A Sparse Ising Model with Covariates

Jie Cheng,[1] Elizaveta Levina,[1] Pei Wang,[2] and Ji Zhu[1],*

[1]Department of Statistics, University of Michigan, Ann Arbor, Michigan, U.S.A.
[2]Icahn School of Medicine at Mount Sinai, New York, U.S.A.
*_email:_ jizhu@umich.edu

SUMMARY.   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 ℓ_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: Binary Markov network; Graphical model; Ising model; Lasso; Pseudo-likelihood; Tumor suppressor genes.

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, Chao, and Hsu, 2011), image processing (Woods, 1978; Hassner and Sklansky, 1980), and natural language processing (Manning and 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, that is, 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 articles on estimating the inverse covariance matrix have appeared in recent years, with a focus on the high-dimensional framework, for example, Meinshausen and Bühlmann (2006), Yuan and Lin (2007), d'Aspremont, Banerjee, and El Ghaoui (2008), Ravikumar et al. (2008), Rocha, Zhao, and Yu et al. (2008), Rothman et al. (2008), Lam and Fan (2009), Peng et al. (2009), Yuan (2010), Cai, Liu, and Luo (2011). Most of these articles 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, Hastie, and Tibshirani (2008), which was recently improved further by Witten, Friedman, and Simon (2011) and Mazumder and 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, Wainwright, and Lafferty (2010) proposed an approach in the spirit of Meinshausen and Bühlmann (2006)'s work for the Gaussian case, fitting separate ℓ_1-penalized logistic regressions for each node to infer the graph structure. A pseudo-likelihood based algorithm was developed by Höefling and Tibshirani (2009) and analyzed by Guo et al. (2010).

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., 2010, 2014). 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 articles on graphical models that incorporate covariates, but they do so in ways quite different from ours. Yin and Li (2011) and Cai et al. (2013) 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 article 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 $\boldsymbol{y} = (y_1, \ldots, y_q) \in \{0, 1\}^q$ denote a binary random vector. The Ising model specifies the probability mass function $P_{\boldsymbol{\theta}}(\boldsymbol{y})$ as

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

where $\boldsymbol{\theta} = (\theta_{11}, \theta_{12}, \ldots, \theta_{q-1q}, \theta_{qq})$ 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 θ_{jk} equals to θ_{kj} unless otherwise specified. The Markov property is related to the parameter $\boldsymbol{\theta}$ via

$$\theta_{jk} = 0 \Longleftrightarrow y_j \perp y_k \parallel \boldsymbol{y}_{\backslash (j,k)}, \quad \forall j \neq k, \qquad (1)$$

that is, 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 $\mathcal{D}_n = \{(\boldsymbol{x}^1, \boldsymbol{y}^1), \ldots, (\boldsymbol{x}^n, \boldsymbol{y}^n)\}$ with $\boldsymbol{x}^i \in \mathbb{R}^p$ and $\boldsymbol{y}^i \in \{0, 1\}^q$. We assume that given covariates $\boldsymbol{x}$, the binary response $\boldsymbol{y}$ follows the Ising distribution given by

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

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

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

where, $\boldsymbol{y}_{\backslash j} = (y_1, \ldots, y_{j-1}, y_{j+1}, \ldots, y_q)$. Further, conditioning on $\boldsymbol{y}_{\backslash \{j,k\}}$ being 0, we have

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

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

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

$$\boldsymbol{\theta}_{jk}(\boldsymbol{x}) = \theta_{jk0} + \boldsymbol{\theta}_{jk}^T \boldsymbol{x}, \quad \text{where} \quad \boldsymbol{\theta}_{jk}^T = (\theta_{jk1}, \ldots, \theta_{jkp})$$

$$\boldsymbol{\theta}_{jk}(\boldsymbol{x}) = \boldsymbol{\theta}_{kj}(\boldsymbol{x}), \qquad \forall j > k.$$

The model can be expressed in terms of the parameter vector $\boldsymbol{\theta}$ as follows:

$$P_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})$$

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

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

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

Note that for each j, the conditional log-odds in (5) involves $(p+1)q$ parameters; taking into account the symmetry, that is, $\theta_{jk0} = \theta_{kj0}$ and $\boldsymbol{\theta}_{jk} = \boldsymbol{\theta}_{kj}$, we thus have a total of $(p+1)q(q+1)/2$ parameters in the fully parametrized model.

The choice of linear parametrization for $\boldsymbol{\theta}_{jk}(\boldsymbol{x})$ has several advantages. First (5) mirrors the logistic regression model when viewing the x_ℓ'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 effect of that particular predictor. Specifically, for continuous covariates, $\theta_{jk\ell}$ ($\ell \geq 1$) describes the effect of covariate x_ℓ on the conditional log-odds of y_j when y_k is 1, and θ_{jk0} simply describes the main effect of y_k on the conditional log-odds of y_j. For categorical covariates, the interpretation is almost the same except that now the covariate would be represented by several dummy variables, and the parameter $\theta_{jk\ell}$ represents the effect of each level of the covariate relative to the reference level. 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 $\boldsymbol{x}$ and the rest of y_ℓ's, and $\theta_{jk\ell}$ being zero implies that the conditional association between y_j and y_k does not depend on x_ℓ. 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 $P_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})$ in (4) includes the partition function $Z(\boldsymbol{\theta}(\boldsymbol{x}))$, which requires summation of 2^q terms for each data point and makes it intractable to directly maximize the joint conditional likelihood $\sum_{i=1}^n \log P_{\boldsymbol{\theta}}(\boldsymbol{y}^i|\boldsymbol{x}^i)$. 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 $\boldsymbol{y}_{\backslash j}^i$ and $\boldsymbol{x}^i$ rather than the joint log-likelihood of $\boldsymbol{y}^i$.

