The proposed method is used to explore voting dependencies between senators in the 109th Congress; our analysis confirms known political patterns and provides new insights into the US Senate's voting.

KEY WORDS: Graphical model, Ising model, Lasso, ℓ_1-regularization, Markov network, Pseudo likelihood.

1 Introduction

Undirected graphical models have proved useful in a number of application areas, including bioinformatics (Airoldi, 2007), natural language processing (Jung et al., 1996), image analysis (Li, 2001), and many others, due to their ability to succinctly represent dependence relationships among a set of random variables. Such models represent the relationships between p variables $X_1, \cdots, X_p$ through an undirected graph $G = (V, E)$, whose node set V corresponds to the variables and the edge set E characterizes their pairwise relationships. Specifically, variables X_j and $X_{j'}$ are conditionally independent given all other variables if their associated nodes are not linked by an edge.

Two important types of graphical models are the Gaussian model, where the p variables are assumed to follow a joint Gaussian distribution, and the Markov model, which captures relationships between categorical variables. In the former, the structure of the underlying graph can be recovered by estimating the corresponding inverse covariance (precision) matrix, whose off-diagonal elements are proportional to the partial correlations between the variables. A large body of literature has emerged over the past few years addressing this issue, especially for sparse networks. A number of methods focus on estimating a sparse inverse covariance matrix and inferring the network from estimated zeros (Banerjee et al., 2008; Yuan and Lin, 2007; Rothman et al., 2008; Friedman et al., 2008; Lam and Fan, 2009; Rocha et al., 2008; Ravikumar et al., 2008; Peng et al., 2009). Another class of methods focuses on estimating the network directly without first estimating the precision matrix (Drton and Perlman, 2004; Meinshausen and Buhlmann, 2006). There is also some recent literature on directed acyclic graphical models (see, for example, Shojaie and Michailidis (2010) and references therein).

For the Markov model, the estimation problem is significantly harder, since it is computationally infeasible for any realistic size network to directly evaluate the likelihood, due to the intractable constant (the log-partition function). Several methods in the literature over-

come this difficulty by employing computationally tractable approximations. For example, d'Aspremont et al. (2008) proposed estimating the network structure using an ℓ_1-penalized surrogate likelihood, where the log-partition function is approximated by a log-determinant relaxation. Kolar and Xing (2008) improved on this method by incorporating a cutting-plane algorithm to obtain a tighter outer bound on the marginal polytope.

Alternatively, Ravikumar et al. (2010) proposed a neighborhood selection method that approximates the likelihood by a pseudo-likelihood function, in analogy to the Meinshausen and Buhlmann (2006) method for Gaussian graphical models, where p individual ℓ_1-penalized regressions were fitted, regressing each variable on all others, and the network structure was recovered from the regression coefficients. Ravikumar et al. (2010) separately fit p individual penalized logistic regressions, whose coefficients are used to recover the Markov network structure. They also showed that the neighborhood selection method satisfies both estimation consistency and model selection consistency. Recently, Höefling and Tibshirani (2009) used this algorithm to iteratively approximate the full likelihood by a series of pseudo-likelihoods estimated by the neighborhood selection method.

The neighborhood selection method, in spite of its established asymptotic properties, has certain disadvantages in practice. It is impractical to search for an optimal combination of p different tuning parameters for each of the p regressions, and tuning each regression separately can lead to numerical instability, as shown by Peng et al. (2009) in the Gaussian case. Thus, the same tuning parameter λ is used in practice, which implies that implicitly the method assumes a "homogeneous" network; i.e. the same degree of sparsity is encouraged for all nodes. However, there are many real networks with a "hub" like structure (few nodes possessing very high degrees (Barabasi and Albert, 1999)) that would be a challenge for the neighborhood selection method. In addition, a consequence of estimating pairwise interactions by fitting p separate logistic regression is lack of symmetry; the estimate of interaction between X_i and X_j may have a different value and even a different sign from the interac-

tion between X_j and X_i. An ad hoc symmetrizing procedure was employed by Ravikumar et al. (2010) after fitting the separate regressions, but a better solution would be to estimate all the regressions jointly. We propose a JOint Structure Estimation method (JOSE) that simultaneously solves the p logistic regression problems and encourages the sparsity of the interaction parameters, thus automatically ensuring symmetry. The joint application of the ℓ_1 penalty allows for a more flexible degree distribution in the estimated graph, as explained in Section 2. Our proposal is related to the method of Peng et al. (2009) for Gaussian graphical models, but is significantly more challenging to analyze and implement due to the more complicated Markov model structure. We show that the proposed algorithm leads to consistent parameter estimation and model selection under high-dimensional asymptotics, and develop an efficient local quadratic approximation algorithm for computing the estimator. We have also recently become aware of a simultaneous and independent effort to develop a similar algorithm by Wang et al. (2009); however, that paper does not provide any theoretical analysis and focuses on very different biological applications.

The remainder of the paper is organized as follows. Section 2 presents the JOSE method for binary Markov models and discusses algorithmic issues. Section 3 establishes the theoretical properties of the JOSE estimator, including consistency of parameter estimation and network recovery. Section 4 evaluates the performance of the JOSE method by simulation. Section 5 applies the JOSE method to explore voting dependencies between senators in the 109th Congress. An extension to Markov models with general categorical variables is discussed in Section 6.

2 Methodology

We focus initially on a Markov model for binary variables (henceforth called the Ising model) and discuss the extension to general categorical variables in Section 6. We start by setting up the problem and also discuss the neighborhood selection criterion (Ravikumar et al.,

2010), whose shortcomings we overcome through the proposed JOSE method.

2.1 Problem Setup and Neighborhood Selection

Suppose we have p binary random variables $X_1, \ldots, X_p$, with $X_j \in \{1, 0\}$, $1 \leq j \leq p$, whose joint distribution has the following density function:

$$\mathrm{f}(X_1, \ldots, X_p) = \frac{1}{\mathrm{Z}(\boldsymbol{\Theta})} \exp\Big(\sum_{j=1}^{p} \theta_{j,j} X_j + \sum_{1 \leq j < j' \leq p} \theta_{j,j'} X_j X_{j'} \Big), \tag{1}$$

where $\boldsymbol{\Theta} = (\theta_{j,j'})_{p \times p}$ is a symmetric matrix specifying the network structure.

Note that $\theta_{j,j}$, $1 \leq j \leq p$, corresponds to the main effect for variable X_j, whereas $\theta_{j,j'}$, $1 \leq j < j' \leq p$, corresponds to the interaction effect between variables X_j and $X_{j'}$. These $\theta_{j,j'}$'s reflect the structure of the underlying network. Specifically, if $\theta_{j,j'} = 0$, then X_j and $X_{j'}$ are conditionally independent given other variables and hence their corresponding nodes are *not* connected. Ravikumar et al. (2010) pointed out that one could consider *only* the pairwise interaction effects, since higher order interactions can be converted to pairwise ones through the introduction of additional variables and thus retaining the Markovian structure of the network (Wainwright and Jordan, 2008). The partition function $\mathrm{Z}(\boldsymbol{\Theta}) = \sum_{X_j \in \{0,1\}, 1 \leq j \leq p} \exp(\sum_{j=1}^{p} \theta_{j,j} X_j + \sum_{1 \leq j < j' \leq p} \theta_{j,j'} X_j X_{j'})$ ensures that the density function in (1) is a proper one, integrating to one.

The structure of the partition function with its 2^p terms renders optimizing (1) infeasible, except in toy problems. A strategy to overcome this difficulty is to use the pseudo-likelihood function to approximate the joint likelihood function associated with density (1). Specifically, let $x_{i,j}$ be the i-th realization of variable X_j, then the pseudo-likelihood function can be written as follows:

$$\prod_{j=1}^{p} \prod_{i=1}^{n} \phi_{i,j}^{x_{i,j}} (1 - \phi_{i,j})^{1 - x_{i,j}}, \tag{2}$$

where $\phi_{i,j} = \mathrm{P}(x_{i,j} = 1 | x_{i,k}, k \neq j; \theta_{j,k}, 1 \leq k \leq p) = \exp(\theta_{j,j} + \sum_{k \neq j} \theta_{j,k} x_{i,k}) / \{1 + \exp(\theta_{j,j} + \sum_{k \neq j} \theta_{j,k} x_{i,k})\}$. It can be seen that this gives rise to a logistic regression problem where

the j-th variable is taken as the response and is regressed on the remaining variables, and hence decomposes the problem into p separate logistic regressions, which are simple to solve. Ravikumar et al. (2010), in order to recover a *sparse* network structure, consider a penalized version of (2) in log-scale. Specifically, for each $1 \leq j \leq p$, they optimize

$$\max_{\{\theta_{j,k}\}_{k=1}^{p}} \sum_{i=1}^{n} \Big[x_{i,j} \Big(\theta_{j,j} + \sum_{k \neq j} \theta_{j,k} x_{i,k} \Big) - \log \Big\{ 1 + \exp \Big(\theta_{j,j} + \sum_{k \neq j} \theta_{j,k} x_{i,k} \Big) \Big\} \Big] - \lambda_j \sum_{k \neq j} |\theta_{j,k}|, \quad (3)$$

where λ_j is a tuning parameter controlling the number of neighbors associated with the j-th node. They call their procedure the neighborhood selection method, which can be implemented efficiently using a coordinate descent algorithm. As discussed in the Introduction, since the p regularized logistic regressions are fitted separately, the resulting estimates $\widehat{\theta}_{j,j'}$ and $\widehat{\theta}_{j',j}$ will usually be different, especially in the presence of small sample sizes. Kolar and Xing (2008) introduced two rules to aggregate these estimates:

Maximum aggregation rule:

$$\widehat{\theta}^{\max}_{j,j'} = \widehat{\theta}^{\max}_{j',j} = \begin{cases} \widehat{\theta}_{j,j'}, & \text{if } |\widehat{\theta}_{j,j'}| > |\widehat{\theta}_{j',j}|; \\ \widehat{\theta}_{j',j}, & \text{if } |\widehat{\theta}_{j,j'}| \leq |\widehat{\theta}_{j',j}|. \end{cases} \quad (4)$$

Minimum aggregation rule:

$$\widehat{\theta}^{min}_{j,j'} = \widehat{\theta}^{min}_{j',j} = \begin{cases} \widehat{\theta}_{j,j'}, & \text{if } |\widehat{\theta}_{j,j'}| < |\widehat{\theta}_{j',j}|; \\ \widehat{\theta}_{j',j}, & \text{if } |\widehat{\theta}_{j,j'}| \geq |\widehat{\theta}_{j',j}|. \end{cases} \quad (5)$$

The neighborhood selection estimators aggregated by formula (4) and (5) are referred to as NS-MAX and NS-MIN, respectively.

2.2 Joint Structure Estimation

In the proposed JOSE method, we solve the following joint criterion problem:

$$\begin{aligned} \max_{\boldsymbol{\Theta}} \quad & \sum_{j=1}^{p} \sum_{i=1}^{n} \Big[x_{i,j} \Big(\theta_{j,j} + \sum_{k \neq j} \theta_{j,k} x_{i,k} \Big) \\ & - \log \Big\{ 1 + \exp \Big(\theta_{j,j} + \sum_{k \neq j} \theta_{j,k} x_{i,k} \Big) \Big\} \Big] - \lambda \sum_{j<j'} |\theta_{j,j'}| \\ \text{subject to} \quad & \theta_{j,j'} = \theta_{j',j}, \quad 1 \leq j < j' \leq p. \end{aligned} \quad (6)$$

Notice that the penalty *jointly* imposes sparsity over all interaction effects, while the tuning parameter λ controls its degree. However, the JOSE method does not lead to solving p separate logistic problems due to the symmetry constraint $\theta_{j,j'} = \theta_{j',j}$. On the other hand, it reduces the number of parameters to be estimated by half, i.e., $p(p+1)/2$ for the JOSE method vs. p^2 for the neighborhood selection method.

