Mixed and Covariate Dependent Graphical Models

by

Jie Cheng

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2013

Doctoral Committee:

Associate Professor Elizaveta Levina, Co-Chair
Professor Ji Zhu, Co-Chair
Professor Kerby Shedden
Professor Peter X.K. Song

To my parents

ACKNOWLEDGEMENTS

I am most grateful for having Prof. Ji Zhu and Prof. Liza Levina as my academic advisors. I have benefited immensely from their insightful suggestions and caring personalities. Their constant encouragements and attention to details have made my growth as a researcher possible through the past five years. I would also like to thank my committee members Prof. Kerby Shedden and Prof. Peter X.K Song for their time and helpful suggestions throughout my dissertation. I owe additional thanks to Kerby for his patient help and guidance with the course work at the beginning of my PhD. Moreover, I want to express my deepest appreciation to my beloved parents for their constant support. The friends I have made here in Michigan have made my PhD life so much more colorful and enjoyable. Special mentions go out to Yeo Jung Park, with whom I share every story; Xueying Yu, from whom I feel the warmth of an elder sister; Emmy Huitian Lei, Lucy Lu and Cen Guo, whom I had the pleasure of knowing and becoming great friends with. Last but not the least, I feel thankful that I met Ashin Mukherjee, the most kind, delightful and caring person, with whom I will happily share the rest of the journey.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Mixed and Covariate Dependent Graphical Models

by

Jie Cheng

Co-Chairs: Assoc. Prof. Elizaveta Levina and Prof. Ji Zhu

Graphical models have proven to be a useful tool in understanding the conditional dependency structure of multivariate distributions. In Chapters II and III of the thesis, we consider two types of undirected graphical models that are motivated by particular types of applications. The first model we consider is a mixed graphical model, linking both continuous and discrete variables. The proposed model is simple enough to be suitable for high-dimensional data, yet flexible enough to represent all possible graph structures for mixed types of data. We develop a computationally efficient regression-based algorithm for fitting the model by focusing on the conditional log-likelihood of each variable given the rest. The parameters have a natural group structure, and sparsity in the fitted graph is attained by incorporating a group lasso penalty, approximated by a weighted ℓ_1 penalty for computational efficiency. We demonstrate the effectiveness of our method through an extensive simulation study and apply it to a music annotation data set (CAL500), obtaining a sparse and interpretable graphical model relating the continuous features of the audio signal to categorical variables such as genre, emotions, and usage associated with particular songs.

The second model we consider is a sparse covariate dependent Ising model which allows us 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 subjects' covariates influence the strength of association between two 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 proved. The results on the genetic tumor data set and their biological significance are discussed in detail.

Another problem of interest that also involves multivariate binary data is the multi-label classification problem, which has broad applications in text mining, media annotation and the study of gene functions. In Chapter IV, we propose a simplified covariate dependent Ising model to model the joint distribution of the binary responses, as well as two extensions adapted to high-dimensional data sets. The proposed methods are compared on a number of benchmark data sets. We also address the issue of choosing the best prediction rule for a particular measure of classification performance, since in mutli-label classification there are multiple ways to define classification error.

CHAPTER I

Introduction

1.1 Background and Literature Review

Graphical models have proven to be a useful tool in representing the conditional dependency structure of multivariate distributions. The undirected graphical model in particular, sometimes also referred to as the Markov network, has drawn a lot of attention over the past decade. They 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 and Sklansky*, 1980; *Woods*, 1978), 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.

The goal is to recover the graph structure, i.e., the edge set E, from an i.i.d. sample drawn from the underlying Markov network. 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. For the multivariate Gaussian case, let us denote the random vector by $\boldsymbol{X} = (X_1, \ldots, X_p) \sim N(\boldsymbol{\mu}, \Sigma)$. We observe n i.i.d.