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

$$\ell_j(\boldsymbol{\theta}; \mathcal{D}_n) = -\frac{1}{n}\sum_{i=1}^n \log P(y_j^i|\boldsymbol{x}^i, \boldsymbol{y}_{\backslash j}^i)$$

$$= -\frac{1}{n}\sum_{i=1}^n \left(\log(1 + e^{\eta_j^i}) - y_j^i \eta_j^i\right), \qquad (6)$$

where

$$\eta_j^i = \log\left(\frac{P(y_j^i = 1|\boldsymbol{y}_{\backslash j}^i, \boldsymbol{x}^i)}{1 - P(y_j^i = 1|\boldsymbol{y}_{\backslash j}^i, \boldsymbol{x}^i)}\right)$$

$$= \theta_{jj0} + \boldsymbol{\theta}_{jj}^T \boldsymbol{x}^i + \sum_{k \neq j}(\theta_{jk0} + \boldsymbol{\theta}_{jk}^T \boldsymbol{x}^i)y_k^i.$$

Note that this conditional log-likelihood involves the parameter vector $\boldsymbol{\theta}$ only through its subvector $\boldsymbol{\theta}_j = (\theta_{j10},$ $\boldsymbol{\theta}_{j1}^T, \ldots, \theta_{jq0}, \boldsymbol{\theta}_{jq}^T) \in \mathbb{R}^{(p+1)q}$, 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 ℓ_1 regularization to encourage sparsity, and propose two approaches to maximize the conditional likelihood (6).

2.2.1. Separate regularized logistic regressions. The first approach is to estimate each $\boldsymbol{\theta}_j$, $j = 1, \ldots, q$ separately using the following criterion,

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

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

In this approach, $\boldsymbol{\theta}_{jk}$ and $\boldsymbol{\theta}_{kj}$ are estimated from the jth and kth 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 and Bühlmann (2006), where the initial estimates are combined by comparing their magnitudes. Specifically, let $\hat{\theta}_{jk\ell}$ denote the final estimate and $\hat{\theta}_{jk\ell}^0$ denote the initial estimate from the separate regularized logistic regressions. Then for any $1 \leq j < k \leq q$ and any $\ell = 0, \ldots, p$, we can use one of the two symmetrizing approaches:

separate-max: $\hat{\theta}_{jk\ell} = \hat{\theta}_{kj\ell} = \hat{\theta}_{jk\ell}^0 \mathbb{I}_{(|\hat{\theta}_{jk\ell}^0| > |\hat{\theta}_{kj\ell}^0|)} + \hat{\theta}_{kj\ell}^0 \mathbb{I}_{(|\hat{\theta}_{jk\ell}^0| < |\hat{\theta}_{kj\ell}^0|)}$

separate-min: $\hat{\theta}_{jk\ell} = \hat{\theta}_{kj\ell} = \hat{\theta}_{jk\ell}^0 \mathbb{I}_{(|\hat{\theta}_{jk\ell}^0| < |\hat{\theta}_{kj\ell}^0|)} + \hat{\theta}_{kj\ell}^0 \mathbb{I}_{(|\hat{\theta}_{jk\ell}^0| > |\hat{\theta}_{kj\ell}^0|)}.$

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 q is relatively small, the separate-min approach is often too conservative to effectively identify non-zero parameters, while when q is relatively large, the separate-min approach performs better than the separate-max approach. More details are given in Section 4.

2.2.2. 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,

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

where $\boldsymbol{\theta}_{\backslash 0} = \boldsymbol{\theta} \backslash \{\theta_{110}, \theta_{220}, \ldots, \theta_{qq0}\}$. 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, Hastie, and Tibshirani (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. The spirit of the proof is similar to Ravikumar et al. (2010), but since their model does not include covariates x, both our assumptions and conclusions are different.

In this analysis, we treat the covariates 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

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

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

$$\boldsymbol{U}_j^* = \mathbb{E}_{\theta^*}\big((\boldsymbol{x}\otimes\boldsymbol{y}_{\setminus j})(\boldsymbol{x}\otimes\boldsymbol{y}_{\setminus j})^T\big), \tag{8}$$

where

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

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

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

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

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

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

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

$$\Lambda_{\min}\left(\boldsymbol{I}_{\mathcal{S}_j\mathcal{S}_j}^*\right) \geq \Delta_{\min}$$

$$\Lambda_{\max}(\boldsymbol{U}_j^*) \leq \Delta_{\max}$$

A3 There exists $\delta > 0$, such that

$$P(\|\boldsymbol{x}\|_\infty \geq M) \leq \exp(-M^\delta), \quad \text{for all} \quad M \geq M_0 > 0. \tag{9}$$

