

# Sparse Ising Models with Covariates

Jie Cheng<sup>1</sup>, Elizaveta Levina<sup>1</sup>, Pei Wang<sup>2</sup> and Ji Zhu<sup>1</sup>

<sup>1</sup>Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.

<sup>2</sup>Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, U.S.A.

September 28, 2012

## Abstract

There has been a lot of work fitting Ising models to multivariate binary data in order to understand the conditional dependency relationships between the variables. However, additional covariates are frequently recorded together with the binary data, and may influence the dependence relationships. Motivated by such a dataset on genomic instability collected from tumor samples of several types, we propose a sparse covariate dependent Ising model to study both the conditional dependency within the binary data and its relationship with the additional covariates. This results in subject-specific Ising models, where the subject's covariates influence the strength of association between the genes. As in all exploratory data analysis, interpretability of results is important, and we use  $\ell_1$  penalties to induce sparsity in the fitted graphs and in the number of selected covariates. Two algorithms to fit the model are proposed and compared on a set of simulated data, and asymptotic results are established. The results on the tumor dataset and their biological significance are discussed in detail.

**KEY WORDS:** Graphical model, Lasso, Ising model, Binary Markov network, covariates.

# 1 Introduction

Markov networks have been applied in a wide range of scientific and engineering problems to infer the local conditional dependency of the variables. Examples include gene association studies (Peng et al., 2009; Wang et al., 2011), image processing (Hassner & Sklansky, 1980; Woods, 1978), and natural language processing (Manning & Schutze, 1999). A pairwise Markov network can be represented by an undirected graph  $G = (V, E)$ , where  $V$  is the node set representing the collection of random variables, and  $E$  is the edge set where the existence of an edge is equivalent to the conditional dependency between the corresponding pair of variables, given the rest of the graph.

Previous studies have focused on the case where an i.i.d. sample is drawn from an underlying Markov network, and the goal is to recover the graph structure, i.e., the edge set  $E$ , from the data. Two types of graphical models have been studied extensively: the multivariate Gaussian model for continuous data, and the Ising model (Ising, 1925) for binary data. In the multivariate Gaussian case, the graph structure  $E$  is completely specified by the off-diagonal elements of the inverse covariance matrix, also known as the precision matrix. Therefore, estimating the edge set  $E$  is equivalent to identifying the non-zero off-diagonal entries of the precision matrix. Many papers on estimating the inverse covariance matrix have appeared in recent years, with a focus on the high-dimensional framework, for example, Meinshausen & Bühlmann (2006); Yuan & Lin (2007); Rothman et al. (2008); d’Aspremont et al. (2008); Rocha et al. (2008); Ravikumar et al. (2008); Lam & Fan (2009); Peng et al. (2009); Yuan (2010); Cai et al. (2011b). Most of these papers focus on penalized likelihood methods, and many establish asymptotic properties such as consistency and sparsistency. Many have also proposed fast computational algorithms, the most popular of which is perhaps glasso by Friedman et al. (2008), which was recently improved further by Witten et al. (2011) and Mazumder & Hastie (2012).

In the Ising model, the network structure can be identified from the coefficients of the interaction terms in the probability mass function. The problem is, however, considerably more difficult due to the intractable normalizing constant, which makes the penalized likelihood methods popular for the Gaussian case extremely computationally demanding. Ravikumar et al. (2010) proposed an approach in the spirit of Meinshausen & Bühlmann (2006)'s work for the Gaussian case, fitting separate  $\ell_1$ -penalized logistic regressions for each node to infer the graph structure. A pseudo-likelihood based algorithm was developed by Höfling & Tibshirani (2009) and analyzed by Guo et al. (2010c).

The existing literature mostly assumes that the data are an i.i.d. sample from one underlying graphical model, although the case of data sampled from several related graphical models on the same nodes has been studied both for the Gaussian and binary cases Guo et al. (2010b,a). However, in many real-life situations, the structure of the network may further depend on other extraneous factors available to us in the form of explanatory variables or covariates, which result in subject-specific graphical models. For example, in genetic studies, deletion of tumor suppressor genes plays a crucial role in tumor initiation and development. Since genes function through complicated regulatory relationships, it is of interest to characterize the associations among various deletion events in tumor samples. However, in practice we observe not only the deletion events, but also various clinical phenotypes for each subject, such as tumor category, mutation status, and so on. These additional factors may influence the regulatory relationships, and thus should be included in the model. Motivated by situations like this, here we propose a model for the conditional distribution of binary network data given covariates, which naturally incorporates covariate information into the Ising model, allowing the strength of the connection to depend on the covariates. With high-dimensional data in mind, we impose sparsity in the model, both in the network structure and in covariate effects. This allows us to select important covariates that have influence on the network structure.

There have been a few recent papers on graphical models that incorporate covariates, but they do so in ways quite different from ours. Yin & Li (2011) and Cai et al. (2011a) proposed to use conditional Gaussian graphical models to fit the eQTL (gene expression quantitative loci) data, but only the mean is modeled as a function of covariates, and the network remains fixed across different subjects. Liu et al. (2010) proposed a graph-valued regression, which partitions the covariate space and fits separate Gaussian graphical models for each region using glasso. This model does result in different networks for different subjects, but lacks interpretation of the relationship between covariates and the graphical model. Further, there is a concern about stability, since the so built graphical models for nearby regions of the covariates are not necessarily similar. In our model, covariates are incorporated directly into the conditional Ising model, which leads to straightforward interpretation and “continuity” of the graphs as a function of the covariates, since in our model it is the strength of the edges rather than the edges themselves that change from subject to subject.

The rest of the paper is organized as follows. In Section 2, we describe the conditional Ising model with covariates, and two estimation procedures for fitting it. Section 3 establishes asymptotic properties of the proposed estimation method. We evaluate the performance of our method on simulated data in Section 4, and apply it to a dataset on genomic instability in breast cancer samples in Section 5. Section 6 concludes with a summary and discussion.

## 2 Conditional Ising model with covariates

### 2.1 Model set-up

We start from a brief review of the Ising model, originally proposed in statistical physics by Ising (1925). Let  $\mathbf{y} = (y_1, \dots, y_q) \in \{0, 1\}^q$  denote a binary random vector. The Ising model specifies the probability mass function  $P_{\boldsymbol{\theta}}(\mathbf{y})$  as

$$P_{\boldsymbol{\theta}}(\mathbf{y}) = \frac{1}{Z(\boldsymbol{\theta})} \exp \left( \sum_j \theta_{jj} y_j + \sum_{k>j} \theta_{jk} y_j y_k \right),$$

where  $\boldsymbol{\theta} = (\theta_{11}, \theta_{12}, \dots, \theta_{q-1q}, \theta_{qq})$  is a  $q(q+1)/2$ -dimensional parameter vector and  $Z(\boldsymbol{\theta})$  is the partition function ensuring the  $2^q$  probabilities summing up to 1. Note that from now on we assume  $\theta_{jk}$  equals to  $\theta_{kj}$  unless otherwise specified. The Markov property is related to the parameter  $\boldsymbol{\theta}$  via

$$\theta_{jk} = 0 \iff y_j \perp y_k \parallel \mathbf{y}_{\setminus(j,k)}, \quad \forall j \neq k, \quad (1)$$

i.e.,  $y_j$  and  $y_k$  are independent given all other  $y$ 's if and only if  $\theta_{jk} = 0$ .

Now suppose we have additional covariate information, and the data are a sample of  $n$  i.i.d. points  $\mathcal{D}_n = \{(\mathbf{x}^1, \mathbf{y}^1), \dots, (\mathbf{x}^n, \mathbf{y}^n)\}$  with  $\mathbf{x}^i \in \mathbb{R}^p$  and  $\mathbf{y}^i \in \{0, 1\}^q$ . We assume that given covariates  $\mathbf{x}$ , the binary response  $\mathbf{y}$  follows the Ising distribution given by

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\boldsymbol{\theta}(\mathbf{x}))} \exp \left( \sum_{j=1}^q \boldsymbol{\theta}_{jj}(\mathbf{x})y_j + \sum_{(j,k):1 \leq k < j \leq q} \boldsymbol{\theta}_{jk}(\mathbf{x})y_j y_k \right). \quad (2)$$

We note that for any covariates  $\mathbf{x}^i$ , the conditional Ising model is fully specified by the vector  $\boldsymbol{\theta}(\mathbf{x}^i) = (\boldsymbol{\theta}_{11}(\mathbf{x}^i), \boldsymbol{\theta}_{12}(\mathbf{x}^i), \dots, \boldsymbol{\theta}_{q-1q}(\mathbf{x}^i), \boldsymbol{\theta}_{qq}(\mathbf{x}^i))$ , and by setting  $\boldsymbol{\theta}_{kj}(\mathbf{x}) = \boldsymbol{\theta}_{jk}(\mathbf{x})$  for all  $j > k$ , the functions  $\boldsymbol{\theta}_{jk}(\mathbf{x})$  can be connected to conditional log-odds in the following way,

$$\log \left( \frac{P(y_j = 1 | \mathbf{y}_{\setminus j}, \mathbf{x})}{1 - P(y_j = 1 | \mathbf{y}_{\setminus j}, \mathbf{x})} \right) = \boldsymbol{\theta}_{jj}(\mathbf{x}) + \sum_{k:k \neq j} \boldsymbol{\theta}_{jk}(\mathbf{x})y_k, \quad (3)$$

where,  $\mathbf{y}_{\setminus j} = (y_1, \dots, y_{j-1}, y_{j+1}, \dots, y_q)$ . Further, conditioning on  $\mathbf{y}_{\setminus\{j,k\}}$  being 0, we also have

$$\log \left( \frac{P(y_j = 1, y_k = 1 | \mathbf{y}_{\setminus\{j,k\}}, \mathbf{x})P(y_j = 0, y_k = 0 | \mathbf{y}_{\setminus\{j,k\}}, \mathbf{x})}{P(y_j = 1, y_k = 0 | \mathbf{y}_{\setminus\{j,k\}}, \mathbf{x})P(y_j = 0, y_k = 1 | \mathbf{y}_{\setminus\{j,k\}}, \mathbf{x})} \right) = \boldsymbol{\theta}_{jk}(\mathbf{x}).$$

Similarly to (1), this implies  $y_j$  and  $y_k$  are conditionally independent given covariates  $\mathbf{x}$  and all other  $y$ 's if and only if  $\boldsymbol{\theta}_{jk}(\mathbf{x}) = 0$ .