We present next an algorithm to optimize the objective function in (6). It consists of two nested loops. In the outer loop, we follow the strategy in Friedman et al. (2010) to approximate the logistic log-likelihood in (6) by its Taylor series expansion. Specifically, we denote the estimate of $\theta_{j,j'}$ in the t-th iteration by $\theta_{j,j'}^{(t)}$, and write

$$\begin{aligned} & x_{i,j}\Big(\theta_{j,j} + \sum_{k\neq j}\theta_{j,k}x_{i,k}\Big) - \log\Big\{1 + \exp\Big(\theta_{j,j} + \sum_{k\neq j}\theta_{j,k}x_{i,k}\Big)\Big\} \\ \approx\ & -\frac{1}{2}w_{i,j}^{(t)}\Big(y_{i,j}^{(t)} - \theta_{j,j} - \sum_{k\neq j}\theta_{j,k}x_{i,k}\Big)^2 + C_{i,j}^{(t)}\ , \end{aligned} \tag{7}$$

where

$$\begin{aligned} p_{i,j}^{(t)} &= \frac{\exp(\theta_{j,j}^{(t)} + \sum_{k\neq j}\theta_{j,k}^{(t)}x_{i,k})}{1 + \exp(\theta_{j,j}^{(t)} + \sum_{k\neq j}\theta_{j,k}^{(t)}x_{i,k})}\ , \\ y_{i,j}^{(t)} &= \theta_{j,j}^{(t)} + \sum_{k\neq j}\theta_{j,k}^{(t)}x_{i,k} - \frac{p_{i,j}^{(t)} - x_{i,j}}{w_{i,j}^{(t)}}\ , \\ w_{i,j}^{(t)} &= p_{i,j}^{(t)}(1 - p_{i,j}^{(t)})\ , \end{aligned}$$

and $C_{i,j}^{(t)}$ is some constant unrelated to $\boldsymbol{\Theta}$. We define next the following quantities:

$$\begin{aligned} \boldsymbol{\theta} &= (\theta_{1,2}, \ldots, \theta_{j,j'}, \ldots, \theta_{p-1,p})^{\mathsf{T}}\ , \\ \boldsymbol{y}_j^* &= (\sqrt{w_{1,j}^{(t)}}y_{1,j}, \ldots, \sqrt{w_{n,j}^{(t)}}y_{n,j})^{\mathsf{T}}\ , \\ \boldsymbol{y}_j^{**} &= \boldsymbol{y}_j^* - \bar{y}_j, \text{where } \bar{y}_j = \frac{1}{n}\sum_{i=1}^{n}\sqrt{w_{i,j}^{(t)}}y_{i,j}^{(t)}\ , \\ \boldsymbol{x}_j^* &= (\sqrt{w_{1,j}^{(t)}}x_{1,j}, \ldots, \sqrt{w_{n,j}^{(t)}}x_{n,j})^{\mathsf{T}}\ , \\ \boldsymbol{x}_j^{**} &= \boldsymbol{x}_j^* - \bar{x}_j, \text{where } \bar{x}_j = \frac{1}{n}\sum_{i=1}^{n}\sqrt{w_{i,j}^{(t)}}x_{i,j}^{(t)}\ . \end{aligned} \tag{8}$$

We further define an $np \times 1$ column vector

$$\boldsymbol{\mathcal{X}}_{j,j'} = (\mathbf{0}_n{}^\top, \ldots, \mathbf{0}_n{}^\top, \underbrace{\boldsymbol{x}_{j'}^{**\top}}_{j\text{-th block}}, \mathbf{0}_n{}^\top, \ldots, \mathbf{0}_n{}^\top, \underbrace{\boldsymbol{x}_{j}^{**\top}}_{j'\text{-th block}}, \mathbf{0}_n{}^\top, \ldots, \mathbf{0}_n{}^\top)^\top, \tag{9}$$

where $\mathbf{0}_n$ is an n-dimensional column vector of zeros. $\boldsymbol{\mathcal{X}}_{j,j'}$ consists of p blocks of size n, where the j-th block and the j'-th block are $\boldsymbol{x}_{j'}^{**}$ and $\boldsymbol{x}_{j}^{**}$, respectively, and all other blocks are zeros. Finally, let $\boldsymbol{\mathcal{Y}} = (\boldsymbol{y}_1^{**\top}, \ldots, \boldsymbol{y}_p^{**\top})^\top$ (an $np \times 1$ column vector) and $\boldsymbol{\mathcal{X}} = (\boldsymbol{\mathcal{X}}_{1,2}, \ldots, \boldsymbol{\mathcal{X}}_{j,j'}, \ldots, \boldsymbol{\mathcal{X}}_{p-1,p})$ (an $np \times p(p-1)/2$ matrix). Then, (6) can be rewritten as the following lasso problem:

$$\min_{\boldsymbol{\theta}} \frac{1}{2}\|\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|_1. \tag{10}$$

In the inner loop of the algorithm, criterion (10) can be efficiently solved by shooting-type algorithms Friedman et al. (2007)). Letting $\widehat{\boldsymbol{\theta}}$ be the estimate obtained from (10), then for each $1 \le j \le p$, the main effects $\theta_{j,j}$'s in (7) are calculated as follows:

$$\widehat{\theta}_{j,j} = \frac{\bar{y}_j - \sum_{k \ne j} \widehat{\theta}_{j,k}\bar{x}_k}{\frac{1}{n}\sum_{i=1}^n \sqrt{w_{i,j}^{(t)}}}. \tag{11}$$

3 Theoretical Properties

In this section, we present the asymptotic properties of the JOSE estimator; the proofs can be found in the Appendices. Since in the Ising model the structure of the underlying network only depends on the interaction effects, we focus on the variant of the model with no main effects, which gives rise to the criterion

$$\max_{\boldsymbol{\theta}} \sum_{j=1}^{p}\sum_{i=1}^{n}\Big[x_{i,j}\Big(\sum_{j' \ne j}\theta_{j,j'}x_{i,j'}\Big) - \log\Big\{1 + \exp\Big(\sum_{j' \ne j}\theta_{j,j'}x_{i,j'}\Big)\Big\}\Big] - \lambda\sum_{j<j'}|\theta_{j,j'}|, \tag{12}$$

where $\theta_{j,j'} = \theta_{j',j}$, $1 \le j < j' \le p$, and $\boldsymbol{\theta}$ has been defined in Section 2.2.

Let $\boldsymbol{\theta}^0$ be the true value of $\boldsymbol{\theta}$, and let $\boldsymbol{Q}^0$ be the population Fisher information matrix of the model in criterion (12) at $\boldsymbol{\theta}^0$ (refer to Appendix I for details). Let $\boldsymbol{\mathcal{X}}^{(i,j)}$ be the $[(j-1)n+i]$-th row of $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{X}}^{(i)} = (\boldsymbol{\mathcal{X}}^{(i,1)}, \ldots, \boldsymbol{\mathcal{X}}^{(i,p)})^\top$, and let $\boldsymbol{U}^0 = E\boldsymbol{\mathcal{X}}^{(i)\top}\boldsymbol{\mathcal{X}}^{(i)}$. In

addition, let $S = \{(j, j') : \theta^0_{j,j'} \neq 0, 1 \leq j < j' \leq p\}$ be the index set of all nonzero components of $\boldsymbol{\theta}^0$, whose cardinality is denoted by q, and let S^c be the complement of S. Finally, for any matrix $\boldsymbol{W}$ and subsets of row and column indices $\mathcal{U}$ and $\mathcal{V}$, let $\boldsymbol{W}_{\mathcal{U},\mathcal{V}}$ be the matrix consisting of rows $\mathcal{U}$ and columns $\mathcal{V}$ in $\boldsymbol{W}$, and let $\Lambda_{\min}(\cdot)$ and $\Lambda_{\max}(\cdot)$ denote the smallest and largest eigenvalue of a matrix.

Our results rely on the following regularity conditions:

(A) Dependency: There exist positive constants $\tau_{\min}$ and $\tau_{\max}$ such that

$$\Lambda_{\min}(\boldsymbol{Q}^0_{S,S}) \geq \tau_{\min} \quad \text{and} \quad \Lambda_{\max}(\boldsymbol{U}^0_{S,S}) \leq \tau_{\max} \; ; \tag{13}$$

(B) Incoherence: There exists a constant $\tau \in (0, 1)$ such that

$$\|\boldsymbol{Q}^0_{S^c,S}(\boldsymbol{Q}^0_{S,S})^{-1}\|_\infty \leq 1 - \tau \; . \tag{14}$$

Similar conditions have been assumed by Meinshausen and Buhlmann (2006), Ravikumar et al. (2010) and Peng et al. (2009). The most closely related conditions for binary data are those of Ravikumar et al. (2010), but because they fit regressions separately, their conditions are on the $p \times p$ matrices corresponding to the individual regressions, whereas ours are on the $p(p-1)/2 \times p(p-1)/2$ matrices corresponding to all the parameters combined. These conditions can be interpreted as a bound on the amount of dependence (A), and a bound on influence non-neighbors can have on a given node (B). Under these conditions, we establish the following results:

Theorem 1 *(Parameter estimation). Suppose conditions (A) and (B) hold and* $\widehat{\boldsymbol{\theta}}$ *is the maximizer of the JOSE criterion* (12)*. If the tuning parameter* $\lambda = C_\lambda \sqrt{(\log p)/n}$ *for some constant* $C_\lambda > 16(2-\tau)/\tau$ *and if* $n > (4/C)q^3 \log(p)$ *for some constant* $C < \tau^2_{min}\tau^2 / \max\{288(1-\tau)^2, 72\}$*, then with probability tending to 1,*

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\|_2 \leq M\sqrt{\frac{q \log p}{n}} \; , \tag{15}$$

for some constant $M > (2C_\lambda/\tau_{min})[1 + \tau/(8 - 4\tau)]$*.*

Theorem 2 *(Structure estimation). Under conditions of Theorem 1, if we further assume* $\theta^0_{min} = \min_{(j,j')\in S} |\theta^0_{j,j'}| \geq 2M\sqrt{q\log(p)/n}$*, then with probability tending to 1,*

$$\widehat{\theta}_{j,j'} \neq 0 \text{ for all } (j,j') \in S \text{ and } \widehat{\theta}_{j,j'} = 0 \text{ for all } (j,j') \in S^c \ .$$

The proofs of Theorems 1 and 2 are given in Appendix I.

4 Simulated Examples

In this section, we compare the performance of the JOSE method to that of the neighborhood selection method (Ravikumar et al., 2010), whose estimates are further aggregated as NS-MAX (4) and NS-MIN (5), respectively. We use four different types of network structures: a chain, a lattice, a nearest-neighbor and a scale-free network. Each network consists of $p = 100$ nodes. The details of these networks are described below:

Example 1: Chain Network. In this example, we consider a chain network, which connects nodes 1 to p sequentially. Figure 1 (A) illustrates the chain graph.

Example 2: Lattice Network. In this example, we generate a lattice network with $p = 100$ nodes laid out in a 10×10 array on a plane. Each node only connects to its closest neighbors in four directions (east, south, west and north). Figure 1 (B) illustrates the lattice graph.

Example 3: Nearest-neighbor Network. To generate the nearest neighbor networks, we slightly modify the data generating mechanism described in Li and Gui (2006). Specifically, we generate p points randomly on a unit square, calculate all $p(p-1)/2$ pairwise distances, and find the m nearest neighbors of each point in terms of these distances. The nearest neighbor network is obtained by linking any two points that are m-nearest neighbors of each other. The integer m controls the degree of sparsity of the network and the value $m = 5$ was chosen in the simulation study. Figure 1 (C) exhibits one realization of the nearest-neighbor network.

Example 4: Scale-free Network. A scale-free network has a power-law degree distribution and can be simulated by the Barabasi-Albert algorithm (Barabasi and Albert, 1999). A realization of a scale-free network is depicted in Figure 1 (D).