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

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

Assume A1, A2, and A3 hold for $\boldsymbol{I}_j^*$ and $\boldsymbol{U}_j^*$. Let $d = \max_j \|\mathcal{S}_j\|_0$ and $C > 0$ be a constant independent of (n, p, q). If

$$M_n = M \geq (C\lambda_n^2 n)^{\frac{1}{1+\delta}}, \tag{11}$$

$$\lambda_n \geq CM_n\sqrt{\frac{\log p + \log q}{n}}, \tag{12}$$

$$n \geq CM_n^2 d^3(\log p + \log q), \tag{13}$$

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

(1) *Uniqueness: $\hat{\boldsymbol{\theta}}_j$ is the unique optimal solution for any $j \in \{1, \ldots, q\}$.*
(2) *ℓ_2 consistency: $\|\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j^*\|_2 \leq 5\lambda_n\sqrt{d}/\Delta_{\min}$ for any $j \in \{1, \ldots, q\}$*
(3) *Sign consistency: $\hat{\boldsymbol{\theta}}_j$ correctly identifies all the zeros in $\boldsymbol{\theta}_j^*$ for any $j \in \{1, \ldots, q\}$; 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 A3, which requires the distribution of $\boldsymbol{x}$ to have a fast decay on large values, was not in Ravikumar et al. (2010) as the article 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 Online Supplementary Materials Appendix A.

Note that the properties in Theorem 1 hold on the original fit for each separate regression; in practice, we still need to post-process the fitted parameters by symmetrizing them using either the separate-max or the separate-min approach described earlier in Section 2.2.1. Here, we briefly investigate the three properties after this symmetrization. For property (1), the symmetrized estimates might not necessarily be the optimal solution for each regression, thus it is not sensible to consider uniqueness anymore; next, it is not difficult to see that property (3) still holds for both sets of symmetrized estimates since it characterizes the relationship between the fitted value and the true value and the true value remains the same for different fits of the same parameters; as for property (2), we will have to enlarge the upper bound on the right hand side by a factor of $d^{1/2}$, that is,

$$\|\hat{\boldsymbol{\theta}}_j^{\max} - \boldsymbol{\theta}_j^*\|_2 \leq 5\lambda_n d/\Delta_{\min},$$

$$\|\hat{\boldsymbol{\theta}}_j^{\min} - \boldsymbol{\theta}_j^*\|_2 \leq 5\lambda_n d/\Delta_{\min}, \quad \text{for any } j \in \{1, \ldots, q\}. \tag{14}$$

A simple proof of this adapted results can be found in the Online Supplementary Materials Appendix B.

4. Empirical Performance Evaluation

In this section, we present four 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 (Barabási and Albert, 1999); in the Online Supplementary Materials Appendix E, we also study the effect of sparsity for k-nearest neighbor graphs). For any (j, k), $1 \le j \le k \le q$, $(\theta_{jk0}, \boldsymbol{\theta}_{jk}^T)$ 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 θ_{jj0}, where ρ is always set to 1. Covariates $\boldsymbol{x}^i$'s are generated independently from the multivariate Gaussian distribution $N_p(0, I_p)$. Given each $\boldsymbol{x}^i$ and $\boldsymbol{\theta}$, we use Gibbs sampling to generate the $\boldsymbol{y}^i$, where we iteratively generate a sequence of y_j^i's $(j = 1, \ldots q)$ from a Bernoulli distribution with probability $P_{\boldsymbol{\theta}}(y_j^i = 1 | \boldsymbol{y}_{\backslash j}^i, \boldsymbol{x}^i)$ 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 when the dimension of the response is relatively small (results omitted for lack of space). Thus we only present results for the separate-max and the joint methods for the first three simulation studies.

4.1. Effect of Sparsity

First, we investigate how the selection performance is affected by the sparsity of the true model. The sparsity of $\boldsymbol{\theta}$ 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, ρ. 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 ρ 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.

In practice, one often needs to select a value for the tuning parameter. We compared different methods for selecting the optimal tuning parameter: validating the conditional likelihood on a separate data set of the same size, cross-validation, AIC, and BIC. The selected models corresponding to each method are marked on the ROC curves. As we can see, cross-validation seems to provide a reasonable choice in terms of balancing the true positive rate and the false positive rate.

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 β takes values in the set $\{0.5, 1, 2, 4, 8, 16\}$. For each setting, the non-zero entries of the parameter vectors $\boldsymbol{\theta}$ 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 β increases, both the separate and the joint methods show improved selection performance, but the improvement levels off eventually. The separate-max method performs better overall given that the optimal points selected by cross-validation have lower false discovery rate.

4.3. Effect of Noise Covariates

In this 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_{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_{total}, the total number of non-zero parameters remains fixed for each value of p_{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.