data points $\{\boldsymbol{x}_i\}_{i=1}^n$ from the aforementioned distribution. The graph structure E is completely specified by the off-diagonal elements of the inverse covariance matrix, also known as the precision matrix, $\Omega = \Sigma^{-1}$. The (i,j)th element of Ω, $\omega_{ij} = 0$ implies that X_i is independent ofX_j given the rest of the variables. Therefore, estimating the edge set E is equivalent to identifying the non-zero off-diagonal entries of the precision matrix. Early works of *Dempster* (1972) proposed setting elements of the concentration matrix to zero to encourage selection of simpler models. But this work relies on the classical setting where $n > p$ and fails to guarantee positive-definiteness of the final estimate. In recent years this topic of inverse covariance matrix estimation have received a great deal of attention, with a focus on the high-dimensional framework. *Meinshausen and Bühlmann* (2006) proposed a neighborhood selection approach by regressing each variable on the rest using a lasso (*Tibshirani*, 1996) penalty to encourage sparsity, with a slight misuse of the notation, let $\boldsymbol{X}$ denote the $n \times p$ data matrix from a Gaussian graphical model, the above mentioned approach solves the following

$$\widehat{\boldsymbol{\beta}}_i^{\lambda} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p-1}}{\operatorname{argmin}} \left\{ \frac{1}{n} \|\boldsymbol{X}_i - \boldsymbol{X}_{-i}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}, \quad i = 1, \ldots, p.$$

where $\boldsymbol{X}_{-i}$ denotes the data matrix after removing the i-th column $\boldsymbol{X}$; $\|\cdot\|_2$ refers to the squared root of the Euclidean norm of a vector and $\|\cdot\|_1$ refers to the sum of absolute value of all elements of a vector. The authors show that their methods are model selection consistent for sparse graphs under appropriate choice of tuning parameter. *Yuan and Lin* (2007) and *Banerjee et al.* (2008) used penalized likelihood approach to directly estimate a sparse precision matrix in the Gaussian graphical models. After simplification of the log-likelihood the optimization problem can be written as

$$\widehat{\Omega} = \underset{C \succ 0}{\operatorname{argmin}} \left\{ -\log\det(C) + \operatorname{trace}(SC) + \lambda \|C\|_1 \right\},$$

where S denotes the sample covariance matrix and $C \succ 0$ implies that the optimization is restricted over the space of positive definite matrices. Here we should also mention that *Yuan and Lin* (2007) does not penalize the diagonal entries of the precision matrix. The optimization problem is non-trivial due to the positive-definiteness constraint on the precision matrix. *Yuan and Lin* (2007) used the max-det algorithm to solve the optimization problem which restricts their application to small dimensions. On the other hand *Banerjee et al.* (2008) proposed a faster semi-definite programing approach based on Nesterov's method for interior point optimization. Both works note that the simpler approach by *Meinshausen and Bühlmann* (2006) can be viewed as an approximate solution to the exact problem. *Friedman et al.* (2008) used this as a launching pad to develop an extremely fast computational algorithm based on co-ordinate descent method called Graphical Lasso (glasso). Recent works by *Witten et al.* (2011) and *Mazumder and Hastie* (2012) have proposed further improvements and insights to the graphical lasso algorithm. Several other methodological works also focused on the penalized likelihood or pseudo-likelihood approach which include *Rocha et al.* (2008); *Rothman et al.* (2008); *Peng et al.* (2009); *Yuan* (2010); *Cai et al.* (2011b). Many of them also establish asymptotic properties such as consistency and sparsistency under high-dimensional settings. *Lam and Fan* (2009) and *Ravikumar et al.* (2008) proved model selection consistency and convergence rates for covariance and inverse covariance matrix estimation under high-dimensional settings.

The Ising model which originated in statistical physics literature has been a popular choice for multivariate binary data. It can be thought of as a counterpart of multivariate Gaussian distribution. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_q)' \in \{0,1\}^q$ denote the random binary vector. The probability mass function is given by

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

where $\boldsymbol{\theta}$ denotes the parameter vector and $Z(\boldsymbol{\theta})$ stands for the normalization constant. $\theta_{jk} = 0$ implies that $Y_j \perp Y_k$ given the rest of the variables which bears a strong similarity to the Gaussian case for continuous variables. However, penalized log-likelihood approaches are significantly more difficult due to the intractable normalizing constant $Z(\boldsymbol{\theta})$ also known as the partition function. *Banerjee et al.* (2008) tried to overcome this by solving an approximate problem where they replace the log-partition function by a log-determinant relaxation. This allows them to use interior point methods to solve the optimization problem. Further improvement of this approach is possible by considering more accurate bounds for the log-partition function at the cost of more complex optimization problems. *Ravikumar et al.* (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. They also prove theoretical guarantees of their method such as, estimation and model selection consistency. Another pseudo-likelihood based algorithm was developed by *Höfling and Tibshirani* (2009) and analyzed by *Guo et al.* (2010c). *Cai et al. (2012)* used composite likelihoood methods to fit Ising models.