A natural way to model  $\boldsymbol{\theta}_{jk}(\mathbf{x})$  is to parametrize it as a linear function of  $\mathbf{x}$ . Specifically, for  $1 \leq j \leq k \leq q$ , we let

$$\begin{aligned} \boldsymbol{\theta}_{jk}(\mathbf{x}) &= \theta_{jk0} + \boldsymbol{\theta}_{jk}^T \mathbf{x}, & \text{where } \boldsymbol{\theta}_{jk}^T &= (\theta_{jk1}, \dots, \theta_{jkp}) \\ \boldsymbol{\theta}_{jk}(\mathbf{x}) &= \boldsymbol{\theta}_{kj}(\mathbf{x}), & \forall j > k \end{aligned}$$

The model can be expressed in terms of the parameter vector  $\boldsymbol{\theta} = (\theta_{110}, \boldsymbol{\theta}_{11}^T, \theta_{120}, \boldsymbol{\theta}_{12}^T, \dots, \theta_{qq0}, \boldsymbol{\theta}_{qq}^T)$  as follows:

$$P_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\boldsymbol{\theta}(\mathbf{x}))} \exp \left( \sum_{j=1}^q (\theta_{jj0} + \boldsymbol{\theta}_{jj}^T \mathbf{x}) y_j + \sum_{k>j} (\theta_{jk0} + \boldsymbol{\theta}_{jk}^T \mathbf{x}) y_j y_k \right). \quad (4)$$

Instead of (3), we now have the log-odds that depend on the covariates, through

$$\log \left( \frac{P(y_j = 1 | \mathbf{y}_{\setminus j}, \mathbf{x})}{1 - P(y_j = 1 | \mathbf{y}_{\setminus j}, \mathbf{x})} \right) = \theta_{jj0} + \boldsymbol{\theta}_{jj}^T \mathbf{x} + \sum_{k:k \neq j} (\theta_{jk0} + \boldsymbol{\theta}_{jk}^T \mathbf{x}) y_k. \quad (5)$$

The choice of linear parametrization for  $\boldsymbol{\theta}_{jk}(\mathbf{x})$  has several advantages. First, (5) mirrors the logistic regression model when viewing the  $x_\ell$ 's,  $y_k$ 's and  $x_\ell y_k$ 's ( $k \neq j$ ) as predictors. Thus the model has the same interpretation as the logistic regression model, where each parameter describes the size of the conditional contribution of that particular predictor. Second, this parametrization has a straightforward relationship to the Markov network. One can tell which edges exist and on which covariates they depend by simply looking at  $\boldsymbol{\theta}$ . Specifically, the vector  $(\theta_{jk0}, \boldsymbol{\theta}_{jk}^T)$  being zero implies that  $y_k$  and  $y_j$  are conditionally independent given any  $\mathbf{x}$  and the rest of  $y_\ell$ 's, and  $\theta_{jk\ell}$  being zero implies that the conditional association between  $y_j$  and  $y_k$  does not depend on  $x_\ell$ . Third, the continuity of linear functions ensures the similarity among the conditional models for similar covariates, which is a desirable property. Finally, the linear formulation promises the convexity of the negative log-likelihood function, allowing efficient algorithms for fitting the model discussed next.

## 2.2 Fitting the model

The probability model  $P_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})$  in (4) includes the partition function  $Z(\boldsymbol{\theta}(\mathbf{x}))$ , which requires summation of  $2^q$  terms for each data point and makes it intractable to directly maximize the joint conditional likelihood  $\sum_{i=1}^n \log P_{\boldsymbol{\theta}}(\mathbf{y}^i | \mathbf{x}^i)$ . However, (5) suggests we can use logistic regression to estimate the parameters, an approach in the spirit of Ravikumar et al. (2010). The idea is essentially to maximize the conditional log-likelihood of  $y_j^i$  given  $\mathbf{y}_{\setminus j}^i$  and  $\mathbf{x}^i$  rather than the joint log-likelihood of  $\mathbf{y}^i$ .

Specifically, the negative conditional log-likelihood for  $y_j$  can be written as follows

$$\ell_j(\boldsymbol{\theta}; \mathcal{D}_n) = -\frac{1}{n} \sum_{i=1}^n \log P(y_j^i | \mathbf{x}^i, \mathbf{y}_{\setminus j}^i) = -\frac{1}{n} \sum_{i=1}^n \left( \log(1 + e^{\eta_j^i}) - y_j^i \eta_j^i \right), \quad (6)$$

where

$$\eta_j^i = \log \left( \frac{P(y_j^i = 1 | \mathbf{y}_{\setminus j}^i, \mathbf{x}^i)}{1 - P(y_j^i = 1 | \mathbf{y}_{\setminus j}^i, \mathbf{x}^i)} \right) = \boldsymbol{\theta}_{jj}^T \mathbf{x}^i + \sum_{k \neq j} (\theta_{jk0} + \boldsymbol{\theta}_{jk}^T \mathbf{x}^i) y_k^i.$$

Note that this conditional log-likelihood involves the parameter vector  $\boldsymbol{\theta}$  only through its subvector  $\boldsymbol{\theta}_j = (\theta_{j10}, \boldsymbol{\theta}_{j1}^T, \dots, \theta_{jq0}, \boldsymbol{\theta}_{jq}^T) \in \mathbb{R}^{(p+1)q}$ , thus we sometimes write  $\ell_j(\boldsymbol{\theta}_j; \mathcal{D}_n)$  when the rest of  $\boldsymbol{\theta}$  is not relevant.

There are  $(p+1)q(q+1)/2$  parameters to be estimated, so even for moderate  $p$  and  $q$  the dimension of  $\boldsymbol{\theta}$  can be large. For example, with  $p = 10$  and  $q = 10$ , the model has 605 parameters. Thus there is a need to regularize  $\boldsymbol{\theta}$ . Empirical studies of networks as well as the need for interpretation suggest that a good estimate of  $\boldsymbol{\theta}$  should be sparse. Thus we adopt the  $\ell_1$  regularization to encourage sparsity, and propose two approaches to maximize the conditional likelihood (6).

### Separate regularized logistic regressions

The first approach is to estimate each  $\boldsymbol{\theta}_j$ ,  $j = 1, \dots, q$  separately using the following criterion,

$$\min_{\boldsymbol{\theta}_j \in \mathbb{R}^{(p+1)q}} \ell_j(\boldsymbol{\theta}_j; \mathcal{D}_n) + \lambda \|\boldsymbol{\theta}_{j \setminus 0}\|_1,$$

where  $\boldsymbol{\theta}_{j \setminus 0} = \boldsymbol{\theta}_j \setminus \{\theta_{jj0}\}$ , that is, we do not penalize the intercept term  $\theta_{jj0}$ .

In this approach,  $\boldsymbol{\theta}_{jk}$  and  $\boldsymbol{\theta}_{kj}$  are estimated from the  $j$ th and  $k$ th regressions, respectively, thus the symmetry  $\hat{\boldsymbol{\theta}}_{jk} = \hat{\boldsymbol{\theta}}_{kj}$  is not guaranteed. To enforce the symmetry in the final estimate, we post-process the estimates following Meinshausen & Bühlmann (2006), where the initial estimates are combined by comparing their magnitudes. Specifically, let  $\hat{\boldsymbol{\theta}}_{jkl}$  denote the final estimate and  $\hat{\boldsymbol{\theta}}_{jkl}^0$  denote the initial estimate from the separate regularized logistic regressions. Then for any  $1 \leq j < k \leq q$  and any  $l = 0, \dots, p$ , we can use one of the

two symmetrizing approaches:

$$\begin{aligned} \text{separate-max:} \quad & \hat{\theta}_{jk\ell} = \hat{\theta}_{kj\ell} = \hat{\theta}_{jk\ell}^0 \mathbb{I}_{(|\hat{\theta}_{jk\ell}^0| > |\hat{\theta}_{kj\ell}^0|)} + \hat{\theta}_{kj\ell}^0 \mathbb{I}_{(|\hat{\theta}_{jk\ell}^0| < |\hat{\theta}_{kj\ell}^0|)} \\ \text{separate-min:} \quad & \hat{\theta}_{jk\ell} = \hat{\theta}_{kj\ell} = \hat{\theta}_{jk\ell}^0 \mathbb{I}_{(|\hat{\theta}_{jk\ell}^0| < |\hat{\theta}_{kj\ell}^0|)} + \hat{\theta}_{kj\ell}^0 \mathbb{I}_{(|\hat{\theta}_{jk\ell}^0| > |\hat{\theta}_{kj\ell}^0|)} \end{aligned}$$

The separate-min approach is always more conservative than separate-max in the sense that the former provides more zero estimates. It turns out that when the sample size is small, the separate-min approach is often too conservative to effectively identify non-zero parameters. More details are given in Section 4.

### Joint regularized logistic regression

The second approach is to estimate the entire vector  $\boldsymbol{\theta}$  simultaneously instead of estimating the  $\boldsymbol{\theta}_j$ 's separately, using the criterion,

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^{(p+1)q(q+1)/2}} \sum_{j=1}^q \ell_j(\boldsymbol{\theta}; \mathcal{D}_n) + \lambda \|\boldsymbol{\theta}_{\setminus 0}\|_1,$$

where  $\boldsymbol{\theta}_{\setminus 0} = \boldsymbol{\theta} \setminus \{\theta_{110}, \theta_{220}, \dots, \theta_{qq0}\}$ . The joint approach criterion can be written as one large penalized logistic regression by careful rearranging of terms. One obvious benefit of the joint approach is that  $\hat{\boldsymbol{\theta}}$  can be automatically symmetrized by treating  $\boldsymbol{\theta}_{jk}$  and  $\boldsymbol{\theta}_{kj}$  as the same during estimation. The price, however, is that it is computationally much less efficient than the separate approach.

To fit the model using either the separate or the joint approach, we adopt the coordinate shooting algorithm in Fu (1998), where we update one parameter at a time and iterate until convergence. The implementation is similar to the glmnet algorithm of Friedman et al. (2010), and we omit the details here.

## 3 Asymptotics: consistency of model selection

In this section we present the model selection consistency property for the separate regularized logistic regression. Results for the joint approach can be derived in the same fashion



by treating the joint regression as a single large logistic regression. The spirit of the proof is similar to Ravikumar et al. (2010), but since their model does not include covariates  $\mathbf{x}$ , both our assumptions and conclusions are different.

In this analysis, we treat the covariates  $\mathbf{x}_i$ 's as random vectors. With a slight change of notation, we now use  $\boldsymbol{\theta}_j$  to denote  $\boldsymbol{\theta}_{j \setminus 0}$ , dropping the intercept which is irrelevant for model selection. The true parameter is denoted by  $\boldsymbol{\theta}^*$ . Without loss of generality we assume that  $\theta_{jj0}^* = 0$ , and we also assume that  $\hat{\theta}_{jj0} = 0$ .