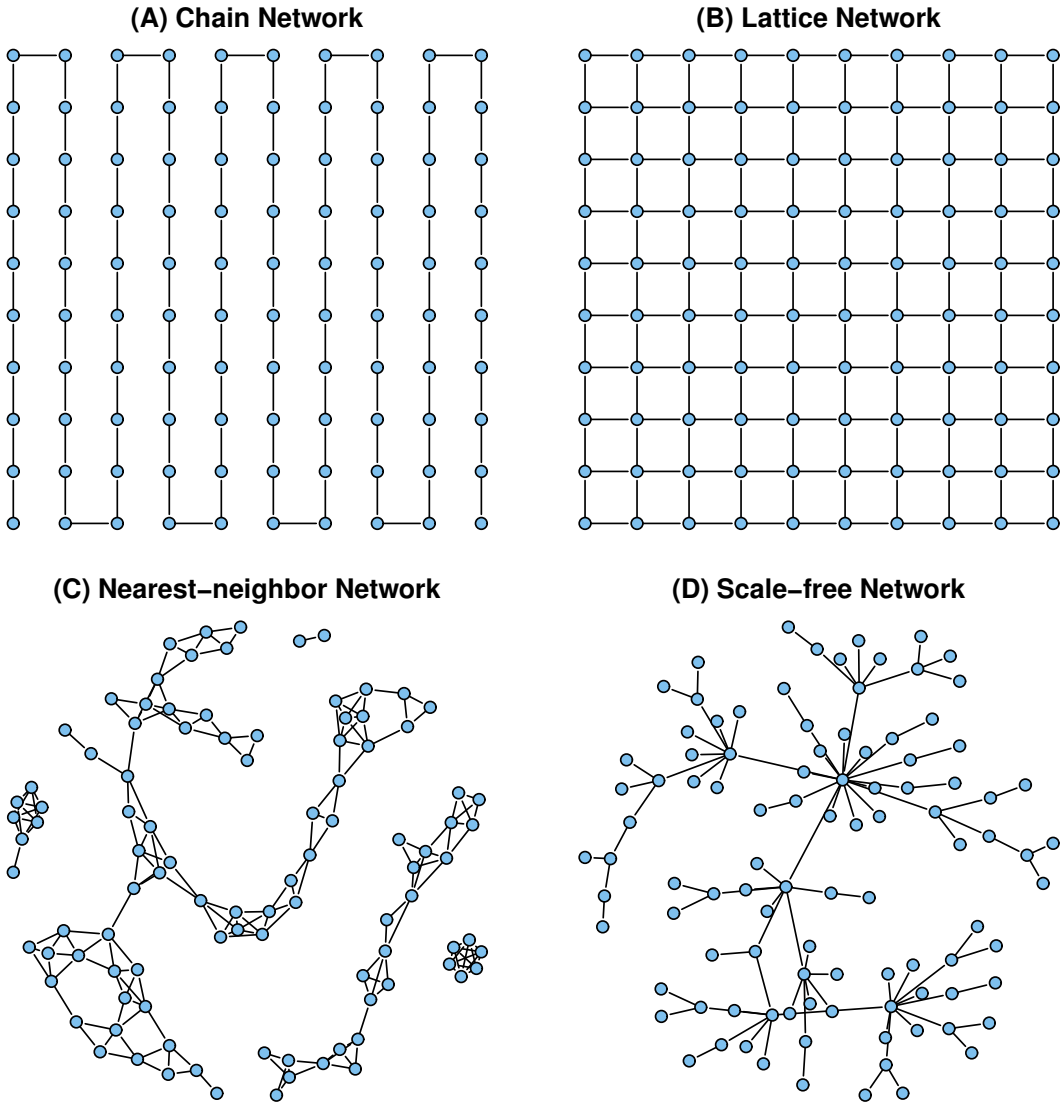

Figure 1: The networks used in simulations: chain, lattice, nearest-neighbor and scale-free networks.

The symmetric parameter matrix $\mathbf{\Theta}$ for each network is generated as follows. Each off-diagonal element $\theta_{j,j'}$ is drawn uniformly from $[-1,-0.5]\cup[0.5,1]$ if nodes j and j' are linked by an edge, otherwise $\theta_{j,j'}=0$. Further, the diagonal elements $\theta_{j,j}$ are drawn uniformly from $[-1,-0.5]\cup[0.5,1]$. Given $\mathbf{\Theta}$, we iteratively generate the data using Gibbs sampling. Specif-

ically, suppose that the MCMC samples from the t-th iteration are available and denoted by $x_1^{(t)}, \ldots, x_p^{(t)}$. Then, in the $(t+1)$-th iteration, $x_j^{(t+1)}$, $1 \leq j \leq p$, is drawn from the following Bernoulli distribution:

$$x_j^{(t+1)} \sim \text{Bernoulli}\Big(\frac{\exp(\theta_{j,j} + \sum_{k \neq j} \theta_{j,k} x_k^{(t)})}{1 + \exp(\theta_{j,j} + \sum_{k \neq j} \theta_{j,k} x_k^{(t)})}\Big). \tag{16}$$

To ensure that the simulated observations are close to i.i.d. from the target distribution, we discard the samples from the first 10^6 iterations (burn-in), and then collect samples every 100 iterations.

The structure estimation results of JOSE, NS-MAX and NS-MIN are represented by ROC curves, which dynamically characterize the sensitivity (proportion of correctly identified edges) and specificity (proportion of correctly excluded edges) by varying the tuning parameter λ. Figure 2 shows the average ROC curves over 50 replications for different sample sizes (n=100, 200 and 500). In all examples, it can be seen that the curves estimated by the JOSE method dominate those based on NS-MAX and NS-MIN. The JOSE method has a more pronounced advantage over its competitors for larger sample sizes. When the false discovery rate (false positive rate) is controlled at the 10% and 5% levels, the power (sensitivity) of the estimates from JOSE are higher than those from the two neighborhood selection estimators (Table 1).

While the ROC curves and the power results measure the performance of the structure estimation, the ℓ_2 loss can be used to characterize the goodness-of-fit in terms of parameter estimation. We define the ℓ_2 loss as $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\| = (\sum_{1 \leq j < j' \leq p} (\widehat{\theta}_{j,j'} - \theta^0_{j,j'})^2)^{1/2}$, where $\widehat{\boldsymbol{\theta}}$ is the estimate of $\boldsymbol{\theta}$ using selected tuning parameter(s) and $\boldsymbol{\theta}^0$ is the true value of $\boldsymbol{\theta}$.

The tuning parameter in JOSE is selected using K-fold cross-validation. Specifically, we randomly split the data set into K subsets (folds) of similar sizes and denote the index set of the observations in the k-th fold ($1 \leq k \leq K$) by $\mathcal{T}_k$. Then, λ is selected by maximizing

the following criterion:

$$\sum_{k=1}^{K}\sum_{j=1}^{p}\sum_{i\in\mathcal{T}_k} x_{i,j}\Big\{\widehat{\theta}_{j,j}^{[-k]}(\lambda)+\sum_{j'\neq j}\widehat{\theta}_{j,j'}^{[-k]}(\lambda)x_{i,j'}\Big\}$$
$$-\log\Big[1+\exp\Big\{\widehat{\theta}_{j,j}^{[-k]}(\lambda)+\sum_{j'\neq j}\widehat{\theta}_{j,j'}^{[-k]}(\lambda)x_{i,j'}\Big\}\Big], \tag{17}$$

where $\widehat{\theta}_{j,j'}^{[-k]}(\lambda)$ is the JOSE estimate of $\theta_{j,j'}$ using all observations except those in the k-th fold and using the tuning parameter λ. These results, summarized in Table 2, show that though JOSE has an advantage in structure estimation, the three methods are comparable in terms of parameter estimation.

Table 1: Power (sensitivity) of the estimates in Examples 1–4 when the false discovery rate (1−specificity) is controlled at 10% and 5%.

Example	n	FDR=10%			FDR=5%		
		JOSE	NS-MAX	NS-MIN	JOSE	NS-MAX	NS-MIN
Chain	100	0.42	0.38	0.40	0.33	0.28	0.31
	200	0.61	0.55	0.60	0.51	0.47	0.48
	500	0.88	0.86	0.86	0.83	0.79	0.81
Lattice	100	0.37	0.34	0.36	0.28	0.24	0.28
	200	0.56	0.50	0.54	0.46	0.41	0.45
	500	0.84	0.76	0.78	0.77	0.69	0.72
Nearest-neighbor	100	0.35	0.31	0.34	0.27	0.23	0.26
	200	0.52	0.45	0.49	0.43	0.36	0.41
	500	0.77	0.67	0.71	0.69	0.61	0.65
Scale-free	100	0.37	0.34	0.35	0.29	0.26	0.27
	200	0.57	0.54	0.53	0.46	0.45	0.42
	500	0.85	0.82	0.80	0.8	0.75	0.75

5 Application to the Senate Voting Record

The dataset was obtained from the website of the US Congress (`http://www.senate.gov`). It contains the voting records of the 100 senators of the 109th Congress (January 3, 2005 — January 3, 2007) on 645 bills, resolutions, motions, debates and roll call votes that the Senate deliberated and voted on. The votes are recorded as one for "yes" and zero

Table 2: ℓ_2 loss in Examples 1–4 (averages over 50 replications with corresponding standard deviations in parentheses).

Example	n	JOSE	NS-MAX	NS-MIN
Chain	100	10.13 (0.12)	9.94 (0.22)	9.91 (0.14)
	200	9.17 (0.24)	9.04 (0.23)	9.07 (0.21)
	500	6.83 (0.21)	6.87 (0.20)	6.91 (0.23)
Lattice	100	13.89 (0.16)	13.63 (0.17)	13.72 (0.14)
	200	12.45 (0.27)	12.38 (0.23)	12.63 (0.21)
	500	9.59 (0.20)	9.72 (0.20)	10.08 (0.17)
Nearest-neighbor	100	14.20 (0.19)	13.99 (0.20)	14.06 (0.14)
	200	12.93 (0.22)	12.88 (0.18)	13.11 (0.18)
	500	10.36 (0.26)	10.51 (0.23)	10.89 (0.22)
Scale-free	100	10.64 (0.13)	10.39 (0.2)	10.37 (0.15)
	200	9.80 (0.18)	9.40 (0.21)	9.59 (0.21)
	500	7.57 (0.22)	7.31 (0.25)	7.58 (0.22)

for "no". Missing values (missed votes) for each senator were imputed with the majority vote of that senator's party on that particular bill; the missing votes for the Independent Senator Jeffords were imputed with the Democratic majority vote. The number of imputed votes is fairly small, less than 5% of the total and less than 3% of the total votes for 90% of the senators, and we do not expect this imputation to have a significant effect on the analysis. Finally, we excluded bills from the analysis if the "yes/no" proportion fell outside the interval $[0.3, 0.7]$, since the Senate votes on many procedural and other uncontroversial motions that do not reflect the real political dynamics in the Senate. This resulted in a total of 387 observations (votes) on 100 variables (senators). The tuning parameter for the JOSE method was selected through cross-validation. The results are shown in Figure 3. A richer structure than that dictated by the presence of two political parties emerges, with four distinct communities, two Republican and two Democratic. As expected, the two political parties are well separated, with many positive dependence links within their members (green solid lines) and negative links across parties (red dashed lines). The two communities on the left side of the plot can be broadly described as representing the

cores of the two parties, although there is additional structure. For example, a number of the more liberal Democrats (Obama, Boxer, Kennedy, Bingaman, Stabenow, Kerry, Lautenberg, Sarbanes, Mikulski, Wyden, Leahy, Dorgan) have the strongest negative associations with the more conservative Republicans (Roberts, Sessions, Hutchison, Coburn, Burr, Shelby, Allen, Cornyn), mostly from Southern states (see also related analysis of earlier congresses in Clinton et al. (2004) and de Leeuw (2006)). Further, a number of positive associations are detected between some of the more centrist Democrats (Lieberman, Nelson, Baucus, Landrieu, Schumer, Clinton); a detailed inspection of the votes suggests that these are mostly due to their positions on issues of national security and the economy. Similarly, there is a separate cluster of moderate Republicans (Grassley, Lugar, Alexander, Warner, Frist, Voinovich). A separate community of Republicans and Democrats emerges on the right side of the plot. An inspection of the votes suggests that they differ from the core members of their respective parties because of their voting record on several issues, including national security, confirmation votes on nominations, and certain regulatory and budget measures. Also of interest is the strong agreement between pairs of senators coming from the same state and party (Schumer-Clinton, Murray-Cantwell, Stevens-Murkowski, Hatch-Bennett, Collins-Snowe). Further, moderate Republicans DeWine, Chafee and Specter and the pro-life Democrat Nelson are represented as isolated nodes, thus confirming results of previous analysis by Clinton et al. (2004) and de Leeuw (2006) (albeit based on data from the 105th Congress). We also note that the Senate voting record from the 109th Congress was analyzed by Banerjee et al. (2008); however, the dataset they used turned out to have been contaminated with many votes from earlier Congresses starting from the 1990s, which led to a large number of missing votes for senators elected later. Since their imputation method was to impute "no" for all missing votes, the validity of their analysis is unclear and their results cannot be directly compared to ours. Overall, our analysis confirms known political patterns and provides new insights into the U.S. Senate's voting.