4.4. Stability Selection When q is Large

Stability selection was proposed by Meinshausen and Bühlmann (2010) to improve potential lack of stability of a solution and to quantify the significance of a selected structure
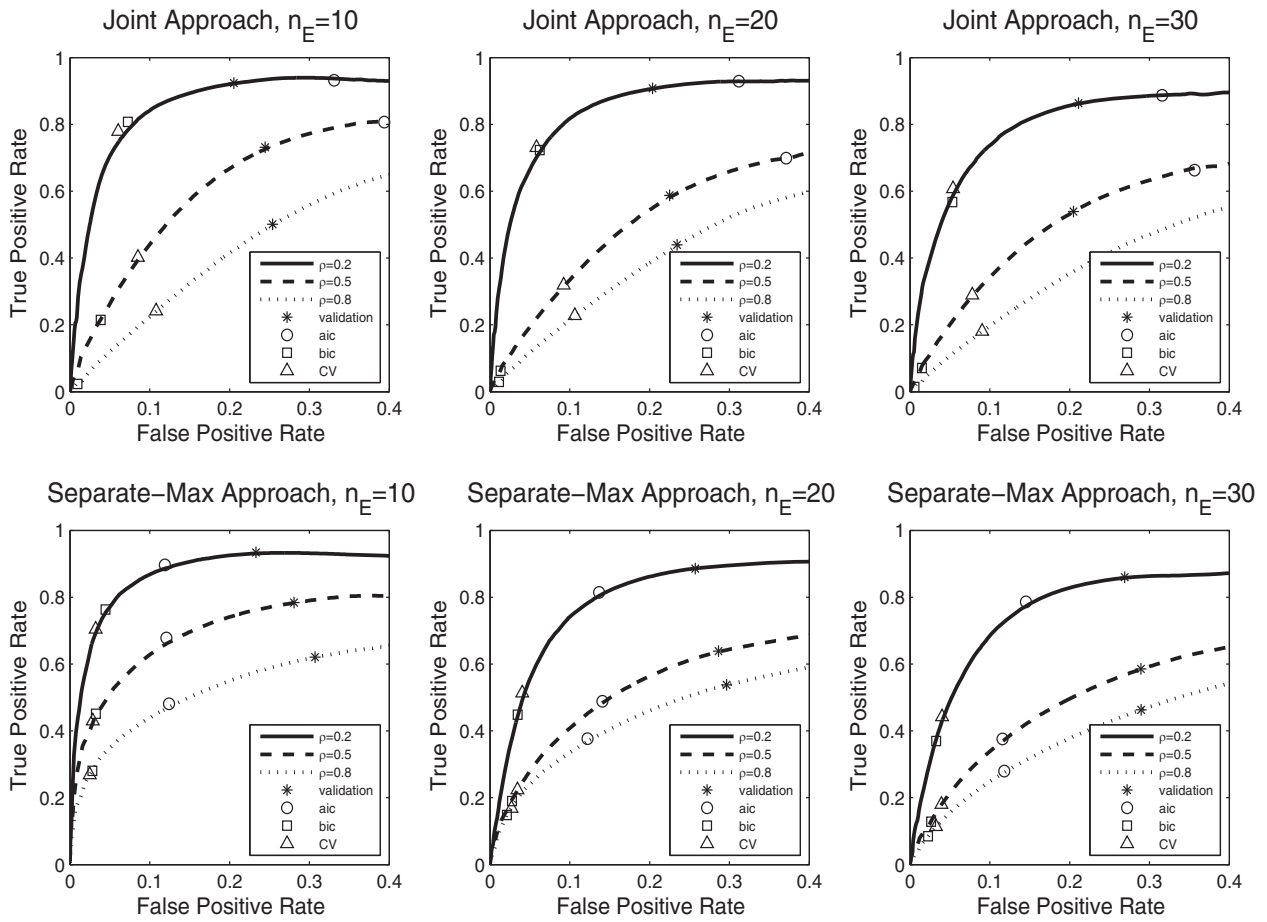
Figure 1. ROC curves for varying levels of sparsity, as measured by the number of edges (n_E) and expected proportion of non-zero covariates (ρ). Optimal points are marked for different model selection methods.
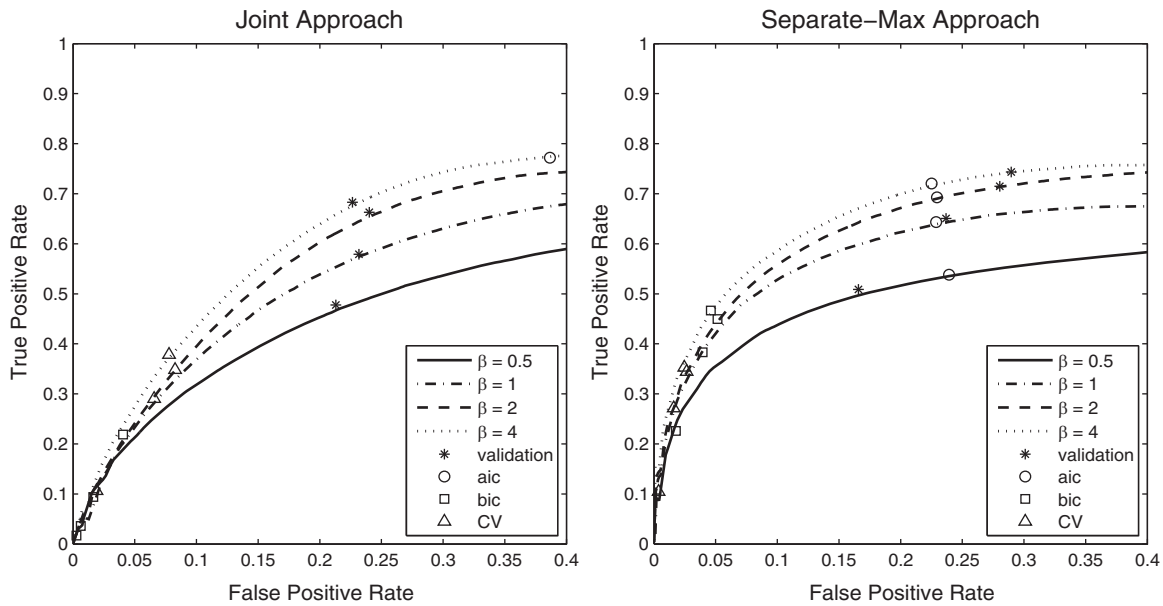