First, we introduce additional notation to be used throughout this section. Let

$$\mathbf{I}_j^* = \mathbb{E}_{\boldsymbol{\theta}^*}(\nabla^2 \log P_{\boldsymbol{\theta}}(y_j | \mathbf{x}, \mathbf{y}_{\setminus j})) \quad (7)$$

$$= \mathbb{E}_{\boldsymbol{\theta}^*} (p_j(1-p_j)(\mathbf{x} \otimes \mathbf{y}_{\setminus j})(\mathbf{x} \otimes \mathbf{y}_{\setminus j})^T) \quad (\text{Information matrix}) \quad (8)$$

$$\mathbf{U}_j^* = \mathbb{E}_{\boldsymbol{\theta}^*} ((\mathbf{x} \otimes \mathbf{y}_{\setminus j})(\mathbf{x} \otimes \mathbf{y}_{\setminus j})^T) \quad (9)$$

where

$$p_j = p_j(\mathbf{x}, \mathbf{y}_{\setminus j}) = P_{\boldsymbol{\theta}^*}(y_j = 1 | \mathbf{x}, \mathbf{y}_{\setminus j}) ,$$

$$\mathbf{x} \otimes \mathbf{y}_{\setminus j} = (1, x_1, \dots, x_p)^T \otimes (y_1, \dots, y_{j-1}, 1, y_{j+1}, \dots, y_q)^T \setminus \{1\} .$$

Let  $\mathcal{S}_j$  denote the index set of the non-zero elements of  $\boldsymbol{\theta}_j^*$ , and let  $\mathbf{I}_{\mathcal{S}_j \mathcal{S}_j}^*$  be the submatrix of  $\mathbf{I}_j^*$  indexed by  $\mathcal{S}_j$ . Similarly defined are  $\mathbf{I}_{\mathcal{S}_j^c \mathcal{S}_j}$  and  $\mathbf{I}_{\mathcal{S}_j \mathcal{S}_j^c}$ , where  $\mathcal{S}_j^c$  is the compliment set of  $\mathcal{S}_j$ . Moreover, for any matrix  $A$ , let  $\|A\|_{\infty} = \max_i \sum_j |A_{ij}|$  be the matrix  $L_{\infty}$  norm, and let  $\Lambda_{\min}(A)$  and  $\Lambda_{\max}(A)$  be the minimum and maximum eigenvalues of  $A$ , respectively.

For our main results to hold, we make the following two assumptions for all  $q$  logistic regressions.

**A1** There exists a constant  $\alpha \in (0, 1]$ , such that

$$\|\mathbf{I}_{\mathcal{S}_j^c \mathcal{S}_j}^* \left( \mathbf{I}_{\mathcal{S}_j \mathcal{S}_j}^* \right)^{-1}\|_{\infty} \leq (1 - \alpha) .$$

**A2** There exist constants  $\Delta_{\min} > 0$  and  $\Delta_{\max} > 0$ , such that

$$\begin{aligned}\Lambda_{\min}(\mathbf{I}_{\mathcal{S}_j}^*) &\geq \Delta_{\min} \\ \Lambda_{\max}(\mathbf{U}_j^*) &\leq \Delta_{\max}\end{aligned}$$

These assumptions bound the correlation among the effective covariates, and the amount of dependence between the group of effective covariates and the rest. Under these assumptions, we have the following result:

**Theorem 1** For any  $j = 1, \dots, q$ , let  $\hat{\boldsymbol{\theta}}_j$  be a solution of the problem

$$\min_{\boldsymbol{\theta}_j} -\ell_j(\boldsymbol{\theta}_j; \mathcal{D}_n) + \lambda_n \|\boldsymbol{\theta}_j\|_1. \quad (10)$$

Assume **A1** and **A2** hold for  $\mathbf{I}_j^*$  and  $\mathbf{U}_j^*$ , and further assume that for some  $\delta > 0$

$$P(\|\boldsymbol{x}\|_{\infty} \geq M) \leq \exp(-M^{\delta}), \quad \text{for all } M \geq M_0 > 0, \quad (11)$$

Let  $d = \max_j \|\mathcal{S}_j\|_0$  and  $C > 0$  a constant independent of  $(n, p, q)$ . If

$$M_n \geq (C\lambda_n^2 n)^{\frac{1}{1+\delta}}, \quad (12)$$

$$\lambda_n \geq CM_n \sqrt{\frac{\log p + \log q}{n}}, \quad (13)$$

$$n \geq CM_n^2 d^3 (\log p + \log q), \quad (14)$$

the following hold with probability at least  $1 - \exp^{-C(\lambda_n^2 n)^{\delta^*}}$  ( $\delta^*$  is a constant in  $(0, 1)$ ),

1. *Uniqueness:*  $\hat{\boldsymbol{\theta}}_j$  is the unique optimal solution for any  $j \in \{1, \dots, q\}$ .
2.  *$\ell_2$  consistency:*  $\|\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j^*\|_2 \leq 5\lambda_n \sqrt{d} / \Delta_{\min}$  for any  $j \in \{1, \dots, q\}$
3. *Sign consistency:*  $\hat{\boldsymbol{\theta}}_j$  correctly identifies all the zeros in  $\boldsymbol{\theta}_j^*$  for any  $j \in \{1, \dots, q\}$ ; moreover,  $\hat{\boldsymbol{\theta}}_j$  identifies the correct sign of non-zeros in  $\boldsymbol{\theta}_j^*$  whose absolute value is at least  $10\lambda_n \sqrt{d} / \Delta_{\min}$ .

Theorem 1 establishes the consistency of model selection allowing both of the dimensions  $p(n)$  and  $q(n)$  to grow to infinity with  $n$ . The extra condition, which requires the distribution of  $\mathbf{x}$  to have a fast decay on large values, was not in Ravikumar et al. (2010) as the paper does not consider covariates. The new condition is, however, quite general; for example, it is satisfied by the Gaussian distribution and all categorical covariates. The proof of the theorem can be found in the Appendix.

## 4 Empirical performance evaluation

In this section, we present three sets of simulation studies designed to test the model selection performance of our methods. We vary different aspects of the model, including sparsity, signal strength and proportion of relevant covariates. The results are presented in the form of ROC curves, where the rate of estimated true non-zero parameters (sensitivity) is plotted against the rate of estimated false non-zero parameters (1-specificity) across a fine grid of the regularization parameter. Each curve is smoothed over 20 replications.

The data generation scheme is as follows. For each simulation, we fix the dimension of the covariates  $p$ , the dimension of the response  $q$ , the sample size  $n$  and a graph structure  $E$  in the form of a  $q \times q$  adjacency matrix (randomly generated scale-free networks (Barabasi & Albert, 1999)). For any  $(j, k)$ ,  $1 \leq j \leq k \leq q$ ,  $(\theta_{jk0}, \boldsymbol{\theta}_{jk}^T)$  consists of  $(p + 1)$  independently generated and selected from three possible values:  $\beta > 0$  (with probability  $\rho/2$ ),  $-\beta$  (with probability  $\rho/2$ ), and 0 (with probability  $1 - \rho$ ). An exception is made for the intercept terms  $\theta_{jj0}$ , where  $\rho$  is always set to 1. Covariates  $\mathbf{x}^i$ 's are generated independently from the multivariate Gaussian distribution  $N_p(0, I_p)$ . Given each  $\mathbf{x}^i$  and  $\boldsymbol{\theta}$ , we use Gibbs sampling to generate the  $\mathbf{y}^i$ , where we iteratively generate a sequence of  $y_j^i$ 's ( $j = 1, \dots, q$ ) from a Bernoulli distribution with probability  $P_{\boldsymbol{\theta}}(y_j^i = 1 | \mathbf{y}_{\setminus j}^i, \mathbf{x}^i)$  and take the last value of the sequence when a stopping criterion is satisfied.

We compared three estimation methods: the separate-min method, the separate-max

method and the joint method. Our simulation results indicate that performance of the separate-min method is substantially inferior to that of the separate-max method in almost all cases (results omitted for lack of space). Thus we only present results for the separate-max and the joint methods in this section.

## 4.1 Effect of sparsity

First, we investigate how the selection performance is affected by the sparsity of the true model. The sparsity of  $\theta$  can be controlled by two factors: the number of edges in  $E$ , denoted by  $n_E$ , and the average proportion of effective covariates for each edge,  $\rho$ . We fix the dimensions  $q = 10$ ,  $p = 20$  and the sample size  $n = 200$ , and set the signal size to  $\beta = 4$ . Under this setting, the total number of parameters is 1155. The sparsity parameter  $n_E$  takes values in the set  $\{10, 20, 30\}$ , and  $\rho$  takes values in  $\{0.2, 0.5, 0.8\}$ .

The resulting ROC curves are shown in Figure 1. The first row shows the results of the joint approach and the second row of the separate-max approach. As the true model becomes less sparse, the performance of both the joint and the separate methods deteriorates, since sparse models have the smallest effective number of parameters to estimate and benefit the most from penalization. Note that the model selection performance seems to depend on the total number of non-zero parameters  $((q+n_E)(p+1)\rho)$ , not just on the number of edges ( $n_E$ ). For example, both approaches perform better in case  $n_E = 20, \rho = 0.2$  than  $n_E = 10, \rho = 0.5$ , even though the former has a more complicated network structure. Comparing the separate-max method and the joint method, we observe that the two methods are quite comparable, with the joint method being slightly less sensitive to increasing the number of edges.

Note that the “\*” point on each curve represents the average sensitivity and (1-specificity) over the replications based on an “optimal”  $\lambda$ , selected by maximizing the conditional log-likelihood on an independent validation dataset of the same size as the training data.