6 Extension to General Markov Networks

The JOSE method can be extended to model general Markov networks consisting of categorical variables. Let $(x_{i,1}, \ldots, x_{i,p})$ be the i-th observation, where $x_{i,j}$, $1 \le j \le p$, takes values in the discrete set $\{1, 2, \ldots, D\}$ for some positive integer D. Denote by $z_{i,j}^{(1)}, \ldots, z_{i,j}^{(D-1)}$ the dummy variables associated with $x_{i,j}$, i.e., $z_{i,j}^{(d)} = \mathrm{I}(x_{i,j} = d)$, $1 \le d \le D-1$, where $\mathrm{I}(\cdot)$ denotes the indicator function. Notice that we omit $z_{i,j}^{(D)}$ because it is redundant given the constraint $\sum_{d=1}^{D} z_{i,j} = 1$.

The JOSE criterion can be modified as follows:

$$\begin{aligned}
\max_{\{\boldsymbol{\theta}_j^*:1\le j\le p\}\bigcup\{\boldsymbol{\theta}_{j,j'}^*:1\le j<j'\le p\}} \quad & \sum_{j=1}^{p}\sum_{i=1}^{n}\Big[\sum_{d=1}^{D-1} z_{i,j}^{(d)}\Big(\theta_j^{(d)} + \sum_{k\neq j}\sum_{d'=1}^{D-1}\theta_{j,k}^{(d,d')} z_{i,k}^{(d')}\Big) \\
- \quad & \log\Big\{\sum_{d=1}^{D-1}\exp\Big(\theta_j^{(d)} + \sum_{k\neq j}\sum_{d'=1}^{D-1}\theta_{j,k}^{(d,d')} z_{i,k}^{(d')}\Big)\Big\}\Big] \\
- \quad & \lambda\sum_{j<j'}\sqrt{\sum_{d=1}^{D-1}\sum_{d'=1}^{D-1}(\theta_{j,j'}^{(d,d')})^2} \\
\text{subject to} \quad & \theta_{j,j'}^{(d,d')} = \theta_{j',j}^{(d,d')}, \quad 1 \le j < j' \le p, 1 \le d, d' \le D-1. \qquad (18)
\end{aligned}$$

In (18), $\theta_j^{(d)}$ corresponds to the main effect of variable j in class d and $\theta_{j,j'}^{(d,d')}$ to the interaction effect between variable j in class d and variable j' in class d'. Further, $\boldsymbol{\theta}_j^* = \{\theta_j^{(d)} : 1 \le d \le D-1\}$ collects all main effects associated with variable j and $\boldsymbol{\theta}_{j,j'}^* = \{\theta_{j,j'}^{(d,d')} : 1 \le d, d' \le D-1\}$ collects all interaction effects associated with variables j and j'. Here, we remove the edge between nodes j and j' only if *all* the elements in $\boldsymbol{\theta}_{j,j'}^*$ are zero. To achieve this, we use the group penalty proposed by Yuan and Lin (2007), where all elements in $\boldsymbol{\theta}_{j,j'}^*$ are regarded as a group and simultaneously estimated as zeros or nonzeros. Criterion (18) can be estimated by a modified LQA-shooting algorithm, in which the inner loop is replaced by a modified shooting algorithm for group lasso (Friedman et al., 2007).

Acknowledgments

This research is partially supported by NSF grant DMS-0805798 (E. Levina), NIH grant 1RC1CA145444-0110 and MEDC grant GR-687 (G. Michailidis), and NSF grants DMS-0705532 and DMS-0748389 (J. Zhu).

References

Airoldi, E. (2007), "Getting Started in Probabilistic Graphical Models," *PLoS Computational Biology*, 3, e252.

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

Barabasi, A.-L. and Albert, R. (1999), "Emergence of scaling in random networks," *Science*, 286, 509–512.

Clinton, J., Jackman, S., and Rivers, D. (2004), "The statistical analysis of roll call data," *American Political Science Review*, 98, 355–370.

d'Aspremont, A., Banerjee, O., and El Ghaoui, L. (2008), "First-order methods for sparse covariance selection," *SIAM Journal on matrix Analysis and its Applications*, 30, 56–66.

de Leeuw, J. (2006), "Principal component analysis of senate voting patterns," in *Real Data Analysis*, ed. Sawilowski, S., Information Age Publishing, North Carolina, pp. 405–411.

Drton, M. and Perlman, M. (2004), "Model selection for Gaussian concentration graphs," *Biometrika*, 91, 591–602.

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

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

— (2010), "Regularized paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, 33.

Hoeffding, W. (1963), "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Asscociation*, 58, 13–30.

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

Jung, S.-Y., Park, Y., Choi, K.-S., and Kim, Y. (1996), "Markov random field based English part-of-speech tagging system," in *Proceedings of the 16th Conference on Computational Linguistics*, pp. 236–242.

Kolar, M. and Xing, E. (2008), "Improved estimation of high-dimensional Ising models," in *Eprint arXiv:0811.1239*.

Lam, C. and Fan, J. (2009), "Sparsistency and rates of convergence in large covariance matrices estimation," *Annals of Statistics*, 37, 4254–4278.

Li, H. and Gui, J. (2006), "Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks," *Biostatistics*, 7, 302–317.

Li, S. (2001), *Markov Random Field Modeling in Image Analysis*, Springer, New York.

Meinshausen, N. and Buhlmann, P. (2006), "High-dimensional graphs with the lasso," *Annals of Statistics*, 34, 1436–1462.

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009), "Partial correlation estimation by joint sparse regression model," *Journal of the American Statistical Asscociation*, 104, 735–746.

Ravikumar, P., Wainwright, M., and Lafferty, J. (2010), "High-dimensional Ising model selection using ℓ_1-regularized logistic regression," *Annals of Statistics*, To appear.

Ravikumar, P., Wainwright, M., Raskutti, G., and Yu, B. (2008), "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence," Tech. rep., Department of Statistics, University of California, Berkeley.

Rocha, G., Zhao, P., and Yu, B. (2008), "A path following algorithm for Sparse Pseudo-Likelihood Inverse Covariance Estimation (SPLICE)," Tech. rep., Department of Statistics, University of California, Berkeley.

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

Shojaie, A. and Michailidis, G. (2010), "Penalized likelihood methods for estimation of sparse high dimensional directed acyclic graphs," *Biometrika*, to appear.

Wainwright, M. and Jordan, M. (2008), "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, 1, 1–305.

Wang, P., Chao, D., and Hsu, L. (2009), "Learning networks from high dimensional binary data: an application to genomic instability data," *Biometrics*, To appear.

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

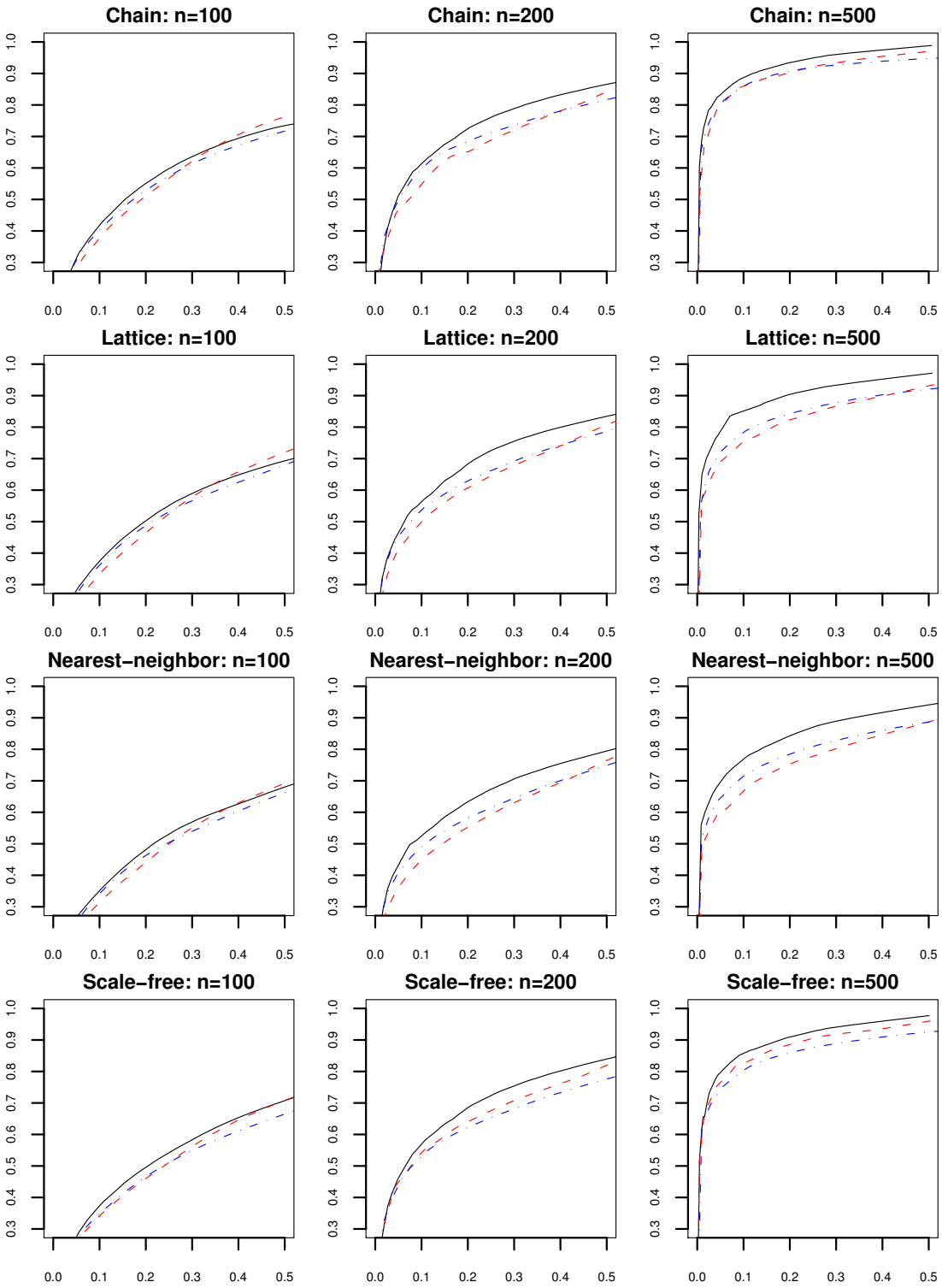

Figure 2: The ROC curves estimated by JOSE, NS-MAX and NS-MAX in simulated examples with sample sizes n=100, 200 and 500. The curves associated with JOSE, NS-MAX and NS-MAX are represented by black solid line, red dashed line and blue dotted-dashed line, respectively. In each panel, the horizontal and the vertical coordinates are 1−specificity and sensitivity, respectively. All results are averages over 50 replications.