In machine learning community, another problem of interest that arises in the context of the multivariate binary data is the multi-label classification problem, which has seen much activity during the past few years. Classification problems arise in a broad spectrum of real life applications where the response variable of a predictive task is categorical. The training data for classification problem are usually in the form of $\{(\mathbf{x}_1, \mathcal{Y}_1), (\mathbf{x}_2, \mathcal{Y}_2), \ldots, (\mathbf{x}_n, \mathcal{Y}_n)\}$, where $\mathbf{x}_i \in \mathbb{R}^p$ is the feature vector of the ith sample and $\mathcal{Y}_i \subseteq \mathcal{L} = \{l_1, l_2, \ldots, l_q\}$ with $\mathcal{L}$ being the set of all the possible labels. The objective is to construct a classification rule $f : \mathbb{R}^p \to \mathcal{L}$ such that for any future input $\mathbf{x}$, we can predict its associated label set with high accuracy. Single-label Classification restricts each instance to only one label, i.e, $\|\mathcal{Y}_i\|_0 = 1$ ($\|\cdot\|_0$ refers to the total number of elements in the set). Within the single-label classification

regime, the tasks are categorized further based on the total number of labels in $\mathcal{L}$, if $\|\mathcal{L}\|_0 = 2$, we call it a binary classification problem; if $\|\mathcal{L}\|_0 > 2$, it is called a multi-class classification problem. In many real applications observations can actually be assigned multiple labels which need not be mutually exclusive; this is known as the multi-label classification problem. Multi-label classification originated in information retrieval and text mining, where each document is possibly associated with a set of relevant labels rather than one. For example, a news article can be categorized both as 'political' and 'financial'. In more recent years, multi-label classification has been applied in a much broader range of applications such as image /audio /video annotation and gene functional analysis. In a multi-label classification task, we allow each subject to be associated with a collection of labels rather than one label, i.e. we assume $\|\mathcal{Y}_i\|_0 \geq 1$. The key difference of a multi-label classification task from single label classification is that the labels may neither be mutually exclusive nor conditionally independent given the features of the samples. There has been a wide range of literature on multi-label classification, where some methods tend to solve individual classification problems of each label, some methods transform the problem into one single-label classification problem and use the existing classification tools, and some other methods take a completely different approach by ranking the labels and output the top of the ranked list. Previous literature will be discussed in detail in the Chapter IV of the thesis.

1.2 Outline of the Thesis

In Chapter II we introduce a novel graphical model for mixed types of data that consist of some binary variables and some continuous ones. This kind of mixed data are common in many scientific applications but the statistical literature on graphical models have surprisingly few works addressing this type of data. The proposed model is based on the general conditional Gaussian density but significantly simplified so

that the new model is able to represent all possible graph structures without sacrificing computational tractability which makes it suitable for the analysis of high-dimensional data. A computationally efficient regression based model fitting method is developed by maximizing the conditional log-likelihood of each variable given the rest. Natural grouping of the parameters allow us to use a group lasso penalty to encourage sparsity. For computational efficiency we approximate the group lasso penalty by weighted ℓ_1 penalty. Simulation studies and an application to a music annotation data set (CAL500) demonstrates the effectiveness of our method. We also discuss an extension of the proposed methodology to general discrete variables.