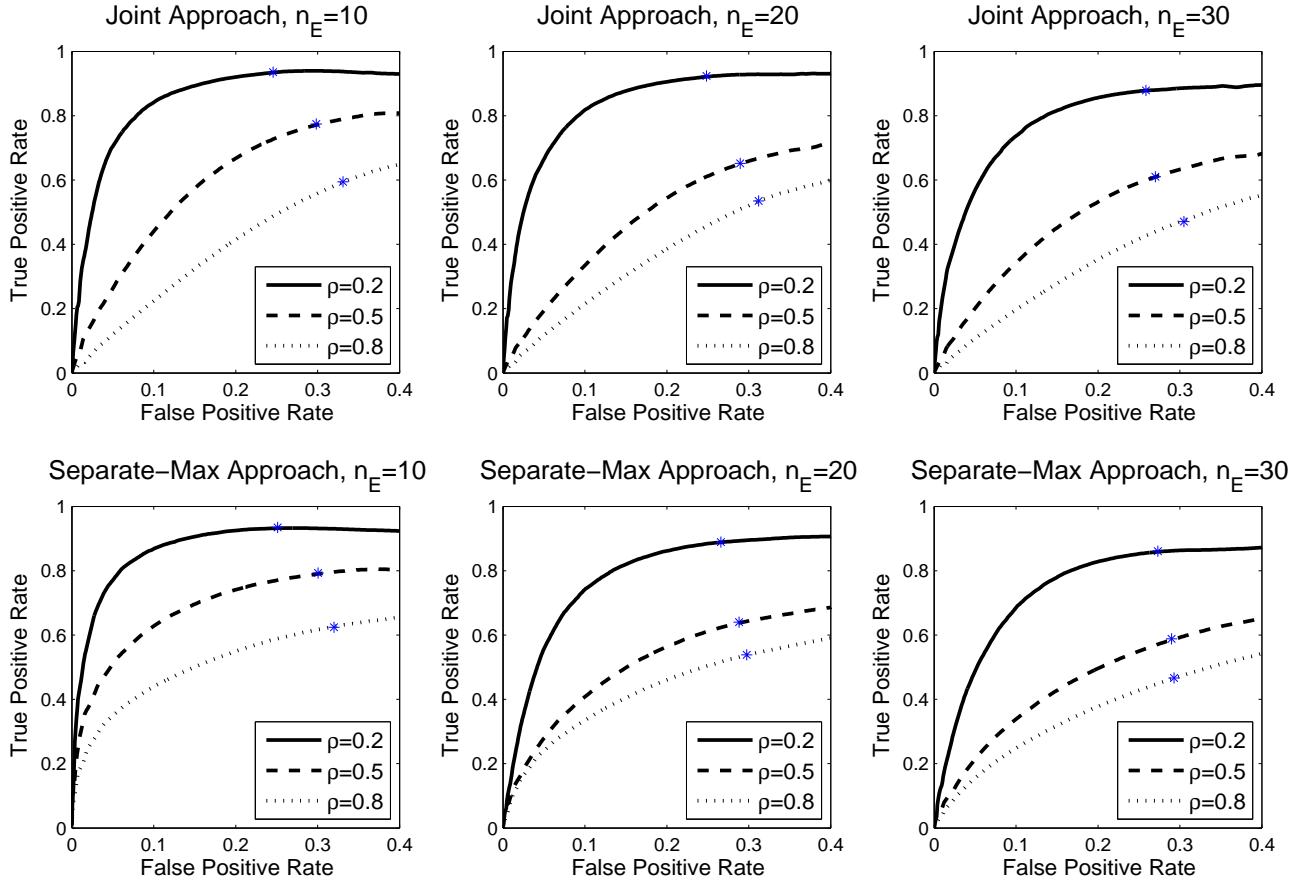


Figure 1: ROC curves for varying levels of sparsity, as measured by the number of edges ( $n_E$ ) and expected proportion of non-zero covariates ( $\rho$ ). The star on each curve corresponds to an optimal value of  $\lambda$  selected on an independent validation set.

## 4.2 Effect of signal size

Second, we assess the effect of signal size. The dimensions are set to be the same as in the previous simulation, that is,  $q = 10$ ,  $p = 20$  and  $n = 200$ , and underlying network is the same. The expected proportion of effective covariates for each edge is  $\rho = 0.5$ . The signal strength parameter  $\beta$  takes values in the set  $\{0.5, 1, 2, 4, 8, 16\}$ . For each setting, the non-zero entries of the parameter vectors  $\theta$  are at the same positions with the same signs, only differing in magnitude. The resulting ROC curves are shown in Figure 2.

As the signal strength  $\beta$  increases, both the separate and the joint methods show improved

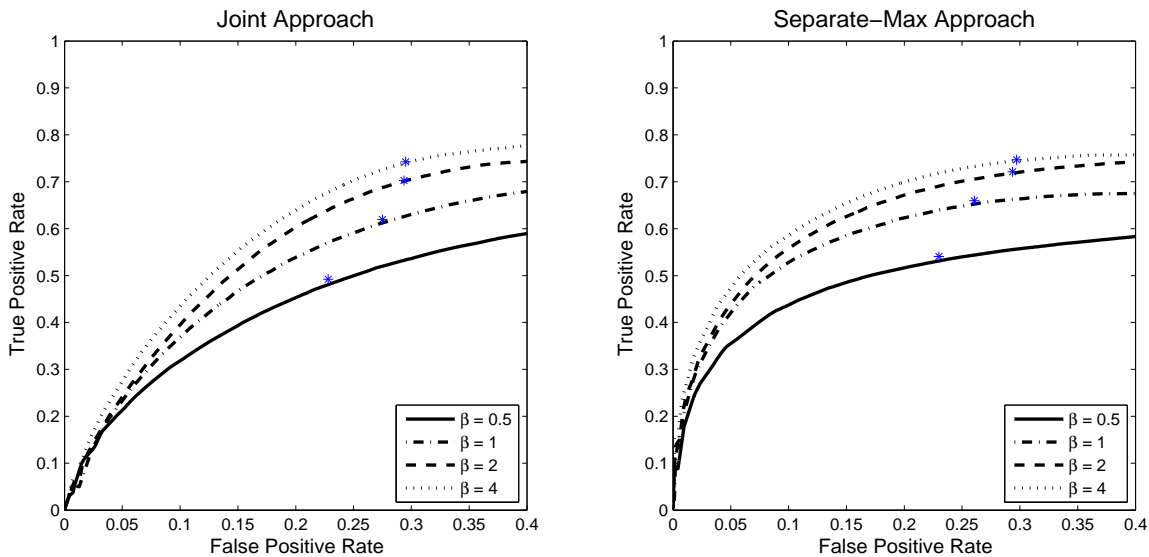


Figure 2: ROC curves for varying levels of signal strength, as measured by the parameter  $\beta$ . The star on each curve corresponds to an optimal value of  $\lambda$  selected on an independent validation set.

selection performance, but the improvement levels off eventually. Both methods achieve almost the same “optimal” sensitivity and specificity (the ‘\*’ point), with the separate-max method performing better overall.

### 4.3 Effect of noise covariates

In the last set of simulations, we study how the model selection performance is affected by adding extra uninformative covariates. At the same time, we also investigate the effect of the number of relevant covariates  $p_{\text{true}}$  and the sample size  $n$ . The dimension of the response is fixed to be  $q = 10$  and the network structure remains the same as in the previous simulation. We take  $p_{\text{true}} \in \{10, 20\}$  and  $n \in \{200, 500\}$ . For each combination, we first fit the model on the original data and then on augmented data with extra uninformative covariates added. The total number of covariates  $p_{\text{total}} \in \{p_{\text{true}}, 50, 200\}$ . The non-zero parameters are generated the same way as before with  $\beta = 4$  and  $\rho = 0.5$ . With the changes in  $p_{\text{total}}$ , the total number of non-zero parameters remains fixed for each value of  $p_{\text{true}}$ , while the total number of zeros is increasing.

To make the results more comparable across setting, we plot the counts rather than rates of true positives and false positives. The resulting curves are shown in Figure 3. Generally, performance improves when the sample size grows and deteriorates when the number of noise covariates increases, particularly with a smaller sample size. The separate-max method dominates the joint method under these settings, but the difference is not large.

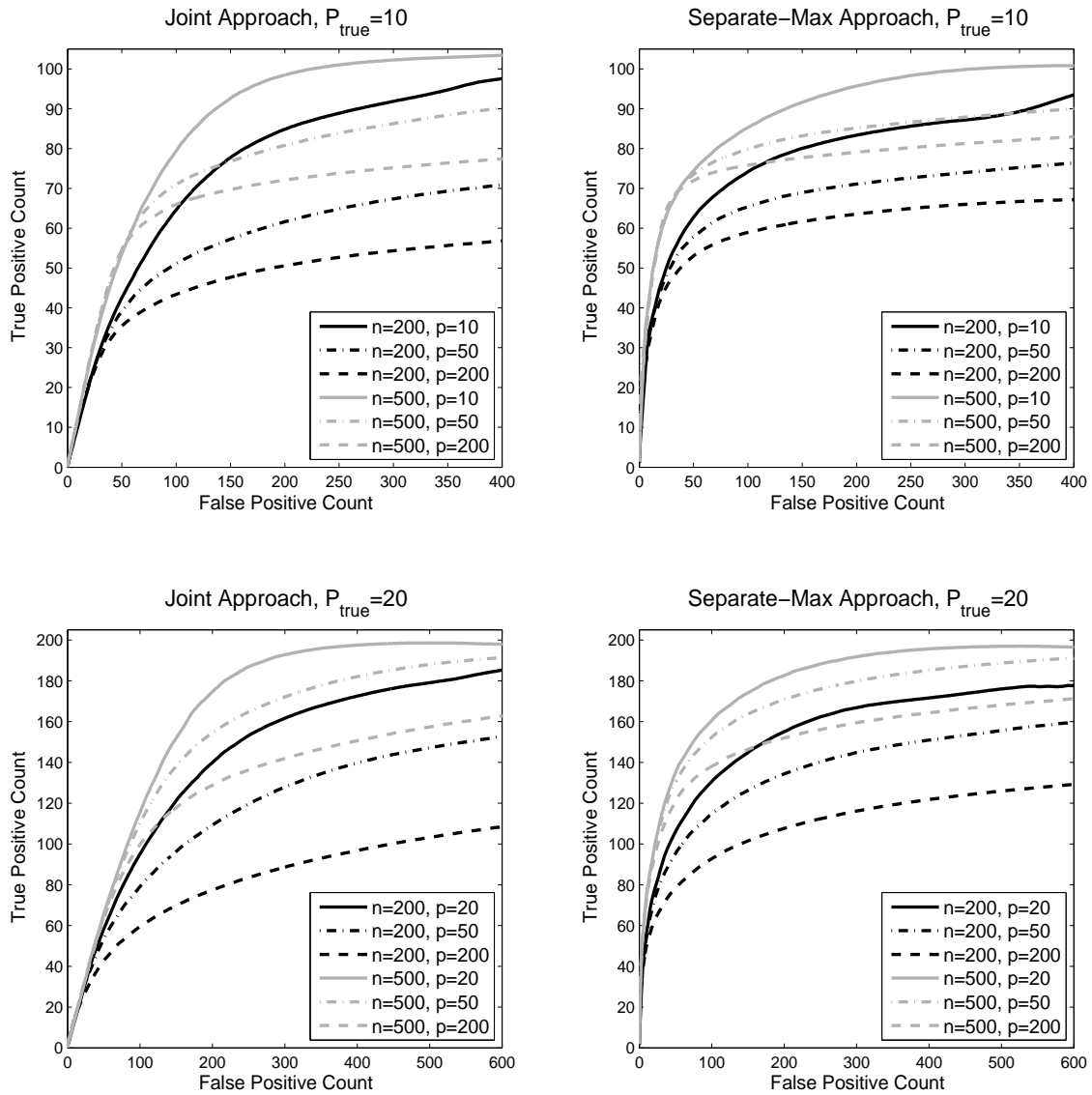


Figure 3: ROC curves for varying dimension, number of noise covariates, and sample size.

## 5 Application to tumor suppressor genes study

In breast cancer, deletion of tumor suppressor genes plays a crucial role in tumor initiation and development. Since genes function through complicated regulatory relationships, it is of interest to characterize the associations among various deletion events in tumor samples, and at the same time to investigate how these association patterns may vary across different tumor subtypes or stages.