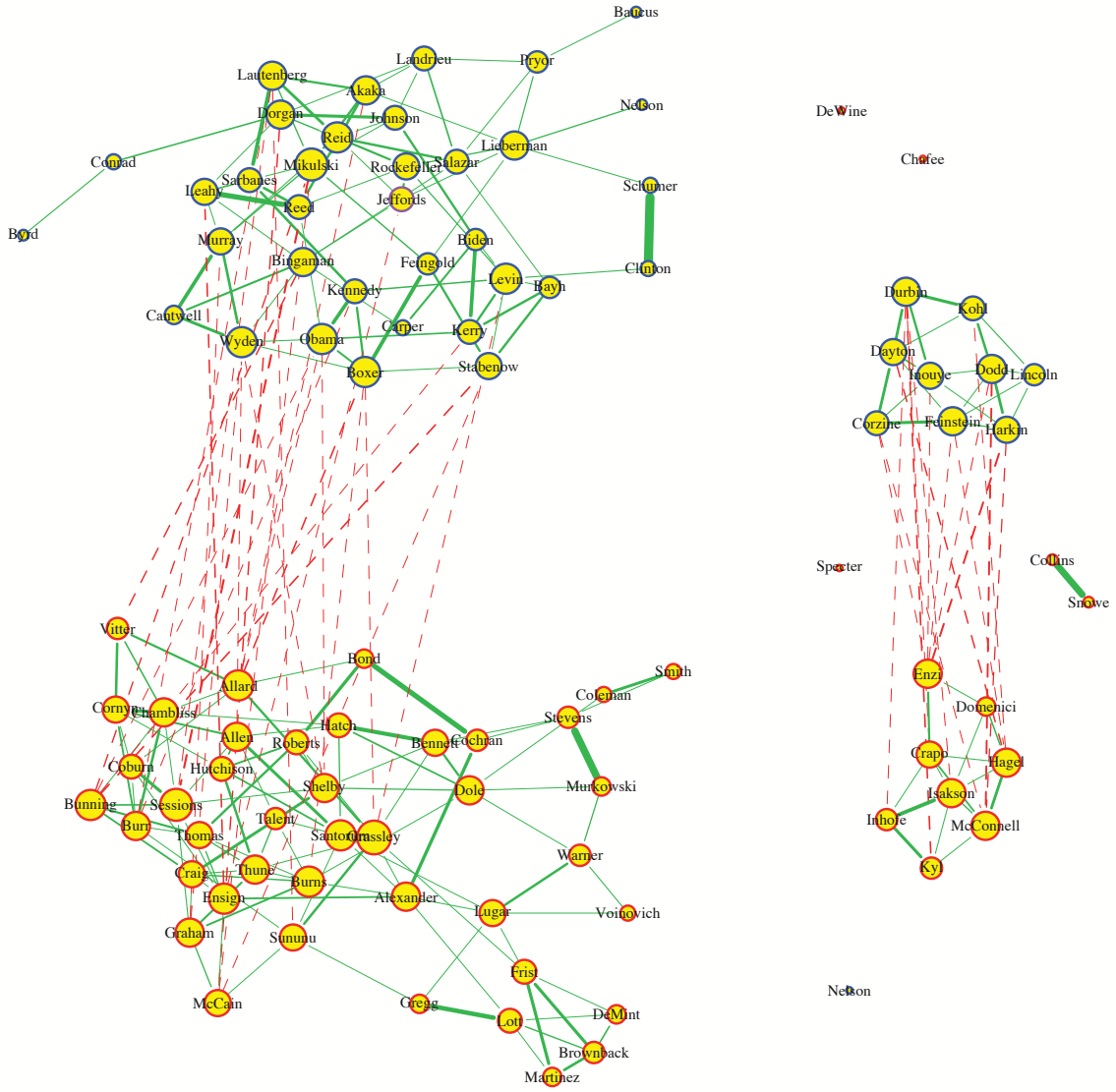

Figure 3: Voting dependencies between senators estimated by the JOSE method. Each red (blue) circle represents a Republican (Democratic) senator, the circle size is proportional to the degree of the node. Senator Jeffords (the purple circle) is an independent senator. A solid green (dashed red) link represents a positive (negative) dependence between two senators. The width of each link is proportional to its associated $|\widehat{\theta}_{j,j'}|$. For clarity, all links with $|\widehat{\theta}_{j,j'}| \leq 0.1$ have the same width.

Appendix I: Main Propositions and Proofs of Theorems

The proof of our main result is divided into many steps; Appendix I presents the main idea of the proof by listing the important propositions and the proofs of Theorems 1 and 2, whereas Appendix II contains additional technical lemmas and proofs of the propositions. The proof bears some similarities to the proof of Ravikumar et al. (2010) for the neighborhood selection method, who in turn adapted the proof from Meinshausen and Buhlmann (2006) to binary data; however, there are also important differences, since all conditions and results are for joint estimation, and many of our bounds need to be more precise than those given by Ravikumar et al. (2010).

The main idea of the proof is as follows. First, we introduce a restricted version of criterion (12), where S is assumed known and all parameters in S^c are set to zero:

$$\widetilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}^{[S]}} \; l(\boldsymbol{\theta}^{[S]}) - \lambda \sum_{(j,j')\in S} |\theta_{j,j'}|. \tag{19}$$

Further, we introduce sample versions of conditions (A) and (B) as follows (see below for detailed definitions of $\boldsymbol{Q}^n$ and $\boldsymbol{U}^n$, the sample analogues of the population quantities $\boldsymbol{Q}^0$ and $\boldsymbol{U}^0$):

(A′) Dependency (sample): There exist positive constants $\tau_{\min}$ and $\tau_{\max}$ such that

$$\Lambda_{\min}(\boldsymbol{Q}^n_{S,S}) \geq \tau_{\min} \quad \text{and} \quad \Lambda_{\max}(\boldsymbol{U}^n_{S,S}) \leq \tau_{\max}. \tag{20}$$

(B′) Incoherence (sample): There exists a constant $\tau \in (0,1)$ such that

$$\|\boldsymbol{Q}^n_{S^c,S}(\boldsymbol{Q}^n_{S,S})^{-1}\|_\infty \leq 1-\tau. \tag{21}$$

The proof consists of the following steps. Proposition 1 and Proposition 2 show that, under sample regularity conditions (A′) and (B′), the conclusions of Theorems 1 and 2 hold for the solution of the restricted problem (19), respectively. Next, Proposition 3 and

Proposition 4 prove that the population regularity conditions (A) and (B) give rise to their sample counterparts (A′) and (B′) with probability tending to 1. Proposition 5 gives the Karush-Kuhn-Tucker (KKT) conditions for the full problem (12), and Proposition 6 shows that, with probability tending to 1, the solution of the restricted problem (19) satisfies the KKT conditions of (12). Thus, the solution of the restricted problem is also the solution of the original problem with probability tending to 1 and both theorems hold.

We start by introducing additional notation. Denote the log-likelihood for the i-th observation by

$$l_i(\boldsymbol{\theta}) = \sum_{j=1}^{p} x_{i,j}\Big(\sum_{k\neq j}\theta_{j,k}x_{i,k}\Big) - \log\Big\{1+\exp\Big(\sum_{k\neq j}\theta_{j,k}x_{i,k}\Big)\Big\}\ , \tag{22}$$

The first derivative of the log-likelihood is $\nabla l_i(\boldsymbol{\theta}) = (\nabla_{1,2}l_i(\boldsymbol{\theta}),\ldots,\nabla_{p-1,p}l_i(\boldsymbol{\theta}))^{\mathsf{T}}$, where

$$\begin{aligned}\nabla_{j,j'}l_i(\boldsymbol{\theta}) &= x_{i,j'}\Big\{x_{i,j} - \frac{\exp(\sum_{k\neq j}\theta_{j,k}x_{i,k})}{1+\exp(\sum_{k\neq j}\theta_{j,k}x_{i,k})}\Big\} \\ &\quad + x_{i,j}\Big\{x_{i,j'} - \frac{\exp(\sum_{k\neq j'}\theta_{j',k}x_{i,k})}{1+\exp(\sum_{k\neq j'}\theta_{j',k}x_{i,k})}\Big\}\ .\end{aligned} \tag{23}$$

The second derivative of $l_i(\boldsymbol{\theta})$ is given by

$$\nabla^2 l_i(\boldsymbol{\theta}) = -\boldsymbol{\mathcal{X}}^{(i)^{\mathsf{T}}}\boldsymbol{\eta}^{(i)}(\boldsymbol{\theta})\boldsymbol{\mathcal{X}}^{(i)}\ , \tag{24}$$

where $\boldsymbol{\eta}^{(i)}(\boldsymbol{\theta}) = \mathrm{diag}(\eta_1^{(i)}(\boldsymbol{\theta}),\ldots,\eta_p^{(i)}(\boldsymbol{\theta}))$ is a $p\times p$ diagonal matrix, and

$$\eta_j^{(i)}(\boldsymbol{\theta}) = \frac{\exp(\sum_{k\neq j}\theta_{j,k}x_{i,k})}{\{1+\exp(\sum_{k\neq j}\theta_{j,k}x_{i,k})\}^2}\ . \tag{25}$$

The first derivative of $\eta_j^{(i)}(\boldsymbol{\theta})$ is given by $\nabla\eta_j^{(i)}(\boldsymbol{\theta}) = \xi_j^{(i)}(\boldsymbol{\theta})(\boldsymbol{\mathcal{X}}^{(i,j)})^{\mathsf{T}}$, where

$$\xi_j^{(i)}(\boldsymbol{\theta}) = \frac{\exp(\sum_{k\neq j}\theta_{j,k}x_{i,k})[1-\exp(\sum_{k\neq j}\theta_{j,k}x_{i,k})]}{[1+\exp(\sum_{k\neq j}\theta_{j,k}x_{i,k})]^3}\ . \tag{26}$$

It is easy to check that $|\nabla_{j,j'}l_i(\boldsymbol{\theta})|\leq 2$, $|\eta_j^{(i)}(\boldsymbol{\theta})|\leq 1$ and $|\xi_j^{(i)}(\boldsymbol{\theta})|\leq 1$. For n observations, the log-likelihood, its first derivative and its second derivative are $l(\boldsymbol{\theta}) = 1/n\sum_{i=1}^n l_i(\boldsymbol{\theta})$, $\nabla l(\boldsymbol{\theta}) = 1/n\sum_{i=1}^n \nabla l_i(\boldsymbol{\theta})$, and $\nabla^2 l(\theta) = 1/n\sum_{i=1}^n \nabla^2 l_i(\theta)$, respectively. Then, the population Fisher

information matrix of (12) at $\boldsymbol{\theta}^0$ can be represented as $\boldsymbol{Q}^0 = \mathrm{E}[\boldsymbol{\mathcal{X}}^{(i)\mathsf{T}}\boldsymbol{\eta}^{(i)}(\boldsymbol{\theta}^0)\boldsymbol{\mathcal{X}}^{(i)}]$, and its sample counterpart $\boldsymbol{Q}^n = -\nabla^2 l(\boldsymbol{\theta}^0) = 1/n\sum_{i=1}^n \boldsymbol{\mathcal{X}}^{(i)\mathsf{T}}\boldsymbol{\eta}^{(i)}(\boldsymbol{\theta}^0)\boldsymbol{\mathcal{X}}^{(i)}$. We also define $\boldsymbol{U}^n = 1/n\sum_{i=1}^n \boldsymbol{\mathcal{X}}^{(i)\mathsf{T}}\boldsymbol{\mathcal{X}}^{(i)}$ as the sample counterpart of $\boldsymbol{U}^0 = \mathrm{E}(\boldsymbol{\mathcal{X}}^{\mathsf{T}}\boldsymbol{\mathcal{X}})$ defined in Section 3. Let $\mathcal{W}$ be any subset of the index set $\{1, 2, \ldots, p(p-1)/2\}$. For any vector $\boldsymbol{\gamma}$, we define $\boldsymbol{\gamma}_{\mathcal{W}}$ as the vector consisting of the elements of $\boldsymbol{\gamma}$ associated with $\mathcal{W}$. Similarly, we define $\boldsymbol{\mathcal{X}}^{(i)}_{\mathcal{W}}$ as the columns of $\boldsymbol{\mathcal{X}}^{(i)}$ associated with $\mathcal{W}$, respectively. Finally, we write $\boldsymbol{\delta} = \boldsymbol{\theta} - \boldsymbol{\theta}^0$, $\widetilde{\boldsymbol{\delta}} = \widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0$ and $\widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0$.

Proposition 1 *Suppose the sample conditions (A′) and (B′) hold. If the tuning parameter* $\lambda = C_\lambda\sqrt{(\log p)/n}$ *for some constant* $C_\lambda > 16(2-\tau)/\tau$ *and* $q\sqrt{(\log p)/n} = o(1)$*, then with probability tending to 1, the optimizer of the restricted criterion* $\widetilde{\boldsymbol{\theta}}$ *satisfies*

$$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\|_2 \leq M\sqrt{\frac{q\log p}{n}} \tag{27}$$

for some constant $M > (2C_\lambda/\tau_{min})\{1 + \tau/(8-4\tau)\}$.

Proposition 2 *Under conditions of Proposition 1, if we further assume* $\theta^0_{min} \geq 2M\sqrt{q(\log p)/n}$*, then with probability tending to 1,* $\widetilde{\theta}_{j,j'} \neq 0$ *for all* $(j,j') \in S$ *and* $\widetilde{\theta}_{j,j'} = 0$ *for all* $(j,j') \in S^c$.

Proposition 3 *(Relationship between sample and population dependency) Suppose the regularity conditions (A) hold, then for any* $\epsilon > 0$,

(i) $\mathrm{P}\{\Lambda_{min}(\boldsymbol{Q}^n_{S,S}) \leq \tau_{min} - \epsilon\} \leq 2\exp\{-(\epsilon^2/2)(n/q^2) + 2\log q\}$;

(ii) $\mathrm{P}\{\Lambda_{\max}(\boldsymbol{U}^n_{S,S}) \geq \tau_{\max} + \epsilon\} \leq 2\exp\{-(\epsilon^2/2)(n/q^2) + 2\log q\}$.