Often additional covariates are available along with the multivariate binary data, which may influence the conditional dependence structure between the binary variables. Motivated by such a data set on genomic instability on tumor samples of several types we propose a sparse covariate dependent Ising model in Chapter III. This results in a subject-specific Ising model where the covariates influence the strength of association between the genes. We introduce an ℓ_1 penalty to achieve sparsity in the estimated graph as well as the number of selected covariates. Two computational algorithms are developed to fit the proposed model using co-ordinate descent approach and their asymptotic theoretical properties are established. We conclude with a detailed analysis of the tumor data set and their biological significance.

Multi-label classification refers to the scenario in classification that each instance is associated with a subset of labels rather than one. The labels are not mutually exclusive and often correlated. In Chaper IV, we propose to use a binary Markov network, i.e. Ising model with covariates, to explicitly model the conditional distribution $P(\boldsymbol{y}|\boldsymbol{x})$ for a multi-label classification problem. Pseudo-likelihood is adopted to develop a computationally efficient estimation procedure. We also investigate the choice of evaluation measures in connection to different prediction rules, which is further illustrated by numerical studies. Moreover, we consider two alternative approaches

motivated from the previously proposed model to handle data sets with large dimensional responses. We then apply the proposed methods on four benchmark multi-label data sets and compare their prediction performance.

CHAPTER II

High-Dimensional Mixed Graphical Models

2.1 Introduction

The vast majority of the graphical models literature has been focusing on either the multivariate Gaussian model (*Meinshausen and Bühlmann*, 2006; *Yuan and Lin*, 2007; *Rothman et al.*, 2008; *Banerjee et al.*, 2008; *Rocha et al.*, 2008; *Ravikumar et al.*, 2008; *Lam and Fan*, 2009; *Peng et al.*, 2009; *Yuan*, 2010; *Cai et al.*, 2011b; *Friedman et al.*, 2008), or the Ising model for binary and discrete data (*Höfling and Tibshirani*, 2009; *Ravikumar et al.*, 2010; *Guo et al.*, 2010c). The properties of these models are by now well understood and studied both in the classical and the high-dimensional settings. Both these models only deal with variables of one kind – either all continuous variables in the Gaussian model or all binary variables in the Ising model (extensions of the Ising model to general discrete data, while possible in principle, are rarely used in practice). In many applications, however, data sources are complex and varied, and frequently result in mixed types of data, with both continuous and discrete variables present in the same dataset. In this paper, we will focus on graphical models for this type of mixed data (mixed graphical models).

The conditional Gaussian distribution was originally proposed (*Lauritzen and Wermuth*, 1989; *Lauritzen*, 1996) to model mixed data and has become the foundation of most developments on this topic. In the original paper, *Lauritzen and Wermuth*

(1989) define a general form of the conditional Gaussian density and characterize the connection between the model parameters and the conditional associations among the variables. The model is fitted by maximum likelihood, but the number of parameters in this model, however, grows exponentially with the number of variables, which renders this approach unsuitable for high-dimensional problems arising in many modern applications. Much more recently, *Lee and Hastie* (2012) and *Fellinghauer et al.* (2011) have studied the mixed graphical model (simultaneously and independently of the present work), under a setting that could be viewed as a simplified special case of our proposal. A more detailed discussion of these papers is postponed to Section 2.6.

In this chapter, we propose a simplified version of the conditional Gaussian distribution which reduces the number of parameters significantly yet maintains flexibility. To fit the model in a high-dimensional setting, we impose a sparsity assumption on the underlying graph structure and develop a node-based regression approach with the group lasso penalty (*Yuan and Lin*, 2006), since edges in the mixed graphical model are associated with groups of parameters. The group lasso penalty in itself is not computationally efficient, and we develop a much faster weighted ℓ_1 approximation to the group penalty which is of independent interest. The simulation results show promising model selection performance in terms of estimating the true graph structure under high-dimensional settings.

To start with, we give a brief introduction to the conditional Gaussian distribution and its Markov properties following *Lauritzen* (1996).