Our data set includes DNA copy number profiles from cDNA microarray experiments on 143 breast cancer specimens (Bergamaschi et al., 2006). Among them, 88 samples are from a cohort of Norwegian patients with locally advanced (T3/T4 and/or N2) breast cancer, receiving doxorubicin (Doxo) or 5 fluorouracil/mitomycin C (FUMI) neoadjuvant therapy (Geisler et al., 2003). The samples were collected before the therapy. The other 55 are from another cohort of Norwegian patients from a population-based series (Zhao et al., 2004). Each copy number profile reports the DNA amounts of 39,632 probes in the sample. The array data was preprocessed and copy number gain/loss events were inferred as described in Bergamaschi et al. (2006). To reduce the spatial correlation in the data, we bin the probes by cytogenetic bands (cytobands). For each sample, we define the deletion status of a cytoband to be 1 if at least three probes in this cytoband show copy number loss. 430 cytobands covered by these probes show deletion frequencies greater than 10% in this group of patients, and they were retained for the subsequent analysis. The average deletion rate for all the 430 cytobands in 143 samples is 19.59%. Our goal is to uncover the association among these cytoband-deletion events and how the association patterns may change with different clinical characteristics, including TP53 mutation status (a binary variable), estrogen receptors (ER) status (a binary variable), and tumor stage (an ordinal variable taking values in  $\{1, 2, 3, 4\}$ ).

For our analysis, denote the array data by  $\mathbf{y}_{143 \times 430}$ , where  $y_j^i$  indicates the deletion status



of the  $j^{\text{th}}$  cytoband in the  $i^{\text{th}}$  sample. Let  $\mathbf{x}^i$  denote the covariate vector containing the three clinical phenotypes of the  $i^{\text{th}}$  sample, and  $\mathbf{x}_l$  the  $l$ th covariate vector. We first standardize the covariate matrix  $\mathbf{x}_{143 \times 3}$  and then fit our Ising model with covariates with the separate-max fitting method. We then apply stability selection (Meinshausen & Bühlmann, 2010) to infer the stable set of important covariates for each pairwise conditional association. Specifically, we repeatedly fit the model 100 times on subsamples containing half the data selected randomly without replacement. For each tuning parameter  $\lambda$  from a fixed grid of values, we record the frequency of  $\hat{\theta}_{jkl}$  being non-zero respectively for each covariate  $\mathbf{x}_l$ ,  $l = 0, 1, 2, 3$  on all pairs of  $(j, k)$ ,  $1 \leq j < k \leq 430$ , and denote it by  $f_{jkl}(\lambda)$ . Note that  $\mathbf{x}_0$  corresponds to the main effect interaction between a pair of  $\mathbf{y}_j$ 's and does not involve any covariates. Then we use  $f_{jkl}^* = \max_{\lambda} f_{jkl}(\lambda)$  as a measure of importance of covariate  $\mathbf{x}_l$  for the edge  $(j, k)$ . Finally, for each covariate  $\mathbf{x}_j$ , we rank the edges based on the selection frequencies  $\{f_{jkl}^* : 1 < j \leq k < q\}$ . At the top of the list are the edges that depend on  $\mathbf{x}_j$  most heavily. We are primarily interested in the pairs of genes belonging to different chromosomes, as the interaction between genes located on the same chromosome is more likely explained by strong local dependency. The results are shown in Table 1, where the rank list of the edges depending on different covariates are recorded. The first two columns of each covariate related columns are the node names and the third columns record the selection frequency.

There are 332 inter-chromosome interactions (between cytobands from different chromosomes) with selection probabilities at least 0.5. Among these, 39 interactions change with the TP53 status; 12 change with the ER status; and another 12 change with the tumor grade (see details in Table 1). These results can be used by biologists to generate hypotheses and design relevant experiments to better understand the molecular mechanism of breast cancer.

The most frequently selected pairwise conditional association is between deletion on cytoband 4q31.3 and deletion on 18q23 (94% selection frequency). Cytoband 4q31.3 harbors

the tumor suppressor candidate gene SCFFbw7, which works cooperatively with gene TP53 to restrain cyclin E-associated genome instability (Minella et al., 2007). Previous studies also support the existence of putative tumor suppressor loci at cytoband 18q23 distal to the known tumor suppressor genes SMAD4, SMAD2 and DCC (Huang et al., 1995; Lassus et al., 2001). Thus the association between the deletion events on these two cytobands is intriguing.

Another interesting finding is that the association between deletion on cytoband 9q22.3 region and cytoband 12p13.31 appears to be stronger in the TP53 positive group than in the TP53 negative group. A variety of chromosomal aberrations at 9p22.3 have been found in different malignancies including breast cancer (Mitelman et al., 1997). This region contains several putative tumor suppressor genes (TSG), including DNA-damage repair genes like FANCC and XPA. Alterations in these TSGs have been reported to be associated with poor patient survival (Sinha et al., 2008). On the other hand, cytoband 12p13.31 harbors another TSG, namely ING4 (inhibitor of growth family member 4), whose protein binds TP53 and contributes to the TP53-dependent regulatory pathway. A recent study also suggests involvement of ING4 deletion in the pathogenesis of HER2-positive breast cancer. In light of these previous findings, it is interesting that our analysis also found the association between the deletion events of 9p22.3 and 12p13.31, as well as the changing pattern of the association under different TP53 status. This result suggests potential cooperative roles for multiple tumor suppressor genes in cancer initiation and progression.

We also searched the network for hubs (highly connected nodes), which often have important roles in genetic regulatory pathways. Since there can be different hubs associated with different covariates, we separate them as follows. For each node  $j$ , covariate  $l$ , and stability selection subsample  $m$ , let the “covariate-specific” degree of node  $j$  be  $d_{j,l}^m = \#\{k : \hat{\theta}_{jkl} \neq 0\}$ . A ranking of nodes can then be produced for each covariate  $l$  and each replication  $m$ , with  $r_{j,l}^m$  being the corresponding rank. Finally, we compute the median rank across all stability

Table 1: Frequency-based ranked list of covariate-dependent inter-chromosomal interactions

Main effect			TP53 mutation status			ER status		
Gene1	Gene2	Freq	Gene1	Gene2	Freq	Gene1	Gene2	Freq
4q31.3	18q23	0.95	3p22.2	22q13.1	0.79	3q26.1	11p14.3	0.69
2p25.2	15q26.2	0.87	3p12.3	12p13.1	0.72	4q34.3	5q32	0.64
2q36.3	3p26.1	0.84	12q22	15q14	0.7	8p11.22	11p14.2	0.63
7q21.13	8q21.13	0.84	2p12	Xp22.33	0.69	3q24	22q11.23	0.57
6p21.32	16q12.2	0.83	6p21.32	8p11.22	0.68	4p14	11p15.3	0.55
3p21.1	17p13.2	0.81	1p34.2	3p24.1	0.67	1q31.1	Xq27.3	0.54
4q24	12q21.1	0.81	2p21	Xp11.22	0.67	13q33.2	22q11.23	0.54
2q23.3	6p12.1	0.79	2p12	7p21.1	0.66	21q21.1	22q11.21	0.54
8p21.3	21q21.1	0.79	12q15	13q12.12	0.63	5q33.1	17q21.31	0.53
2q34	3q13.31	0.78	4q25	8p11.22	0.62	12q21.32	18q22.3	0.51
6p21.32	9q31.3	0.78	8p11.22	Xq23	0.62	8p11.22	22q11.21	0.5
6p21.32	13q21.1	0.78	9p21.2	16q22.1	0.61	8q21.13	Xp22.11	0.5
6p21.31	11p15.2	0.78	3p21.1	11q14.1	0.58			
11p15.1	14q22.2	0.78	3p13	9p24.2	0.58			
1p36.11	2p21	0.77	9q22.32	12p13.31	0.57			
1p31.1	2q32.2	0.76	7q21.3	22q12.3	0.56			
1q31.1	22q11.21	0.76	3q26.1	11p13	0.55			
2q32.1	6q14.1	0.76	4q35.2	22q12.3	0.55			
9q21.11	16q21	0.76	15q22.33	17p11.2	0.55			
9q31.3	14q24.3	0.76	3p22.1	6p21.31	0.54			
10q25.3	12p13.31	0.76	4q28.2	7q21.13	0.54			
4q35.1	15q22.2	0.75	5q13.1	6q22.33	0.54			
3p21.31	17p11.2	0.74	5q23.2	8p21.2	0.54			
6p21.32	13q31.2	0.74	16q22.1	17q21.31	0.54			
10q11.21	12p13.32	0.74	4q28.3	9p21.3	0.53			
9q33.1	14q12	0.73	4q35.1	9p21.3	0.53			
12p13.31	17q11.2	0.73	4q35.2	16q22.1	0.53			
1p34.2	3p22.1	0.72	2q31.3	4q13.2	0.52			
5q33.1	11p15.4	0.72	3p26.1	14q13.1	0.52			
6q12	20p12.1	0.72	4p16.1	13q31.1	0.52			
12p12.2	Xp11.4	0.72	6p21.31	11q14.2	0.52			
4q35.2	9p21.2	0.71	3p25.1	11p15.2	0.51			
11p15.2	18q12.1	0.71	5q14.2	Xq27.1	0.51			
1p21.1	7q21.12	0.7	5q14.2	Xq27.2	0.51			
2p16.1	6p12.3	0.7	8p11.22	15q14	0.51			
2q31.2	3p26.2	0.7	10q23.32	21q21.1	0.51			
2q36.3	9q22.31	0.7	16q22.1	17p13.2	0.51			
3p22.1	15q25.3	0.7	3p22.1	5q33.3	0.5			
6p21.32	Xp11.4	0.7	5q14.2	17q21.2	0.5			

selection subsamples  $r_{j,l} = \text{median}\{r_{j,l}^m, m = 1, \dots, 100\}$ , and order nodes by rank for each covariate. The results are listed in Table 2. Interestingly, cytoband 8p11.22 was ranked close to the top for all three covariates. The 8p11-p12 genomic region plays an important roles in breast cancer, as numerous studies have identified this region as the location of multiple oncogenes and tumor suppressor genes (Yang et al., 2006; Adelaide et al., 1998). High frequency of loss of heterozygosity (LOH) of this region in breast cancer has also been reported (Adelaide et al., 1998). Particularly, cytoband 8p11.22 harbors the candidate tumor suppressor gene TACC1 (transforming, acidic coiled-coil containing protein 1), whose alteration is believed to disturb important regulations and participate in breast carcinogenesis (Conte et al., 2002). From Table 1, we can also see that the deletion of cytoband 8p11.22 region is associated with the deletion of cytoband 6p21.32 and 11p14.2 with relatively high confidence (selection frequency  $> 0.6$ ); and these associations change with both TP53 status and ER status. This finding is interesting because high frequency LOH at 6q and 11p in breast cancer cells are among the earliest findings that led to the discovery of recessive tumor suppressor genes of breast cancer (Ali et al., 1987; Devlce et al., 1991; Negrini et al., 1994). Moreover, there is evidence that allele loss of c-Ha-ras locus at 11p14 correlates with paucity of oestrogen receptor protein, as well as patient survival (MacKay et al., 1988; Garcia et al., 1989). These results together with the associations we detected confirm the likely cooperative roles of multiple tumor suppressor genes involved in breast cancer.