Proposition 4 *(Relationship between sample and population incoherence) Suppose the regularity conditions (A) and (B) hold, then for any* $\epsilon > 0$*, there exists a constant* $C = \min\{\tau^2_{min}\tau^2/288(1-\tau)^2, \tau^2_{min}\tau^2/72, \tau_{min}\tau/48\}$*, such that*

$$\mathrm{P}[\|\boldsymbol{Q}^n_{S^c,S}(\boldsymbol{Q}^n_{S,S})^{-1}\|_\infty \geq 1 - \frac{\tau}{2}] \leq 12\exp\Big(-C\frac{n}{q^3} + 4\log p\Big)\,. \tag{28}$$

Proposition 5 *(KKT conditions) The sufficient and necessary condition for* $\widehat{\boldsymbol{\theta}}$ *to be a solution of problem* (12) *is*

$$\begin{array}{rcll} \nabla_{j,j'} l(\widehat{\boldsymbol{\theta}}) & = & \lambda \mathrm{sgn}(\widehat{\theta}_{j,j'}), & \text{if } \widehat{\theta}_{j,j'} \neq 0; \\ |\nabla_{j,j'} l(\widehat{\boldsymbol{\theta}})| & < & \lambda, & \text{if } \widehat{\theta}_{j,j'} = 0. \end{array} \tag{29}$$

Moreover, this solution is unique due to the strict convexity of problem (12).

Proposition 6 *(The restricted solution satisfies KKT conditions) Under all conditions of Proposition 2, with probability tending to 1, we have,*

(i) $\nabla_{j,j'} l(\widetilde{\boldsymbol{\theta}}) = \lambda \mathrm{sgn}(\widetilde{\theta}_{j,j'})$, *for all* $(j, j') \in S$;

(ii) $|\nabla_{j,j'} l(\widetilde{\boldsymbol{\theta}})| < \lambda$, *for all* $(j, j') \in S^c$.

Proof of Theorem 1. The condition $n > (4/C)q^3 \log(p)$ implies $q\sqrt{(\log p)/n} = o(1)$. In addition, since $n > (4/C)q^3 \log(p)$, we have $-(\epsilon^2/2)(n/q^2) + 2\log q \to -\infty$ and $-Cn/q^3 + 4\log(p)] \to -\infty$. Thus, by Propositions 3 and 4, the sample dependency and incoherence conditions (A′) and (B′) hold with probability 1. Therefore, Proposition 1 holds and, with probability tending to 1, the solution of the restricted problem (19) satisfies parameter estimation consistency.

On the other hand, Proposition 6 shows that, with probability tending to 1, the solution of the restricted problem $\widetilde{\boldsymbol{\theta}}$ satisfies the KKT conditions in Proposition 5. Since the criterion (12) is strictly convex, we conclude $\widetilde{\boldsymbol{\theta}}$ is the unique solution of (12), i.e., $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}$. This proves Theorem 1.

□

Proof of Theorem 2 is analogous to Proof of Theorem 1 and is omitted.

Appendix II: Proofs of Propositions

This appendix contains several additional technical lemmas and proofs of Propositions 1-6.

Lemma 1 *[Bound on $\nabla l(\boldsymbol{\theta}^0)$] With probability tending to 1, $\|\nabla l(\boldsymbol{\theta}^0)\|_\infty \leq C_\nabla \sqrt{(\log p)/n}$ for some constant $C_\nabla > 4$.*

Proof of Lemma 1: Note that $E[\nabla l_i(\boldsymbol{\theta}^0)] = \mathbf{0}$, $1 \leq i \leq n$ and $|\nabla_{j,j'} l_i(\theta_0)| \leq 2$, $1 \leq i \leq n, 1 \leq j < j' \leq p$. By applying the Azuma-Hoeffding inequality (Hoeffding, 1963), we get

$$P[|\nabla_{j,j'} l(\boldsymbol{\theta}^0)| \geq t] \leq 2\exp(-nt^2/8). \tag{30}$$

Letting $t = C_\nabla \sqrt{(\log p)/n}$ for some constant $C_\nabla > 0$, we obtain

$$P\Big[|\nabla_{j,j'} l(\boldsymbol{\theta}^0)| \geq C_\nabla \sqrt{\frac{\log p}{n}}\Big] \leq 2\exp(-C_\nabla^2 \log p/8)\ . \tag{31}$$

Then, by the union-sum inequality we have

$$P[\|\nabla l(\boldsymbol{\theta}^0)\|_\infty \geq C_\nabla \sqrt{\frac{\log p}{n}}] \leq 2\exp(-C_\nabla^2 \log p\,/8 + 2\log p). \tag{32}$$

Setting $C_\nabla > 4$ establishes the lemma. □

Lemma 2 *[Bound on $-\boldsymbol{\delta}_S{}^\mathsf{T}[\nabla^2 l(\boldsymbol{\theta}^0 + \alpha\boldsymbol{\delta}^{[S]})]_{S,S}\boldsymbol{\delta}_S/m$ If the sample dependency condition (A') holds and $q\sqrt{(\log p)/n} = o(1)$, then for any $\alpha \in [0,1]$, with probability tending to 1,*

$$-\boldsymbol{\delta}_S{}^\mathsf{T}[\nabla^2 l(\boldsymbol{\theta}^0 + \alpha\boldsymbol{\delta}^{[S]})]_{S,S}\boldsymbol{\delta}_S \geq \frac{1}{2}\tau_{\min}\|\boldsymbol{\delta}_S\|_2^2\ . \tag{33}$$

Proof of Lemma 2: Applying the mean value theorem, we have $\eta_j(\boldsymbol{\theta}^0 + \alpha\boldsymbol{\delta}^{[S]}) = \eta_j(\boldsymbol{\theta}^0) + \alpha\nabla\eta_j(\boldsymbol{\theta}^0 + \alpha^*\boldsymbol{\delta}^{[S]})^\intercal\boldsymbol{\delta}^{[S]}$, for some constant $\alpha^* \in (0, \alpha)$. Then, we have

$$\begin{aligned}
& -\boldsymbol{\delta}_S{}^\intercal[\nabla^2 l(\boldsymbol{\theta}^0 + \alpha\boldsymbol{\delta}^{[S]})]_{S,S}\boldsymbol{\delta}_S \\
= & \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\mathcal{X}}_S^{(i)}\boldsymbol{\delta}_S)^\intercal\boldsymbol{\eta}(\boldsymbol{\theta}^0 + \alpha\boldsymbol{\delta}^{[S]})(\boldsymbol{\mathcal{X}}_S^{(i)}\boldsymbol{\delta}_S) \\
= & \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}\eta_j(\boldsymbol{\theta}^0)(\boldsymbol{\mathcal{X}}_S^{(i,j)}\boldsymbol{\delta}_S)^2 \\
& +\frac{\alpha}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}\nabla\eta_j(\boldsymbol{\theta}^0 + \alpha^*\boldsymbol{\delta}^{[S]})^\intercal\boldsymbol{\delta}^{[S]}(\boldsymbol{\mathcal{X}}_S^{(i,j)}\boldsymbol{\delta}_S)^2 \\
\geq & -\boldsymbol{\delta}_S{}^\intercal[\nabla^2 l(\boldsymbol{\theta}^0)]_{S,S}\boldsymbol{\delta}_S \\
& -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}|\xi_j^{(i)}(\boldsymbol{\theta}^0 + \alpha^*\boldsymbol{\delta}^{[S]})||\boldsymbol{\mathcal{X}}_S^{(i,j)}\boldsymbol{\delta}_S|(\boldsymbol{\mathcal{X}}_S^{(i,j)}\boldsymbol{\delta}_S)^2 \ . \qquad (34)
\end{aligned}$$

The first term is bounded from below by

$$-\boldsymbol{\delta}_S{}^\intercal[\nabla^2 l(\boldsymbol{\theta}^0)]_{S,S}\boldsymbol{\delta}_S \geq \Lambda_{min}(\boldsymbol{Q}_{S,S}^n)\|\boldsymbol{\delta}_S\|_2^2 \geq \tau_{min}\|\boldsymbol{\delta}_S\|_2^2 \ . \qquad (35)$$

To bound the second term, notice that $|\boldsymbol{\mathcal{X}}_S^{(i,j)}\boldsymbol{\delta}_S| \leq \|\boldsymbol{\mathcal{X}}_S^{(i,j)}\|_\infty\|\boldsymbol{\delta}_S\|_1 \leq \|\boldsymbol{\delta}_S\|_1$ and recall that $|\xi_j^{(i)}| \leq 1$. Then the second term is bounded from above by

$$\|\boldsymbol{\delta}_S\|_1\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}(\boldsymbol{\mathcal{X}}_S^{(i,j)}\boldsymbol{\delta}_S)^2 \leq \tau_{\max}\|\boldsymbol{\delta}_S\|_1\|\boldsymbol{\delta}_S\|_2^2 \leq (\tau_{min}/2)\|\boldsymbol{\delta}_S\|_2^2 \ , \qquad (36)$$

since $\|\boldsymbol{\delta}_S\|_1 \leq \sqrt{q}\|\boldsymbol{\delta}_S\|_2 = Mq\sqrt{(\log p)/n} = o(1)$ and thus when n is large enough, $\|\boldsymbol{\delta}_S\|_1 \leq \tau_{min}/(2\tau_{\max})$. Putting (35) and (36) together establishes the lemma. □

Proof of Proposition 1: The proof relies on the convex function proof method from Rothman et al. (2008). Define

$$\mathrm{G}(\boldsymbol{\delta}_S) = -[l(\boldsymbol{\theta}^0 + \boldsymbol{\delta}^{[S]}) - l(\boldsymbol{\theta}^0)] + \lambda(\|\boldsymbol{\theta}^0 + \boldsymbol{\delta}^{[S]}\|_1 - \|\boldsymbol{\theta}^0\|_1). \qquad (37)$$

It can be seen from (19) that $\widetilde{\boldsymbol{\delta}}_S = \widetilde{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_S^0$ minimizes $\mathrm{G}(\boldsymbol{\delta}_S)$. Moreover, $\mathrm{G}(\mathbf{0}_S) = 0$, thus we must have $\mathrm{G}(\widetilde{\boldsymbol{\delta}}_S) \leq 0$. If we take a ball $\mathcal{A}$ which contains $\mathbf{0}_S$, and show that G is strictly

positive everywhere on the boundary $\partial\mathcal{A}$, then it implies that G has a local minimum inside $\mathcal{A}$, since G is continuous and $\mathrm{G}(\mathbf{0}_S) = 0$. Specifically, we define $\mathcal{A} = \{\boldsymbol{\delta}_S : \|\boldsymbol{\delta}_S\|_2 \leq Ma_n\}$, with boundary $\partial\mathcal{A} = \{\boldsymbol{\delta}_S : \|\boldsymbol{\delta}_S\|_2 = Ma_n\}$, for some constant $M > (2/\tau_{\min})[1+\tau/(8-4\tau)]C_\lambda$ and $a_n = \sqrt{q(\log p)/n}$. For any $\boldsymbol{\delta}_S \in \partial\mathcal{A}$, the Taylor series expansion gives $\mathrm{G}(\boldsymbol{\delta}_S) = I_1 + I_2 + I_3$, where

$$\begin{aligned} I_1 &= -[\nabla l(\boldsymbol{\theta}^0)]_S{}^{\mathsf{T}}\boldsymbol{\delta}_S \ , \\ I_2 &= -\boldsymbol{\delta}_S{}^{\mathsf{T}}[\nabla^2 l(\boldsymbol{\theta}^0 + \alpha\boldsymbol{\delta}^{[S]})]_{S,S}\boldsymbol{\delta}_S, \text{ for some } \alpha \in [0,1] \ , \\ I_3 &= \lambda(\|\boldsymbol{\theta}^0 + \boldsymbol{\delta}^{[S]}\|_1 - \|\boldsymbol{\theta}^0\|_1) = \lambda(\|\boldsymbol{\theta}^0_S + \boldsymbol{\delta}_S\|_1 - \|\boldsymbol{\theta}^0_S\|_1) \ . \end{aligned} \tag{38}$$

Since $C_\lambda > 16(2-\tau)/\tau$, we have $[\tau/(8-4\tau)]C_\lambda > 4$. By Lemma 1,

$$|I_1| \leq \|[\nabla l(\boldsymbol{\theta}^0)]_S\|_\infty \|\boldsymbol{\delta}_S\|_1 \leq \|[\nabla l(\boldsymbol{\theta}^0)]_S\|_\infty \sqrt{q}\|\boldsymbol{\delta}_S\|_2 \leq \frac{\tau}{8-4\tau} C_\lambda M q \frac{\log p}{n} \ .$$

By Lemma 2, $I_2 \geq (\tau_{\min}/2)\|\boldsymbol{\delta}_S\|_2^2 = (\tau_{\min}/2)M^2 q(\log p)/n$. Finally, by the triangular inequality $|I_3| \leq \lambda\|\boldsymbol{\delta}_S\|_1 \leq \lambda\sqrt{q}\|\boldsymbol{\delta}_S\|_2 = C_\lambda M q(\log p)/n$. Then we have

$$\mathrm{G}(\boldsymbol{\delta}_S) \geq M^2 \frac{q\log p}{n}\Big(\frac{\tau_{\min}}{2} - \frac{\tau C_\lambda}{4(2-\tau)M} - \frac{C_\lambda}{M}\Big) > 0. \tag{39}$$

The last inequality uses the condition $M > 2C_\lambda[1 + \tau/(8-4\tau)]/\tau_{\min}$. Therefore, with probability tending to 1, we have $\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\|_F = \|\widetilde{\boldsymbol{\theta}}_S - \boldsymbol{\theta}^0_S\|_F \leq M\sqrt{(q\log p)/n}$.