Figure 2. ROC curves for varying levels of signal strength, as measured by the parameter β. Optimal points are marked for different model selection methods.
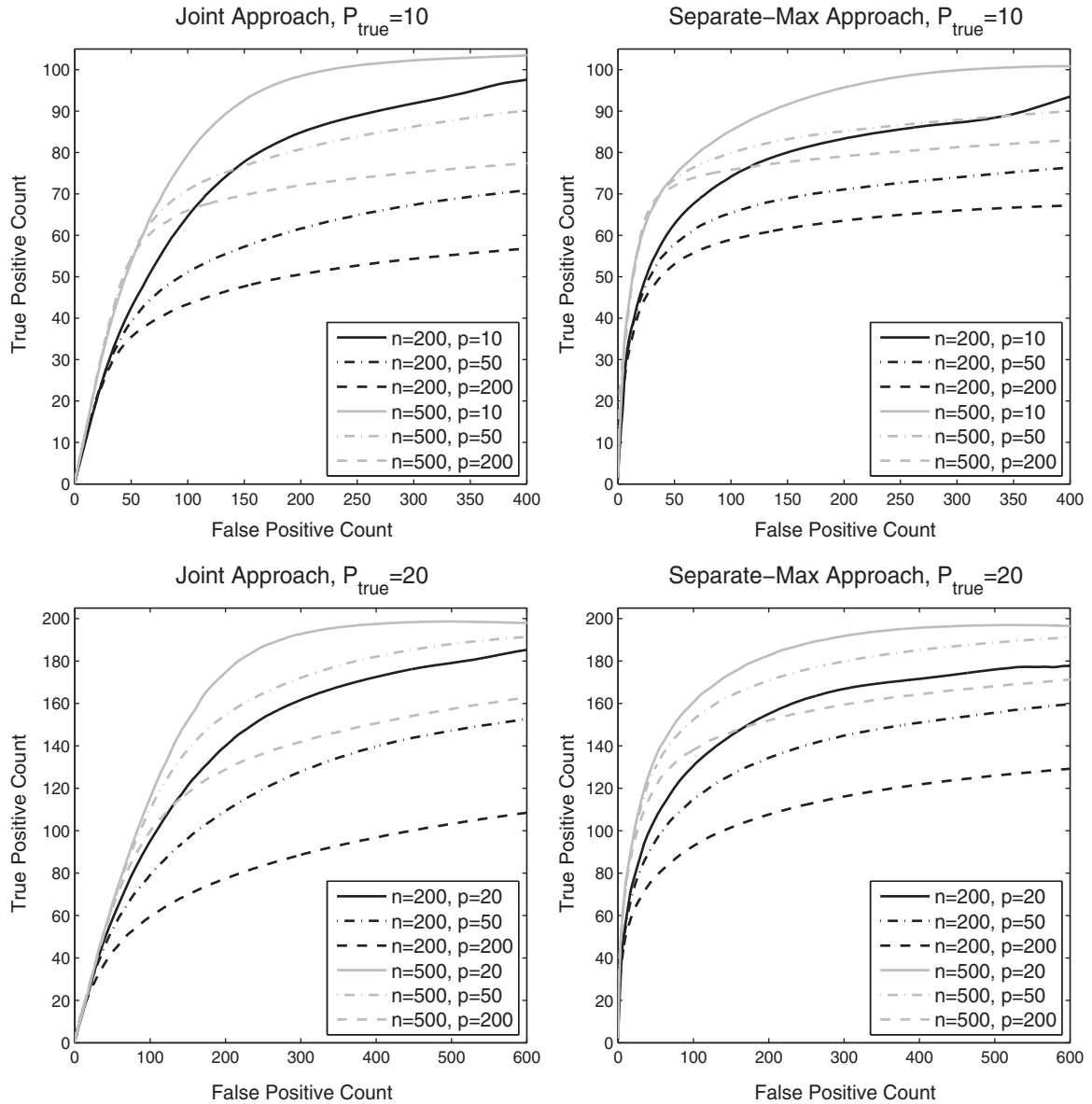
Figure 3. ROC curves for varying dimension, number of noise covariates, and sample size.