## 6 Summary and Discussion

We have proposed a novel Ising graphical model which allows us to incorporate extraneous factors into the graphical model in the form of covariates. Including covariates into the model allows for subject-specific graphical models, where the strength of association between nodes varies smoothly with the values of covariates. One consequence of this is that if all covariates are continuous, there is probability 0 of the graph structure changing with covariates, and

Table 2: Degree-based ranking of nodes

Main effect		TP53 mutation status		ER status		Tumor stage	
Gene	Median rank	Gene	Median rank	Gene	Median rank	Gene	Median rank
1p36.11	16.75	8p11.22	12.75	3q26.1	10	16q23.1	19.25
1q31.1	21	1p31.3	14.5	1q31.1	12	10q11.23	22.25
6p21.31	24.25	3p22.2	25.25	3p22.2	13	16q12.2	23.5
6p21.32	37	1q31.1	28.75	8q21.13	14	9q34.13	27.5
2p12	38.5	12q23.1	32	10q22.1	15.25	22q11.23	27.75
2q32.2	43	2p16.2	33.5	8p11.22	19	12p11.23	33
8q21.13	44.5	4q31.1	41.75	3p21.1	20.25	2q33.1	35.25
6p12.3	45.5	9p21.3	42	11q23.3	22	8p11.22	35.75
2q32.3	53.75	7q21.3	44.25	5q13.1	28	10q25.2	36
3p22.2	54.25	3q26.1	44.75	4p16.1	33	11q14.1	40.5
6p12.1	57.5	12q15	45.5	5q13.3	34	10p12.2	41.5
1p31.3	59.25	12p11.22	51.5	9p22.3	36.25	3q13.13	42
21q21.1	60	15q22.1	51.5	8p21.3	41.25	13q13.2	42.75
3q26.1	73.25	15q23	51.75	3p25.1	42.5	16q12.1	47
12p11.22	73.25	8q21.13	54	10q23.2	42.75	6p21.31	50
6q26	74.5	9p21.2	54.5	5q32	47	11q22.2	53
13q32.1	75.75	21q21.1	55.25	1p36.11	47.5	10q26.3	53.5
17p13.2	78	9q34.13	59	Xp22.22	48.75	9q33.1	55.5
11q14.1	80.25	9p24.2	62	21q21.1	49	4q21.1	56

only the strength of the links is affected. With binary covariates, which is the case in our motivating application, this situation does not arise, but in principle this could be seen as a limitation. On the other hand, this is a necessary consequence of continuity, and small changes in the covariates resulting in large changes in the graph, as can happen with the approach of Liu et al. (2010), make the model interpretation difficult. Further, our approach has the additional advantage of discovering exactly which covariates affect which edges, which can be more important in terms of scientific insight.

While here we focused on binary network data, the idea can be easily extended to categorical and Gaussian data, and to mixed graphical models involving both discrete and continuous data. Another direction of interest is understanding conditions under which methods based on the neighborhood selection principle of running separate regressions are preferable to pseudo-likelihood type methods, and vice versa. This comparison arises frequently in the literature, and understanding this general principle would have applications far beyond our particular method.

## Acknowledgements

E. Levina's research is partially supported by NSF grants DMS-1106772 and DMS-1159005 and NIH grant 5-R01-AR-056646-03. J. Zhu's research is partially supported by NSF grant DMS-0748389 and NIH grant R01GM096194.

## References

- ADELAIDE, J., CHAFFANET, M., IMBERT, A., ALLIONE, F., GENEIX, J., POPOVICI, C., VAN ALEWIJK, D., TRAPMAN, J., ZEILLINGER, R., BØRRENSEN-DALE, A., LIDEREAU, R., BIRNBAUM, D. & PE'BUSQUE, M. (1998). Chromosome region 8p11-p21: Refined mapping and molecular alterations in breast cancer. *Genes, Chromosomes & Cancer* 22 186–199.
- ALI, I., LIDEREAU, R., THEILLET, C. & CALLAHAN, R. (1987). Reduction to homozygosity of genes on chromosome 11 in human breast neoplasia. *Science* 238 185–188.
- BARABASI, A. L. & ALBERT, R. (1999). Emergence of scaling in random networks. *Science* 509–512.
- BERGAMASCHI, A., KIM, Y., WANG, P., SØRLIE, T., HERNANDEZ-BOUSSARD, T., LONNING, P., TIBSHIRANI, R., BØRRESEN-DALE, A. & POLLACK, J. (2006). Distinct patterns of dna copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. *Genes, Chromosomes & Cancer* 45 1033–1040.
- CAI, T., LI, H., LIU, W. & XIE, J. (2011a). Covariate adjusted precision matrix estimation with an application in genetical genomics. *Biometrika* 1–19.

- CAI, T., LIU, W. & LUO, X. (2011b). A constrained l1 minimization approach to sparse precision matrix estimation. *J. American Statistical Association* 106 594–607.
- CONTE, N., CHARAFE-JAUFFRET, E., ADELAIDE, J., GINESTIER, C., GENEIX, J., ISNARDON, D., JACQUEMIER, J. & BIRNBAUM, D. (2002). Carcinogenesis and translational controls: Tacc1 is down-regulated in human cancers and associates with mrna regulators. *Oncogene* 21 5619–5630.
- D’ASPREMONT, A., BANERJEE, O. & EL GHAOU, L. (2008). First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and its Applications* 30 56–66.
- DEVILCE, P., VAN VLIET, M., VAN SLOUN, P., DIJKSHOORN, N., HERMANS, J., PEARSON, P. & CORNELISSE, C. (1991). Allelotype of human breast carcinoma: a second major site for loss of heterozygosity is on chromosome 6q. *Oncogene* 6 1705–1711.
- FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9 432–441.
- FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. (2010). Regularized paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33.
- FU, W. (1998). Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics* 7 397–416.
- GARCIA, I., DIETRICH, P., AAPRO, M., VAUTHIER, G., VADAS, L. & ENGEL, E. (1989). Genetic alterations of c-myc, c-erbB-2, and c-ha-ras protooncogenes and clinical associations in human breast carcinomas. *Cancer Research* 49 6675–6679.
- GEISLER, S., BØRRESEN-DALE, A., JOHNSEN, H., AAS, T., GEISLER, J., AKSLEN, L., ANKER, G. & LONNING, P. (2003). Tp53 gene mutations predict the response to neoad-

- juvant treatment with 5-fluorouracil and mitomycin in locally advanced breast cancer. *Clinical Cancer Research* 9 5582–5588.
- GUO, J., LEVINA, E., MICHAILIDIS, G. & ZHU, J. (2010a). Estimating heterogeneous graphical models for discrete data with an application to roll call voting Manuscript.
- GUO, J., LEVINA, E., MICHAILIDIS, G. & ZHU, J. (2010b). Joint estimation of multiple graphical models. *Biometrika* 98 1–15.
- GUO, J., LEVINA, E., MICHAILIDIS, G. & ZHU, J. (2010c). Joint structure estimation for categorical Markov networks .
- HASSNER, M. & SKLANSKY, J. (1980). The use of markov random fields as models of texture. *Computer Graphics Image Processing* 12 357–370.
- HÖFLING, H. & TIBSHIRANI, R. (2009). Estimation of sparse binary pairwise markov networks using pseudo-likelihoods. *Journal of Machine Learning Research* 10 883–906.
- HUANG, T., YEH, P., MARTIN, M., STRAUB, R., GILLIAM, T., CALDWELL, C. & SKIBBA, J. (1995). Genetic alterations of microsatellites on chromosome 18 in human breast carcinoma. *Diagnostic Molecular Pathology* 4 66–72.
- ISING, E. (1925). Beitrag zur theorie der ferromagnetismus. *Zeitschrift für Physik* 31 253–258.
- LAM, C. & FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrices estimation. *Annals of Statistics* 37 4254–4278.
- LASSUS, H., SALOVAARA, R., ALTONEN, L. & BUTZOW, R. (2001). Allelic analysis of serous ovarian carcinoma reveals two putative tumor suppressor loci at 18q22-q23 distal to smad4, smad2, and dcc. *American Journal of Pathology* 159 35–42.



- LIU, H., CHEN, X., LAFFERTY, J. & WASSERMAN, L. (2010). Graph-valued regression. *Proceedings of Advances in Neural Information Processing Systems (NIPS) 23*.
- MACKEY, J., ELDER, P., PORTEOUS, D., STEEL, C., HAWKINS, R., GOING, J. & CHETTY, U. (1988). Partial deletion of chromosome 11p in breast cancer correlates with size of primary tumour and oestrogen receptor level. *British Journal of Cancer* 6 710–714.
- MANNING, C. & SCHUTZE, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT press.
- MAZUMDER, R. & HASTIE, T. (2012). Exact covariance thresholding into connected components for large-scale graphical lasso. *JMLR* To appear.
- MEINSHAUSEN, N. & BÜHLMANN, P. (2006). High dimensional graphs and variable selection with the Lasso. *Ann. Statist.* 34 1436–1462.
- MEINSHAUSEN, N. & BÜHLMANN, P. (2010). Stability selection. *Journal of the Royal Statistical Society* 72 417–473.
- MINELLA, A., GRIM, J., WELCKER, M. & CLURMAN, B. (2007). p53 and scfbw7 cooperatively restrain cyclin e-associated genome instability. *Oncogene* 26 6948–6953.
- MITELMAN, F., MERTERNS, F. & JOHANSSON, B. (1997). A breakpoint map of recurrent chromosomal rearrangements in human neoplasia. *Nature Genetics* 15 417–474.
- NEGRINI, M., SABBIONI, S., POSSATI, L., RATTAN, S., CORALLINI, A., BARBANTI-BRODANO, G. & CROCE, C. (1994). Suppression of tumorigenicity of breast cancer cells by microcell-mediated chromosome transfer: Studies on chromosomes 6 and 11. *Cancer Research* 54 1331–1336.
- PENG, J., WANG, P., ZHOU, N. & ZHU, J. (2009). Partial correlation estimation by joint sparse regression model. *Journal of the American Statistics Association* 104 735–746.