□

Proof of Proposition 2: Since $\widetilde{\boldsymbol{\theta}}$ is the solution of the restricted problem (19), we have $\widetilde{\boldsymbol{\theta}}_{j,j'} = 0$ for all $(j,j') \in S^c$. To show $\widetilde{\boldsymbol{\theta}}_{j,j'} \neq 0$ for all $(j,j') \in S$, it is sufficient to show

$$\|\widetilde{\boldsymbol{\theta}}_S - \boldsymbol{\theta}^0_S\|_\infty \leq \frac{\theta^0_{\min}}{2} \ , \tag{40}$$

because then $|\widetilde{\theta}_{j,j'}| \geq |\widetilde{\theta}^0_{j,j'}| - |\widetilde{\theta}_{j,j'} - \widetilde{\theta}^0_{j,j'}| \geq \theta^0_{\min}/2$ for all $(j,j') \in S$. With probability tending to 1, by Proposition 1 we have

$$\|\widetilde{\boldsymbol{\theta}}_S - \boldsymbol{\theta}^0_S\|_\infty \leq \|\widetilde{\boldsymbol{\theta}}_S - \boldsymbol{\theta}^0_S\|_2 \leq M\sqrt{\frac{q(\log p)}{n}} \ .$$

The additional condition $\theta^0_{\min} \geq 2M\sqrt{q(\log p)/n}$ implies (40). □

Lemma 3 *For any* $\epsilon > 0$,

(i) $P[\|\boldsymbol{Q}^n_{S^c,S} - \boldsymbol{Q}^0_{S^c,S}\|_\infty \geq \epsilon] \leq 2\exp\{-(\epsilon^2/2)(n/q^2) + \log(q) + \log[p(p-1)/2 - q]\}$,

(ii) $P[\|\boldsymbol{Q}^n_{S,S} - \boldsymbol{Q}^0_{S,S}\|_\infty \geq \epsilon] \leq 2\exp\{-(\epsilon^2/2)(n/q^2) + 2\log(q)\}$.

Proof of Lemma 3: We first prove claim (i). Let $v^{(i)}_{(j,j'),(h,h')}$ be the $[(j,j'),(h,h')]$-th element of matrix $\boldsymbol{\mathcal{X}}^{(i)\mathsf{T}}\boldsymbol{\eta}\boldsymbol{\mathcal{X}}^{(i)} - \boldsymbol{Q}^0$. Note $\mathrm{E}(v^{(i)}_{(j,j'),(h,h')}) = 0$ and $|v^{(i)}_{(j,j'),(h,h')}| \leq 1$, and let $v_{(j,j'),(h,h')} = 1/n\sum_{i=1}^n v^{(i)}_{(j,j'),(h,h')}$. Then

$$\begin{aligned} P[\sum_{(h,h')\in S} |v_{(j,j'),(h,h')}| \geq \epsilon] &\leq \sum_{(h,h')\in S} P[|v_{(j,j'),(h,h')}| \geq \epsilon/q] \\ &\leq q \max_{(h,h')\in S} P[|v_{(j,j'),(h,h')}| \geq \epsilon/q]. \end{aligned} \tag{41}$$

Combining the union-sum inequality with (41), we have

$$P[\|\boldsymbol{Q}^n_{S^c,S} - \boldsymbol{Q}^0_{S^c,S}\|_\infty \geq \epsilon] \leq q\Big(\frac{p(p-1)}{2} - q\Big) \max_{(h,h')\in S} P[|v_{(j,j'),(h,h')}| \geq \epsilon/q]. \tag{42}$$

Then, by the Azuma-Hoeffding inequality (Hoeffding, 1963), we have $P[|v_{(j,j'),(h,h')}| \geq \epsilon/q] \leq 2\exp\{-(\epsilon^2/2)(n/q^2)\}$, and (i) follows. The proof of (ii) is similar. □

Proof of Proposition 3: Note that

$$\begin{aligned} \Lambda_{\min}(\boldsymbol{Q}^n_{S,S}) &= \min_{\|y\|_2=1} [y^\mathsf{T}\boldsymbol{Q}^0_{S,S}y + y^\mathsf{T}(\boldsymbol{Q}^n_{S,S} - \boldsymbol{Q}^0_{S,S})y] \\ &\geq \Lambda_{\min}(\boldsymbol{Q}^0_{S,S}) - \|\boldsymbol{Q}^n_{S,S} - \boldsymbol{Q}^0_{S,S}\|_2 \geq \tau_{\min} - \|\boldsymbol{Q}^n_{S,S} - \boldsymbol{Q}^0_{S,S}\|_\infty \ . \end{aligned}$$

Now claim (i) follows from Lemma 3 (ii). The proof of claim (ii) is similar. □

Lemma 4 *Suppose conditions (A) and (B) hold. Then for any $\epsilon > 0$,*

$$P[\|(\boldsymbol{Q}^n_{S,S})^{-1} - (\boldsymbol{Q}^0_{S,S})^{-1}\|_\infty \geq \epsilon] \leq 4\exp\{-(\tau_{\min}\epsilon^2/8)(n/q^3) + 2\log(q)\}. \tag{43}$$

Proof of Lemma 4: Writing $(\boldsymbol{Q}^n_{S,S})^{-1} - (\boldsymbol{Q}^0_{S,S})^{-1} = (\boldsymbol{Q}^0_{S,S})^{-1}(\boldsymbol{Q}^0_{S,S} - \boldsymbol{Q}^n_{S,S})(\boldsymbol{Q}^n_{S,S})^{-1}$ and applying norm inequalities, we have

$$\begin{aligned}
\|(\boldsymbol{Q}^n_{S,S})^{-1} - (\boldsymbol{Q}^0_{S,S})^{-1}\|_\infty &\leq \sqrt{q}\|(\boldsymbol{Q}^0_{S,S})^{-1}(\boldsymbol{Q}^0_{S,S} - \boldsymbol{Q}^n_{S,S})(\boldsymbol{Q}^n_{S,S})^{-1}\|_2 \\
&\leq \sqrt{q}\|(\boldsymbol{Q}^0_{S,S})^{-1}\|_2\|\boldsymbol{Q}^0_{S,S} - \boldsymbol{Q}^n_{S,S}\|_\infty\|(\boldsymbol{Q}^n_{S,S})^{-1}\|_2 \\
&\leq \frac{\sqrt{q}}{\tau_{\min}}\|\boldsymbol{Q}^0_{S,S} - \boldsymbol{Q}^n_{S,S}\|_\infty\|(\boldsymbol{Q}^n_{S,S})^{-1}\|_2 \ . \qquad (44)
\end{aligned}$$

The last inequality holds because $\|(\boldsymbol{Q}^0_{S,S})^{-1}\|_2 = \{\Lambda_{\min}(\boldsymbol{Q}^0_{S,S})\}^{-1}$. In addition, we have $\|(\boldsymbol{Q}^n_{S,S})^{-1}\|_2 = \{\Lambda_{\min}(\boldsymbol{Q}^n_{S,S})\}^{-1}$. Then by setting $\epsilon = \tau_{\min}/2$ in Proposition 3 (i), we have

$$\begin{aligned}
\mathrm{P}\left[\frac{\|(\boldsymbol{Q}^n_{S,S})^{-1}\|_2}{\tau_{\min}} \geq \frac{2}{\tau^2_{\min}}\right] &= \mathrm{P}[\Lambda_{\min}(\boldsymbol{Q}^n_{S,S}) \\
&\leq \frac{\tau_{\min}}{2}] \leq 2\exp(-\frac{\tau^2_{\min}}{8}\frac{n}{q^2} + 2\log q). \qquad (45)
\end{aligned}$$

By replacing ϵ in Lemma 3 (ii) with $\tau^2_{\min}\epsilon/(2\sqrt{q})$, we have

$$\mathrm{P}[\|\boldsymbol{Q}^0_{S,S} - \boldsymbol{Q}^n_{S,S}\|_\infty \geq \frac{\tau^2_{\min}\epsilon}{2\sqrt{q}}] \leq 2\exp(-\frac{\tau^4_{\min}\epsilon^2}{8}\frac{n}{q^3} + 2\log q) \ . \tag{46}$$

Finally,

$$\mathrm{P}[\|(\boldsymbol{Q}^n_{S,S})^{-1} - (\boldsymbol{Q}^0_{S,S})^{-1}\|_\infty \geq \epsilon] \leq \mathrm{P}[\frac{\|\boldsymbol{Q}^n_{S,S}\|_2}{\tau_{\min}} \geq \frac{2}{\tau^2_{\min}}] + \mathrm{P}[\sqrt{q}\|\boldsymbol{Q}^0_{S,S} - \boldsymbol{Q}^n_{S,S}\|_\infty \geq \frac{\tau^2_{\min}\epsilon}{2}] \ ,$$

and the lemma follows. □

Proof of Proposition 4: we write $\boldsymbol{Q}^n_{S^c,S}(\boldsymbol{Q}^n_{S,S})^{-1} = \boldsymbol{T}_1 + \boldsymbol{T}_2 + \boldsymbol{T}_3 + \boldsymbol{T}_4$, where

$$\begin{aligned}
\boldsymbol{T}_1 &= \boldsymbol{Q}^0_{S^c,S}[(\boldsymbol{Q}^n_{S,S})^{-1} - (\boldsymbol{Q}^0_{S,S})^{-1}] \ , \\
\boldsymbol{T}_2 &= (\boldsymbol{Q}^n_{S^c,S} - \boldsymbol{Q}^0_{S^c,S})(\boldsymbol{Q}^0_{S,S})^{-1} \ , \\
\boldsymbol{T}_3 &= (\boldsymbol{Q}^n_{S^c,S} - \boldsymbol{Q}^0_{S^c,S})[(\boldsymbol{Q}^n_{S,S})^{-1} - (\boldsymbol{Q}^0_{S,S})^{-1}] \ , \\
\boldsymbol{T}_4 &= \boldsymbol{Q}^0_{S^c,S}(\boldsymbol{Q}^0_{S,S})^{-1} \ .
\end{aligned}$$

To bound $\boldsymbol{T}_1$, we write $\boldsymbol{T}_1 = \boldsymbol{Q}^0_{S^c,S}(\boldsymbol{Q}^0_{S,S})^{-1}(\boldsymbol{Q}^0_{S,S} - \boldsymbol{Q}^n_{S,S})(\boldsymbol{Q}^n_{S,S})^{-1}$. Thus,

$$\|\boldsymbol{T}_1\|_\infty \le \|\boldsymbol{Q}^0_{S^c,S}(\boldsymbol{Q}^0_{S,S})^{-1}\|_\infty \|\boldsymbol{Q}^n_{S,S} - \boldsymbol{Q}^0_{S,S}\|_\infty (\sqrt{q}\|(\boldsymbol{Q}^n_{S,S})^{-1}\|_2) \ .$$