under high-dimensional settings. It is based on subsampling in combination with (high-dimensional) selection algorithms. Benefits of stability selection includes its insensitiveness to the amount of the regularization in the original selection algorithms and its improvement of model selection accuracy over them. Specifically, we repeatedly fit the model 100 times on random subsamples (without replacement) of half of the original data size. For each tuning parameter λ from a fixed grid of values, we record the frequency of $\hat{\theta}_{jk\ell}$ being non-zero respectively for each covariate X_ℓ, $\ell = 0, \ldots, p$ on all pairs of (j, k), $1 \leq j < k \leq q$, and denote it by $f_{jk\ell}(\lambda)$. Then we use $f_{jk\ell}^* = \max_\lambda f_{jk\ell}(\lambda)$ as a measure of importance of covariate X_ℓ for the edge (j, k). Finally, for each covariate X_ℓ, we rank the edges based on the selection frequencies $\{f_{jk\ell}^* : 1 \leq j < k \leq q\}$. At the top of the list are the edges that depend on X_ℓ most heavily.

We note that in previous simulation settings, even though the dimensions of the variables are not large, the number of parameters far exceeds the sample size. We applied stability selection coupled with the proposed methods, and found that the results of stability selection are comparable with those of not using stability selection without much improvement (see the Online Supplementary Materials Appendix C).

In this subsection, we mimic the real data example and increase the dimension of the binary responses q to a large value. Specifically, we set the sample size $n = 100$, the dimension of the covariates $p = 5$ and vary the dimension of binary responses $q = \{50, 100, 200\}$. The total number of non-zero parameters is fixed to be 150. The magnitude of the model parameters are set to be $\beta = 8$. We then compare the three proposed methods with stability selection. The results are shown in Figure 3 of the Online Supplementary Materials.

We found that as the dimension q increases, coupling stability selection with the joint approach and the separate-max approach significantly improve model selection performance over the respective original methods. Since the separate-min approach coupled with stability selection had the best performance overall, we adopted it for our data example in Section 5.

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. Four-hundred thirty 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 patterns 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 $\boldsymbol{y}_{143 \times 430}$, where y_j^i indicates the deletion status of the *jth* cytoband in the *ith* sample. Let $\boldsymbol{x}^i$ denote the covariate vector containing the three clinical phenotypes of the *ith* sample, and $\boldsymbol{x}_\ell$ the ℓth covariate vector. We first standardize the covariate matrix $\boldsymbol{x}_{143 \times 3}$ and then fit our Ising model with covariates with the separate-min fitting method. We then apply stability selection (Meinshausen and Bühlmann, 2010) to infer the stable set of important covariates for each pairwise conditional association. 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 of the Online Supplementary Materials, 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 column records the selection frequency.

There are 348 inter-chromosome interactions (between cytobands from different chromosomes) with selection probabilities at least 0.3. Among these, 75 interactions change with the TP53 status; 51 change with the ER status; and another 58 change with the tumor grade (see details in Table 1 of the Online Supplementary Materials). 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 (83% 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, Merterns, and Johansson, 1997). This region contains several putative tumor suppressor genes (TSG), including DNA-damage repair genes such as 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.

For visualization, we constructed separate graphs for each covariate, where each graph includes all the edges depending on that covariate with selection frequency at least 0.3. Specifically, we use the results from Table 1 of the Online Supplementary Materials to create Figure 4. In Figure 4, the "main effect" subplot shows all edges in the first two columns of Table 1 of the Online Supplementary Materials, which correspond to non-zero $\hat{\theta}_{jk0}$ parameters and the edges are weighted by their selection frequency as shown in the same table. The remaining three subplots are created in the same fashion. For nodes that have at least three neighbors in the covariate dependent plots, we also mark the names of them. To make the graph readable, we did not include all 430 nodes but only the nodes with edges.