- RAVIKUMAR, P., WAINWRIGHT, M. & LAFFERTY, J. (2010). High-dimensional ising model selection using l1-regularized logistic regression. *Annals of Statistics* 38 1287–1319.
- RAVIKUMAR, P., WAINWRIGHT, M., RASKUTTI, G. & YU, B. (2008). Model selection in gaussian graphical models: High-dimensional consistency of l1-regularized mle. *Advances in Neural Information Processing Systems(NIPS)* 21.
- ROCHA, G. V., ZHAO, P. & YU, B. (2008). A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (splice). Tech. Rep. 759, Department of Statistics, UC Berkeley.
- ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. & ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* 2 494–515.
- SINHA, S., SINGH, R., ALAM, N., ROY, A., ROUCHOUDHURY, S. & PANDA, C. (2008). Alterations in candidate genes PHF2, FANCC, PTCH1 and XPA at chromosomal 9q22.3 region: Pathological significance in early- and late- onset breast carcinoma. *Molecular Cancer* .
- WANG, P., CHAO, D. & HSU, L. (2011). Learning networks from high dimensional binary data: An application to genomic instability data. *Biometrics* 67 164–173.
- WITTEN, D. M., FRIEDMAN, J. H. & SIMON, N. (2011). New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics* 20 892–900.
- WOODS, J. (1978). Markov image modeling. *IEEE Transactions on Automatic Control* 23 846–850.
- YANG, Z., STREICHER, K., RAY, M., ABRAMS, J. & ETHEIR, S. (2006). Multiple inter-

acting oncogenes on the 8p11-p12 amplicon in human breast cancer. *Cancer Research* 66 11632–11634.

YIN, J. & LI, H. (2011). A sparse conditional gaussian graphical model for analysis of genetical genomics data. *Annals of Applied Statistics* 5 2630–2650.

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

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

ZHAO, H., LANGEROD, A., JI, Y., NOWELS, K., NESSLAND, J., TIBSHIRANI, I., R. BUKHOLM, KARESEN, R., BOTSTEIN, D. & BØRRESEN-DALE, A. (2004). Different gene expression patterns in invasive lobular and ductal carcinomas of the breast. *Molecular Biology of the Cell* 15 2523–2536.

## Appendix: Proof of Theorem 1

For notational convenience, we omit the  $j$  indexing each separate regression. Following the literature, we prove the main theorem in two steps: first, we prove the result holds when assumptions **A1** and **A2** hold for  $\mathbf{I}^n$  and  $\mathbf{U}^n$ , the sample versions of  $\mathbf{I}^*$  and  $\mathbf{U}^*$  defined in (7) (Proposition 1). Then we show that if **A1** and **A2** hold for the population versions  $\mathbf{I}^*$  and  $\mathbf{U}^*$ , they also hold for  $\mathbf{I}^n$  and  $\mathbf{U}^n$  with high probability (Proposition 2). The sample quantities  $\mathbf{I}^n$  and  $\mathbf{U}^n$  are defined as

$$\begin{aligned} \mathbf{I}^n &= \nabla^2 \ell(\boldsymbol{\theta}^*, \mathcal{D}_n) = \frac{1}{n} \sum_{i=1}^n (p_j^i (1 - p_j^i) (\mathbf{x}^i \otimes \mathbf{y}_{\setminus j}^i) (\mathbf{x}^i \otimes \mathbf{y}_{\setminus j}^i)^T) , \\ \mathbf{U}^n &= \frac{1}{n} \sum_{i=1}^n (\mathbf{x}^i \otimes \mathbf{y}_{\setminus j}^i) (\mathbf{x}^i \otimes \mathbf{y}_{\setminus j}^i)^T . \end{aligned}$$

**Proposition 1** *If **A1** and **A2** are satisfied by  $\mathbf{I}^n$  and  $\mathbf{U}^n$ , assume moreover that*

$$\begin{aligned} M_n &= \sup \|\mathbf{x}\|_\infty < \infty \quad a.s., \\ \lambda_n &\geq \frac{8M_n(2-\alpha)}{\alpha} \sqrt{\frac{\log p + \log q}{n}}, \\ n &> Cd^2(\log p + \log q). \end{aligned}$$

*Then with probability at least  $1 - 2 \exp\left(-C \frac{\lambda_n^2 n}{M_n^2}\right)$ , the result of Theorem 1 holds.*

*Proof of Proposition 1.* The proof requires several steps. The uniqueness part follows directly from the following lemma:

**Lemma 1** (Shared sparsity and uniqueness of  $\hat{\boldsymbol{\theta}}$ , Ravikumar et al. (2010)). *Define the sign vector  $\mathbf{t}$  for  $\boldsymbol{\theta}$  to satisfy the following properties,*

$$\begin{cases} \hat{t}_k = \text{sign}(\hat{\theta}_k), & \text{if } \hat{\theta}_k \neq 0, \\ |\hat{t}_k| \leq 1, & \text{if } \hat{\theta}_k = 0. \end{cases}$$

*Suppose there exists an optimal solution  $\hat{\boldsymbol{\theta}}$  with  $\text{sign } \hat{\mathbf{t}}$  defined as above, such that,  $\|\hat{\mathbf{t}}_{SC}\|_\infty < 1$ , then any optimal solution  $\tilde{\boldsymbol{\theta}}$  must have  $\tilde{\boldsymbol{\theta}}_{SC} = 0$ . Furthermore, if the Hessian matrix  $\nabla^2 \ell(\hat{\boldsymbol{\theta}})_{SS}$  is strictly positive definite, then  $\hat{\boldsymbol{\theta}}$  is the unique solution.*

We now proceed to prove the rest of Proposition 1. For  $\hat{\boldsymbol{\theta}}$  to be a solution of (10), the sub-gradient at  $\hat{\boldsymbol{\theta}}$  must be 0, i.e.,

$$\nabla \ell(\hat{\boldsymbol{\theta}}, \mathcal{D}_n) + \lambda_n \hat{\mathbf{t}} = 0. \quad (15)$$

Then we can write  $\nabla \ell(\hat{\boldsymbol{\theta}}, \mathcal{D}_n) - \nabla \ell(\boldsymbol{\theta}^*, \mathcal{D}_n) = -\lambda_n \hat{\mathbf{t}} + W^n$ , where

$$W^n = -\nabla \ell(\boldsymbol{\theta}^*, \mathcal{D}_n) = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}^i \otimes \mathbf{y}_{\setminus j}^i) (y_j^i - p_j^i(\boldsymbol{\theta}^*)).$$

Let  $\tilde{\boldsymbol{\theta}}$  denote a point in the line segment connecting  $\hat{\boldsymbol{\theta}}$  and  $\boldsymbol{\theta}^*$ . Applying the mean value theorem gives

$$\mathbf{I}^n (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = W^n - \lambda_n \hat{\mathbf{t}} + R^n. \quad (16)$$

where  $R^n = \left( \nabla^2 \ell(\boldsymbol{\theta}^*, \mathcal{D}_n) - \nabla^2 \ell(\tilde{\boldsymbol{\theta}}, \mathcal{D}_n) \right) (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ .

Now define  $\hat{\boldsymbol{\theta}}$  as follows: let  $\mathcal{S}$  be the index set of true non-zeros in  $\boldsymbol{\theta}^*$ , let  $\hat{\boldsymbol{\theta}}_{\mathcal{S}}$  be the solution of

$$\min_{(\hat{\boldsymbol{\theta}}_{\mathcal{S}}, 0)} \ell(\hat{\boldsymbol{\theta}}, \mathcal{D}_n) + \lambda_n \|\hat{\boldsymbol{\theta}}_{\mathcal{S}}\|_1, \quad (17)$$

and let  $\hat{\boldsymbol{\theta}}_{\mathcal{S}^c} = 0$ . We will show that this  $\hat{\boldsymbol{\theta}}$  is the optimal solution and is sign consistent with high probability.

We set the corresponding sign vector  $\hat{\mathbf{t}}_{\mathcal{S}}$  for  $\hat{\boldsymbol{\theta}}_{\mathcal{S}}$  similarly defined as in Lemma 1, and  $\hat{\mathbf{t}}_{\mathcal{S}^c} = -\frac{1}{\lambda_n} \nabla_{\mathcal{S}^c} \ell(\hat{\boldsymbol{\theta}}_{\mathcal{S}}, \mathcal{D}_n)$  as obtained in (15). Now we need to show that with high probability,

$$\|\hat{\mathbf{t}}_j\|_{\infty} < 1, \quad \text{for } j \in \mathcal{S}^c \quad (18)$$

$$\hat{\mathbf{t}}_j = \text{sign}(\boldsymbol{\theta}_j^*), \quad \text{for } j \in \mathcal{S} \text{ and } \|\boldsymbol{\theta}_j^*\| \geq \frac{10\lambda_n\sqrt{d}}{\Delta_{\min}} \quad (19)$$

The following three lemmas form the proof.

**Lemma 2** (Control the remainder term  $W^n$ ). *For  $\alpha \in (0, 1]$ , assume  $\|\mathbf{x}\|_{\infty} \leq M_n$  a.s, then,*

$$P\left(\frac{2-\alpha}{\lambda_n} \|W^n\|_{\infty} \geq \frac{\alpha}{4}\right) \leq 4 \exp\left(-\frac{\lambda_n^2 n \alpha^2}{32M_n^2(2-\alpha)^2} + \log p + \log q\right).$$

*This probability goes to 0 as long as  $\lambda_n \geq 8M_n^{\frac{2-\alpha}{\alpha}} \sqrt{\frac{\log p + \log q}{n}}$ .*

*Proof of Lemma 2.* We can write  $W^n = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}^i \otimes \mathbf{y}_{\setminus j}^i)(y_j^i - p_j^i(\boldsymbol{\theta}^*)) = \sum_{i=1}^n Z_i$ , where  $Z_{ik}$  is bounded by  $M_n/n$ . Thus by Azuma-Hoeffding Inequality,

$$\begin{aligned} P\left(\|W^n\|_{\infty} \geq \frac{\lambda_n \alpha}{4(2-\alpha)}\right) &\leq 2pqP\left(\|W_k^n\|_{\infty} \geq \frac{\lambda_n \alpha}{4(2-\alpha)}\right) \\ &\leq 4 \exp\left(-\frac{\lambda_n^2 n \alpha^2}{32M_n^2(2-\alpha)^2} + \log p + \log q\right). \end{aligned}$$

□

**Lemma 3** ( $\ell_2$ -consistency of the sub-vector  $\hat{\boldsymbol{\theta}}_{\mathcal{S}}$ ). *If  $\lambda_n d < \frac{\Delta_{\min}^2}{10\Delta_{\max} M_n}$ , and,  $\|W^n\|_{\infty} \leq \frac{\lambda_n}{4}$ ,*

*then*

$$\|\hat{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_{\mathcal{S}}^*\|_2 \leq \frac{5\lambda_n\sqrt{d}}{\Delta_{\min}}.$$