By condition (B), we have $\|\boldsymbol{Q}^0_{S^c,S}(\boldsymbol{Q}^0_{S,S})^{-1}\|_\infty \le 1-\tau$. By setting $\epsilon = \tau_{\min}/2$ in Proposition 3(i), and $\epsilon = \tau_{\min}\tau/(12(1-\tau)\sqrt{q})$ in Lemma 3(ii), we have

$$\begin{aligned}
& \mathrm{P}[\|\boldsymbol{T}_1\|_\infty \ge \frac{\tau}{6}] \\
\le \ & \mathrm{P}\left[\|\boldsymbol{Q}^n_{S,S} - \boldsymbol{Q}_{S,S}\|_\infty \ge \frac{\tau_{\min}\tau}{12(1-\tau)\sqrt{q}}\right] + \mathrm{P}\left[\|(\boldsymbol{Q}^n_{S,S})^{-1}\|_2 \ge \frac{2}{\tau_{\min}}\right] \\
\le \ & 2\exp\left(-\frac{\tau^2_{\min}\tau^2}{288(1-\tau)^2}\frac{n}{q^3} + 2\log q\right) + 2\exp\left(-\frac{\tau^2_{\min}}{8}\frac{n}{q^2} + 2\log q\right). \qquad (47)
\end{aligned}$$

To bound $\boldsymbol{T}_2$, we write

$$\|\boldsymbol{T}_2\|_\infty \le \|\boldsymbol{Q}^n_{S^c,S} - \boldsymbol{Q}^0_{S^c,S}\|_\infty \sqrt{q}\|(\boldsymbol{Q}^0_{S,S})^{-1}\|_2 \le \frac{\sqrt{q}}{\tau_{\min}}\|\boldsymbol{Q}^n_{S^c,S} - \boldsymbol{Q}^0_{S^c,S}\|_\infty \ .$$

By setting $\epsilon = \tau_{\min}\tau/(6\sqrt{q})$ in Lemma 3 (i), we have

$$\begin{aligned}
\mathrm{P}[\|\boldsymbol{T}_2\|_\infty \ge \frac{\tau}{6}] \ &\le \ \mathrm{P}(\|\boldsymbol{Q}^n_{S^c,S} - \boldsymbol{Q}^0_{S^c,S}\|_\infty \ge \frac{\tau_{\min}\tau}{6\sqrt{q}}) \\
&\le \ 2\exp\{-\frac{\tau^2_{\min}\tau^2}{72}\frac{n}{q^3} + \log q + \log[p(p-1)/2 - q]\}. \qquad (48)
\end{aligned}$$

To bound $\boldsymbol{T}_3$, we set $\epsilon = \sqrt{\tau/6}$ in both Lemma 3 (i) and Lemma 4, so that

$$\begin{aligned}
\mathrm{P}[\|\boldsymbol{T}_3\|_\infty \ge \frac{\tau}{6}] \ &\le \ \mathrm{P}[\|\boldsymbol{Q}^n_{S^c,S} - \boldsymbol{Q}^0_{S^c,S}\|_\infty \ge \sqrt{\frac{\tau}{6}}] \\
&\quad + \mathrm{P}[\|(\boldsymbol{Q}^n_{S,S})^{-1} - (\boldsymbol{Q}^0_{S,S})^{-1}\|_\infty \ge \sqrt{\frac{\tau}{6}}] \\
&\le \ 2\exp\{-\frac{\tau}{12}\frac{n}{q^2} + \log q + \log[p(p-1)/2 - q]\} \\
&\quad + 4\exp\{-\frac{\tau_{\min}\tau}{48}\frac{n}{q^3} + 2\log q\}. \qquad (49)
\end{aligned}$$

Finally, $\|\boldsymbol{T}_4\|_\infty \le 1-\tau$ by condition (B). Since $\log q \le 2\log p$ and $\log[p(p-1)/2-q] \le 2\log p$, we have

$$\begin{aligned}
\mathrm{P}[\|\boldsymbol{Q}^n_{S^c,S}(\boldsymbol{Q}^n_{S,S})^{-1}\|_\infty \ge 1 - \frac{\tau}{2}] \ &\le \ \mathrm{P}[\|\boldsymbol{T}_1\|_\infty \ge \frac{\tau}{6}] + \mathrm{P}[\|\boldsymbol{T}_2\|_\infty \ge \frac{\tau}{6}] + \mathrm{P}[\|\boldsymbol{T}_3\|_\infty \ge \frac{\tau}{6}] \\
&\le \ 12\exp\Big(-C\frac{n}{q^3} + 4\log p\Big) \ , \qquad (50)
\end{aligned}$$

where $C = \min\{\tau_{\min}^2\tau^2/288(1-\tau)^2, \tau_{\min}^2\tau^2/72, \tau_{\min}\tau/48\}$. □

Lemma 5 *[Bound on $[\nabla^2 l(\boldsymbol{\theta}^0 + \alpha\boldsymbol{\delta}) - \nabla^2 l(\boldsymbol{\theta}^0)]\boldsymbol{\delta}$] Suppose (A) holds. For any $\alpha \in [0,1]$,*

$$\|[\nabla^2 l(\boldsymbol{\theta}^0 + \alpha\boldsymbol{\delta}) - \nabla^2 l(\boldsymbol{\theta}^0)]\boldsymbol{\delta}\|_\infty \leq \tau_{\max}\|\boldsymbol{\delta}_S\|_2^2 \ . \tag{51}$$

Proof of Lemma 5: We have

$$\begin{aligned}
& |\{[\nabla^2 l(\boldsymbol{\theta}^0 + \alpha\boldsymbol{\delta})]_{(j,j'),S} - [\nabla^2 l(\boldsymbol{\theta}^0)]_{(j,j'),S}\}\boldsymbol{\delta}_S| \\
\leq & \ \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^p |\boldsymbol{\mathcal{X}}_{j,j'}^{(i,j)^\intercal}||[\eta_j(\boldsymbol{\theta}^0 + \alpha\boldsymbol{\delta}_S) - \eta_j(\boldsymbol{\theta}^0)](\boldsymbol{\mathcal{X}}_S^{(i,j)}\boldsymbol{\delta}_S)| \\
\leq & \ \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^p |\xi_j^{(i)}(\boldsymbol{\theta}^0 + \alpha^*\boldsymbol{\delta}_S)|(\boldsymbol{\mathcal{X}}_S^{(i,j)}\boldsymbol{\delta}_S)^2 \leq \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^p(\boldsymbol{\mathcal{X}}_S^{(i,j)}\boldsymbol{\delta}_S)^2 \\
\leq & \ \Lambda_{\max}(\boldsymbol{U}^n)\|\boldsymbol{\delta}_S\|_2^2 \leq \tau_{\max}\|\boldsymbol{\delta}_S\|_2^2.
\end{aligned} \tag{52}$$

Since $\|[\nabla^2 l(\boldsymbol{\theta}^0 + \alpha\boldsymbol{\delta}) - \nabla^2 l(\boldsymbol{\theta}^0)]\boldsymbol{\delta}\|_\infty = \max_{j<j'}|\{[\nabla^2 l(\boldsymbol{\theta}^0 + \alpha\boldsymbol{\delta})]_{(j,j'),S} - [\nabla l(\boldsymbol{\theta}^0)]_{(j,j'),S}\}\boldsymbol{\delta}_S|$, the lemma follows. □

Proof of Proposition 6: By Proposition 2, with probability tending to 1 $\widetilde{\theta}_{j,j'} \neq 0$ for all $(j,j') \in S$. Since $\widetilde{\boldsymbol{\theta}}$ is the maximizer of the restricted problem (19), with probability tending to 1, $\nabla_{j,j'} l(\widetilde{\boldsymbol{\theta}}) = \lambda \mathrm{sgn}(\widetilde{\theta}_{j,j'})$ for all $(j,j') \in S$, and claim (i) follows.

To show (ii), let $\boldsymbol{u} = \nabla l(\widetilde{\boldsymbol{\theta}})/\lambda$. By (i), $\|\boldsymbol{u}_S\|_\infty = 1$. In addition, by the mean value theorem we have

$$\lambda\boldsymbol{u} - \nabla l(\boldsymbol{\theta}^0) = \nabla^2 l(\boldsymbol{\theta}^0)\widetilde{\boldsymbol{\delta}} = -\boldsymbol{Q}^n\widetilde{\boldsymbol{\delta}} + \boldsymbol{r}^n \ , \tag{53}$$

where $\alpha \in (0,1)$ and $\boldsymbol{r}^n = [\nabla^2 l(\boldsymbol{\theta}^0 + \alpha\widetilde{\boldsymbol{\delta}}) - \nabla^2 l(\boldsymbol{\theta}^0)]\widetilde{\boldsymbol{\delta}}$. Decomposing $\boldsymbol{Q}^n$ and using $\widetilde{\boldsymbol{\delta}}_{S^c} = \boldsymbol{0}$, we have

$$\boldsymbol{Q}_{S,S}^n\widetilde{\boldsymbol{\delta}}_S = -\lambda\boldsymbol{u}_S + [\nabla l(\boldsymbol{\theta}^0)]_S + \boldsymbol{r}_S^n; \tag{54}$$

$$\boldsymbol{Q}_{S^c,S}^n\widetilde{\boldsymbol{\delta}}_S = -\lambda\boldsymbol{u}_{S^c} + [\nabla l(\boldsymbol{\theta}^0)]_{S^c} + \boldsymbol{r}_{S^c}^n. \tag{55}$$

The sample dependency condition implies $\boldsymbol{Q}^n_{S,S}$ is invertible. Thus we can plug (54) into (55) to obtain

$$\boldsymbol{Q}^n_{S^c,S}(\boldsymbol{Q}^n_{S,S})^{-1}(-\lambda \boldsymbol{u}_S + [\nabla l(\boldsymbol{\theta}^0)]_S + \boldsymbol{r}^n_S) = -\lambda \boldsymbol{u}_{S^c} + [\nabla l(\boldsymbol{\theta}^0)]_{S^c} + \boldsymbol{r}^n_{S^c} \ . \tag{56}$$

Extracting $\boldsymbol{u}_{S^c}$, we have

$$\begin{aligned}
\|\boldsymbol{u}_{S^c}\|_\infty &\leq \frac{\|[\nabla l(\boldsymbol{\theta}^0)]_{S^c}\|_\infty}{\lambda} + \frac{\|\boldsymbol{r}^n_{S^c}\|_\infty}{\lambda} \\
&\quad + \|\boldsymbol{Q}^n_{S^c,S}(\boldsymbol{Q}^n_{S,S})^{-1}\|_\infty \left(\|\boldsymbol{u}_S\|_\infty + \frac{\|[\nabla l(\boldsymbol{\theta}^0)]_S\|_\infty}{\lambda} + \frac{\|\boldsymbol{r}^n_S\|_\infty}{\lambda} \right) \\
&\leq \frac{\|\nabla l(\boldsymbol{\theta}^0)\|_\infty}{\lambda} + \frac{\|\boldsymbol{r}^n\|_\infty}{\lambda} \\
&\quad + \|\boldsymbol{Q}^n_{S^c,S}(\boldsymbol{Q}^n_{S,S})^{-1}\|_\infty (\|\boldsymbol{u}\|_\infty + \frac{\|\nabla l(\boldsymbol{\theta}^0)\|_\infty}{\lambda} + \frac{\|\boldsymbol{r}^n\|_\infty}{\lambda}) \\
&\leq 1 - \tau + (2-\tau)(\frac{\|\nabla l(\boldsymbol{\theta}^0)\|_\infty}{\lambda} + \frac{\|\boldsymbol{r}^n\|_\infty}{\lambda}).
\end{aligned} \tag{57}$$

By setting $C_\nabla = \tau(8-4\tau)C_\lambda$ in Lemma 1, $\|\nabla l(\boldsymbol{\theta}^0)\|_\infty/\lambda \leq \tau/(8-4\tau)$. By Lemma 5, we have $\|\boldsymbol{r}^n\|_\infty/\lambda \leq \tau_{\max}\|\widetilde{\delta}_S\|_2^2/\lambda \leq (\tau_{\max}M^2/C_\lambda)q\sqrt{\log p/n} \leq \tau/(8-4\tau)$, where the last inequality holds by the condition $q\sqrt{(\log p)/n} = o(1)$ when n is sufficiently large. Thus

$$\|\boldsymbol{u}_{S^c}\|_\infty \leq 1 - \frac{\tau}{2} < 1 \ , \tag{58}$$

and we have $\|[\nabla l(\widetilde{\boldsymbol{\theta}})]_{S^c}\|_\infty = \lambda\|\boldsymbol{u}_{S^c}\|_\infty < \lambda$. □