Since there are obvious hubs (highly connected nodes) in the graph, which often have important roles in genetic regulatory pathways, we also did extra analysis to confirm the
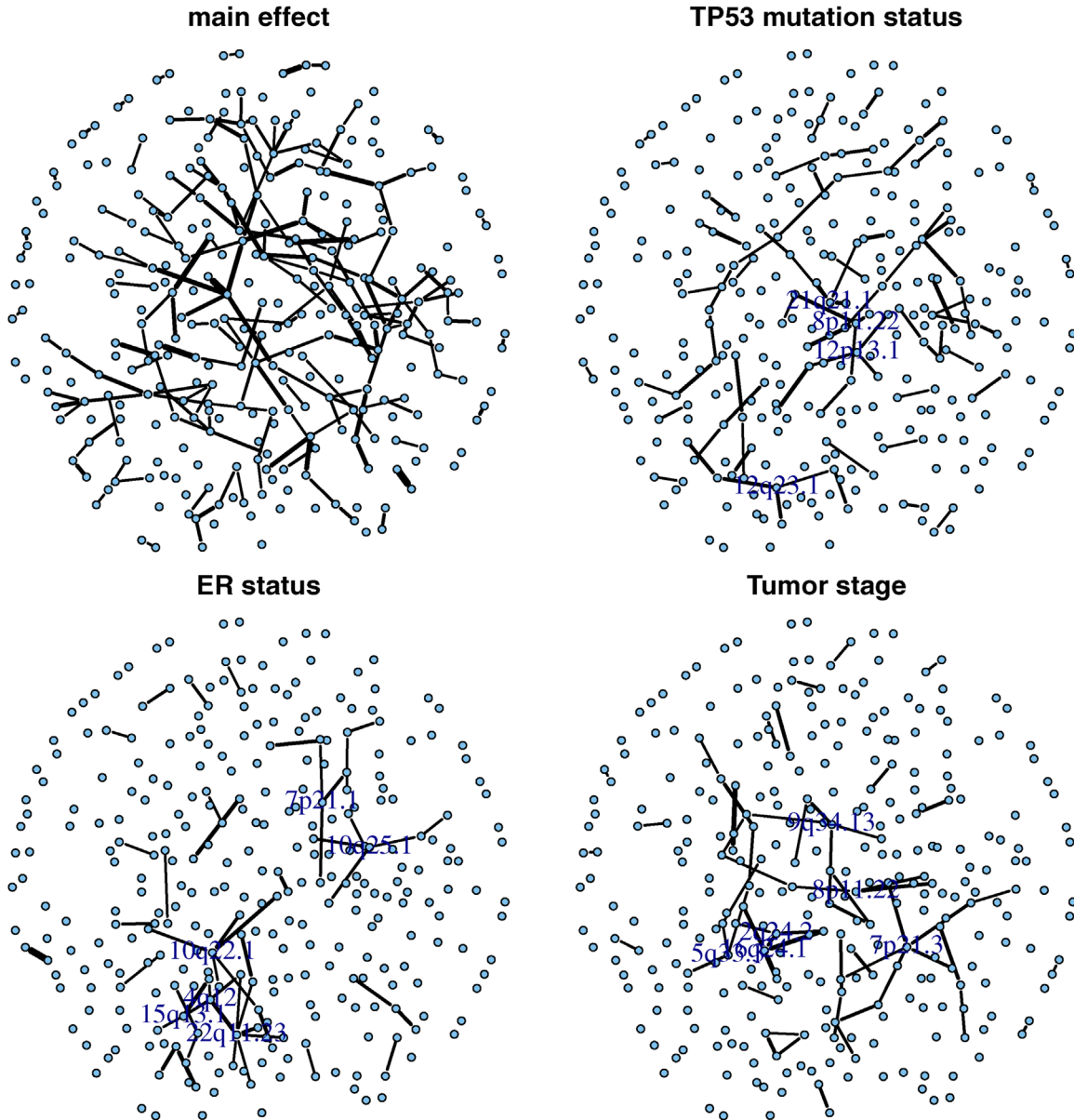
main effect

TP53 mutation status

ER status

Tumor stage

Figure 4. Graphs of edges depending on each covariate (based on Table 1 of the Online Supplementary Materials).

findings. As 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 $d_{j,\ell}^m = \#\{k : \hat{\theta}_{jk\ell} \neq 0\}$. A ranking of nodes can then be produced for each covariate ℓ and each replication m, with $r_{j,\ell}^m$ being the corresponding rank. Finally, we compute the median rank across all stability selection subsamples $r_{j,\ell} = \text{median}\{r_{j,\ell}^m, m = 1, \dots, 100\}$, and order nodes by rank for each covariate. The results are listed in Table 2 of the Online Supplementary Materials. Interestingly, cytoband 8p11.22 was ranked close to the top for all three covariates. The 8p11-p12 genomic region plays an important role in breast cancer, as numerous studies have identified this region as the location of multiple oncogenes and tumor suppressor genes (Adélaïde et al., 1998; Yang et al., 2006). High

frequency of loss of heterozygosity (LOH) of this region in breast cancer has also been reported (Adélaïde 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 of the Online Supplementary Materials, we can also see that the deletion of cytoband 8p11.22 region is associated with the deletion of cytoband 6p21.32 with relatively high confidence (selection frequency = 0.46); and these associations change with both TP53 status. This finding is interesting because high frequency LOH at 6q 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; Devilee et al., 1991; Negrini

et al., 1994). 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, then when the value of a covariate changes, only the strength of the links is affected, and there is zero probability for the graph structure to change. 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. With binary covariates, which is the case in our motivating application, the situation is different; for example, since edge (j, k) depends on the value $\theta_{jk0} + \boldsymbol{\theta}_{jk}^T \boldsymbol{x}$, the graph structure can change in the following situation: when $\theta_{jk0} = 0$ and $\boldsymbol{\theta}_{jk}^T \neq \boldsymbol{0}$, for a subject at the reference level where the entire $\boldsymbol{x}$ is 0, edge (j, k) does not exist, but edge (j, k) will exist if any of the "important" covariate is at the non-reference level. It is the same case with categorical covariates with more than two levels. 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.

7. Supplementary Materials

Web Appendices A–F, referenced in Sections 3, 4, and 5, along with the code, are available with this paper at the *Biometrics* website on Wiley Online Library.

References

Adélaïde, J., Chaffanet, M., Imbert, A., Allione, F., Geneix, J., Popovici, C., van Alewijk, D., Trapman, J., Zeillinger, R., Børrensen-Dale, A. L., Lidereau, R., Birnbaum, D., and Pe'busque, M. J. (1998). Chromosome region 8p11-p21: Refined mapping and molecular alterations in breast cancer. *Genes, Chromosomes & Cancer* **22**, 186–199.