*Proof of Lemma 3.* Let  $G(u_S) = \ell(\boldsymbol{\theta}_S^* + u_S, \mathcal{D}_n) - \ell(\boldsymbol{\theta}_S^*, \mathcal{D}_n) + \lambda_n(\|\boldsymbol{\theta}_S^* + u_S\|_1 - \|\boldsymbol{\theta}_S^*\|_1)$  be a function  $G : \mathbb{R}^d \rightarrow \mathbb{R}$ . It is easy to see that  $G(u_S)$  is convex and it achieves its minimum at  $\hat{u}_S = \hat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_S^*$ . Moreover,  $G(0) = 0$ . Thus if we can show that  $G(u_S)$  is positive on the set  $\|u_S\|_2 = B$ , then we will have  $\hat{u}_S \leq B$  due to convexity of  $G(u_S)$ . Note that

$$G(u_S) = -W_S^{nT} u_S + u_S^T \nabla^2 \ell(\boldsymbol{\theta}_S^* + \alpha u_S) u_S + \lambda_n(\|\boldsymbol{\theta}_S^* + u_S\|_1 - \|\boldsymbol{\theta}_S^*\|_1) .$$

Further,

$$\begin{aligned} |W_S^{nT} u_S| &\leq \|W^n\|_\infty \|u_S\|_1 \leq \frac{\lambda_n}{4} \sqrt{d} \|u_S\|_2 , \\ \Lambda_{\min}(\nabla^2 \ell(\boldsymbol{\theta}_S^* + \alpha u_S)) &\geq \Delta_{\min} - \Delta_{\max} M_n \sqrt{d} \|u_S\|_2 , \\ |\lambda_n(\|\boldsymbol{\theta}_S^* + u_S\|_1 - \|\boldsymbol{\theta}_S^*\|_1)| &\leq \lambda_n \sqrt{d} \|u_S\|_2 . \end{aligned}$$

Combining all of the above, we have

$$G(u_S) \geq \|u_S\|_2 (-\Delta_{\max} M_n \sqrt{d} \|u_S\|_2^2 + \Delta_{\min} \|u_S\|_2 - \frac{5}{4} \lambda_n \sqrt{d}) .$$

Easy algebra shows that if  $\lambda_n d \leq \frac{\Delta_{\min}^2}{10 \Delta_{\max} M_n}$  and  $B = \frac{5 \lambda_n \sqrt{d}}{\Delta_{\min}}$ , the result follows.  $\square$

**Lemma 4** (Control the remainder term  $R^n$ ). *If  $\lambda_n d \leq \frac{\Delta_{\min}^2}{100 M_n \Delta_{\max}} \frac{\alpha}{2-\alpha}$ ,  $\|W^n\|_\infty \leq \frac{\lambda_n}{4}$ , then*

$$\frac{\|R^n\|_\infty}{\lambda_n} \leq \frac{25 \Delta_{\max}}{\Delta_{\min}^2} M_n \lambda_n d \leq \frac{\alpha}{4(2-\alpha)} .$$

*Proof of Lemma 4.* Recall that

$$\begin{aligned} R^n &= \left( \nabla^2 \ell(\boldsymbol{\theta}^*, \mathcal{D}_n) - \nabla^2 \ell(\tilde{\boldsymbol{\theta}}, \mathcal{D}_n) \right) (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \\ &= \frac{1}{n} \sum_{i=1}^n \left( p_j^i(\boldsymbol{\theta}^*) (1 - p_j^i(\boldsymbol{\theta}^*)) - p_j^i(\tilde{\boldsymbol{\theta}}) (1 - p_j^i(\tilde{\boldsymbol{\theta}})) \right) (\mathbf{x}^i \otimes \mathbf{y}_{\setminus j}^i) (\mathbf{x}^i \otimes \mathbf{y}_{\setminus j}^i)^T (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) . \end{aligned}$$

Let  $\omega_j^i(\boldsymbol{\theta}) = p_j^i(\boldsymbol{\theta})(1 - p_j^i(\boldsymbol{\theta}))$ . The  $k$ -th element of  $R^n$  has the form

$$\begin{aligned} R_k^n &= \frac{1}{n} \sum_{i=1}^n (\omega_j^i(\boldsymbol{\theta}^*) - \omega_j^i(\tilde{\boldsymbol{\theta}})) Z_k^i (\mathbf{x}^i \otimes \mathbf{y}_{\setminus j}^i)^T (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \\ &= \frac{1}{n} \sum_{i=1}^n \dot{\omega}_j^i(\tilde{\boldsymbol{\theta}}) Z_k^i (\boldsymbol{\theta}^* - \tilde{\boldsymbol{\theta}})^T (\mathbf{x}^i \otimes \mathbf{y}_{\setminus j}^i) (\mathbf{x}^i \otimes \mathbf{y}_{\setminus j}^i)^T (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) , \end{aligned}$$

where  $Z_k^i = x_l^i y_m^i$ , for some  $(l, m)$ . By **A1** and Lemma 3, we have

$$|R_k^n| \leq M_n \Delta_{\max} \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2 \leq M_n \Delta_{\max} \left( \frac{5\lambda_n \sqrt{d}}{\Delta_{\min}} \right)^2.$$

□

Putting all the lemmas together, we are ready to prove Proposition 1.

*Proof of Proposition 1.* Set  $\lambda_n = \frac{8M_n(2-\alpha)}{\alpha} \sqrt{\frac{\log p + \log q}{n}}$ . By Lemma 2, we have  $\|W^n\|_\infty \leq \frac{\lambda_n \alpha}{4(2-\alpha)} \leq \frac{\lambda_n}{4}$  with probability at least  $1 - 4 \exp(-C\lambda_n^2 n / M_n^2)$ . Choosing  $n \geq \frac{100^2 \Delta_{\max}^2 (2-\alpha)^2}{\Delta_{\min}^4 \alpha^2} d^2 (\log p + \log q)$ , we have  $\lambda_n d \leq \frac{\Delta_{\min}^2}{100M_n \Delta_{\max}} \frac{\alpha}{2-\alpha}$ , thus the conditions of Lemmas 3 and 4 hold.

By rewriting (16) and utilizing the fact that  $\hat{\boldsymbol{\theta}}_{SC} = \boldsymbol{\theta}_{SC}^* = 0$ , we have

$$\mathbf{I}_{SCS}^n (\hat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_S^*) = W_{SC}^n - \lambda_n \hat{\mathbf{t}}_{SC} + R_{SC}^n, \quad (20)$$

$$\mathbf{I}_{SS}^n (\hat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_S^*) = W_S^n - \lambda_n \hat{\mathbf{t}}_S + R_S^n. \quad (21)$$

Since  $\mathbf{I}_{SS}^n$  is invertible by assumption, combining (20) and (21) gives

$$\mathbf{I}_{SCS}^n (\mathbf{I}_{SS}^n)^{-1} (W_S^n - \lambda_n \hat{\mathbf{t}}_S + R_S^n) = W_{SC}^n - \lambda_n \hat{\mathbf{t}}_{SC} + R_{SC}^n. \quad (22)$$

To show (18), we reorganize (22) and use results from Lemmas 2 and 4:

$$\begin{aligned} \lambda_n \|\hat{\mathbf{t}}_{SC}\|_\infty &= \|\mathbf{I}_{SCS}^n (\mathbf{I}_{SS}^n)^{-1} (W_S^n - \lambda_n \hat{\mathbf{t}}_S + R_S^n) - W_{SC}^n - R_{SC}^n\|_\infty \\ &\leq \|\mathbf{I}_{SCS}^n (\mathbf{I}_{SS}^n)^{-1}\|_\infty (\|W^n\|_\infty + \lambda_n + \|R^n\|_\infty) + \|W^n\|_\infty + \|R^n\|_\infty \\ &\leq \lambda_n \left(1 - \frac{\alpha}{2}\right). \end{aligned}$$

To show (19), it suffices to show that  $\|\hat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_S^*\|_\infty \leq \frac{\boldsymbol{\theta}_{\min}^*}{2}$ . By Lemma 3,

$$\|\hat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_S^*\|_\infty \leq \frac{5\lambda_n \sqrt{d}}{\Delta_{\min}} \leq \frac{\boldsymbol{\theta}_{\min}^*}{2}.$$

The last inequality follows as long as  $\boldsymbol{\theta}_{\min}^* \geq \frac{10\lambda_n \sqrt{d}}{\Delta_{\min}}$ . This completes the proof of Proposition

1. □

**Proposition 2** *If  $\mathbf{I}^*$  and  $\mathbf{U}^*$  satisfy **A1** and **A2**, and  $M_n = \sup\|\mathbf{x}\|_\infty < \infty$  a.s., the following hold for any  $\delta > 0$ .  $A$  and  $B$  are some positive constants.*

$$P \left\{ \Lambda_{\max} \left( \frac{1}{n} \sum_{i=1}^n (\mathbf{x}^i \otimes \mathbf{y}_{\setminus j}^i)(\mathbf{x}^i \otimes \mathbf{y}_{\setminus j}^i)^T \right) \geq D_{\max} + \delta \right\} \leq 2 \exp \left( -A \frac{\delta^2 n}{M_n^2 d^2} + B(\log p + \log q) \right)$$

$$P(\Lambda_{\min}(\mathbf{I}_{\mathcal{S}\mathcal{S}}^n) \leq C_{\min} - \delta) \leq 2 \exp \left( -A \frac{\delta^2 n}{M_n^2 d^2} + B \log d \right)$$

$$P \left( \|\mathbf{I}_{\mathcal{S}^c \mathcal{S}}^n (\mathbf{I}_{\mathcal{S}\mathcal{S}}^n)^{-1}\|_\infty \geq 1 - \frac{\alpha}{2} \right) \leq \exp \left( -A \frac{n}{M_n^2 d^3} + B(\log p + \log q) \right)$$

We omit the proof of Proposition 2, which is very similar to Lemmas 5 and 6 in Ravikumar et al. (2010).

*Proof of Theorem 1.* With Propositions 1 and 2, the proof of Theorem 1 is straightforward. Given that A1 and A2 are satisfied by  $\mathbf{I}^*$  and  $\mathbf{U}^*$  and that conditions (13) and (14) hold, on the set  $\mathcal{A} = \{\mathbf{x} : M_n = \sup\|\mathbf{x}\|_\infty < \infty\}$  the assumptions in Proposition 2 are satisfied. Thus with probability at least  $1 - \exp(-\frac{C\lambda_n^2 n}{M_n^2})$ , the conditions of Proposition 1 hold, and therefore the results in Theorem 1 hold. Finally, let  $\mathcal{T}$  stand for the set where the results of Theorem 1 hold. Then by (11) and (12), we have

$$P(\mathcal{T}^c) \leq P(\mathcal{T}^c | \mathcal{A}) + P(\mathcal{A}^c) \leq \exp(-\frac{C\lambda_n^2 n}{M_n^2}) + \exp(-M_n^\delta) \leq \exp-(C'\lambda_n^2 n)^{\delta^*}, \text{ where } 0 < \delta^* < 1.$$

□