Ali, I. U., Lidereau, R., Theillet, C., and Callahan, R. (1987). Reduction to homozygosity of genes on chromosome 11 in human breast neoplasia. *Science* **238**, 185–188.

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

Bergamaschi, A., Kim, Y. H., Wang, P., Sørlie, T., Hernandez-Boussard, T., Lonning, P. E., Tibshirani, R., Børrensen-Dale, A. L., and Pollack, J. R. (2006). Distinct patterns of DNA copy number alteration are associated with different clinico-pathological features and gene-expression subtypes of breast cancer. *Genes, Chromosomes & Cancer* **45**, 1033–1040.

Cai, T. T., Li, H., Liu, W., and Xie, J. (2013). Covariate adjusted precision matrix estimation with an application in genetical genomics. *Biometrika* **100**, 139–156.

Cai, T. T., Liu, W., and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of American Statistical Association* **106**, 594–607.

Conte, N., Charafe-Jauffret, E., Adélaïde, J., Ginestier, C., Geneix, J., Isnardon, D., Jacquemier, J., and 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., and El Ghaoui, L. (2008). First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and its Applications* **30**, 56–66.

Devilee, P., van Vliet, M., van Sloun, P., Kuipers Dijkshoorn, N., Hermans, J., Pearson, P. L., and Cornelisse, C. J. (1991). Allelotype of human breast carcinoma: A second major site for loss of helcrozygosity is on chromosome 6q. *Oncogene* **6**, 1705–1711.

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

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularized paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1–22.

Fu, W. J. (1998). Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics* **7**, 397–416.

Geisler, S., Børresen-Dale, A. L., Johnsen, H., Aas, T., Geisler, J., Akslen, L. A., Anker, G., and Lønning, P. E. (2003). Tp53 gene mutations predict the response to neoadjuvant treatment with 5-fluorouracil and mitomycin in locally advanced breast cancer. *Clinical Cancer Research* **9**, 5582–5588.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2010). Joint estimation of multiple graphical models. *Biometrika* **98**, 1–15.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2010). Joint structure estimation for categorical Markov networks. Technical Report #507, Department of Statistics, University of Michigan, Ann Arbor.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2014). Estimating heterogeneous graphical models for discrete data with an application to roll call voting. *Annals of Applied Statistics.* Accepted.

Hassner, M. and Sklansky, J. (1980). The use of Markov random fields as models of texture. *Computer Graphics Image Processing* **12**, 357–370.

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

Huang, T. H., Yeh, P. L., Martin, M. B., Straub, R. E., Gilliam, T. C., Caldwell, C. W., and Skibba, J. L. (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. and 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. A., and 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., and Wasserman, L. (2010). Graph-valued regression. *Proceedings of Advances in Neural Information Processing Systems (NIPS)* **23**, 1423−1431. Curran Associates Inc.

Manning, C. and Schutze, H. (1999). *Foundations of Statistical Natural Language Processing.* Boston, MA: MIT Press.

Mazumder, R. and Hastie, T. (2012). Exact covariance thresholding into connected components for large-scale graphical lasso. *Journal of Machine Learning Research* **13**, 781−794.

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

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Soceity* **72**, 417−473.

Minella, A. C., Grim, J. E., Welcker, M., and Clurman, B. E. (2007). p53 and scffbw7 cooperatively restrain cyclin e-associated genome instability. *Oncogene* **26**, 6948−6953.

Mitelman, F., Merterns, F., and Johansson, B. (1997). A break-point 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., and Croce, C. M. (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., and Zhu, J. (2009). Partial correlation estimation by joint sparse regression model. *Journal of the American Statistics Association* **104**, 735−746.

Ravikumar, P., Raskutti, G., Wainwright, M. J., and Yu, B. (2008). Model selection in Gaussian graphical models: High-dimensional consistency of l1-regularized MLE. *Advances in Neural Information Processing Systems (NIPS)* **21**, 1329−1336. Curran Associates Inc.

Ravikumar, P., Wainwright, M. J., and Lafferty, J. D. (2010). High-dimensional ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics* **38**, 1287−1319.

Rocha, G. V., Zhao, P., and Yu, B. (2008). A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (splice). Technical Report 759, Department of Statistics, UC Berkeley.

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

Sinha, S., Singh, R. K., Alam, N., Roy, A., Rouchoudhury, S., and Panda, C. K. (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* **7**, 84.

Wang, P., Chao, D. L., and Hsu, L. (2011). Learning oncogenic pathways from binary genomic instability data. *Biometrics* **67**, 164−173.

Witten, D. M., Friedman, J. H., and 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. Q., Streicher, K. L., Ray, M. E., Abrams, J., and Ethier, S. P. (2006). Multiple interacting oncogenes on the 8p11-p12 amplicon in human breast cancer. *Cancer Research* **66**, 11632−11634.

Yin, J. and 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. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94**, 19−35.

Zhao, H., Langerød, A., Ji, Y., Nowels, K. W., Nessland, J. M., Tibshirani, R., Bukholm, I. K., Kåresen, R., Botstein, D., and Børresen-Dale, A. L. (2004). Different gene expression patterns in invasive lobular and ductal carcinomas of the breast. *Molecular Biology of the Cell* **15**, 2523−2536.