Conditional Gaussian (CG) density: let $X = (Z, Y)$ be a mixed random vector, where $Z = (Z_j)_{j\in\Delta}$ is a q-dimensional discrete sub-vector, and $Y = (Y_\gamma)_{\gamma\in\Gamma}$ is a p-dimensional continuous sub-vector. The conditional Gaussian density $f(x)$ is defined as

$$f(x) = f(z, y) = \exp\left(g_z + h_z^T y - \frac{1}{2} y^T K_z y\right), \tag{2.1}$$

where $\{(g_z, h_z, K_z), g_z \in \mathbb{R}, h_z \in \mathbb{R}^p, K_z \in \mathbb{R}^+_{p\times p}, z \in \text{Range}(Z)\}$ are the canonical parameters of the distribution. The following equations connect the canonical parameters in (2.1) to the moments (P_z, ξ_z, Σ_z):

$$\begin{aligned} P_z &= P(Z=z) = (2\pi)^{p/2}(\det(K_z))^{-1/2}\exp\left(g_z + h_z^T K_z^{-1} h_z/2\right), \\ \xi_z &= \mathbb{E}(Y|Z=z) = K_z^{-1}h_z, \\ \Sigma_z &= \text{Var}(Y|Z=z) = K_z^{-1}, \end{aligned}$$

and the conditional distribution of Y given $Z = z$ is $\mathcal{N}(\xi_z, \Sigma_z)$.

The next theorem relates the graphical Markov property of the model to its canonical parameters and serves as the backbone of the subsequent analysis.

Theorem II.1. *Represent the canonical parameters from* (2.1) *by the following expansions,*

$$g_z = \sum_{d:d\subseteq\Delta} \lambda_d(z), \quad h_z = \sum_{d:d\subseteq\Delta} \eta_d(z), \quad K_z = \sum_{d:d\subseteq\Delta} \Phi_d(z)\ , \tag{2.2}$$

where functions indexed by d only depend on z through z_d. Then a CG distribution is Markovian with respect to a graph $\mathcal{G}$ if and only if the density has an expansion that satisfies

$$\begin{aligned} \lambda_d(z) &\equiv 0 \qquad \textit{unless } d \textit{ is complete in } \mathcal{G}, \\ \eta_d^\gamma(z) &\equiv 0 \qquad \textit{unless } d\cup\{\gamma\} \textit{ is complete in } \mathcal{G}, \\ \Phi_d^{\gamma\mu}(z) &\equiv 0 \qquad \textit{unless } d\cup\{\gamma,\mu\} \textit{ is complete in } \mathcal{G}. \end{aligned}$$

where $\eta_d^\gamma(z)$ is the γ-th element of $\eta_d(z)$, $\Phi_d^{\gamma\mu}(z)$ is the $\gamma\mu$-th element of $\Phi_d(z)$, and a subgraph is called complete if it is fully connected.

The rest of the chapter is organized as follows. Section 3.2 introduces the simplified mixed graphical model which has just enough parameters to cover all possible graph

structures, proposes an efficient estimation algorithm for the model, and discusses theoretical guarantees for the proposed method under the high-dimensional setting. Section 2.3 uses several sets of simulation studies to evaluate the model selection performance and compare some alternative choices of regularization. In Section 2.4, the proposed model is applied to a music annotation data set CAL500 with binary labels and continuous audio features. In Section 2.5, we describe the generalization of the model from binary to discrete variables. Finally, we conclude with discussion in Section 2.6.

2.2 Methodology

Our main contribution is a simplified but flexible conditional Gaussian model for mixed data. Model fitting is based on maximizing the conditional log-likelihood of each variable given the rest for computational tractability. This leads to penalized regression problems with an overlapping group structure of the parameters, the natural solution to which is to fit separate regressions with an overlapping group lasso penalty. This is computationally quite expensive, so we approximate the overlapping group lasso penalty by an appropriately rescaled ℓ_1 penalty.

2.2.1 The simplified mixed graphical model

Recall we partition the random vector $\mathrm{X} = (Z_1, Z_2, \ldots, Z_q, Y_1, Y_2, \ldots, Y_p)$ into the binary part $Z_j \in \{0, 1\}$, $j = 1, \ldots, q$, and the continuous $Y_\gamma \in \mathbb{R}$, $\gamma = 1, \ldots, p$. We

propose to consider the conditional Gaussian distribution with the density function

$$\begin{aligned}
\log f(z,y) &= \sum_{d:d\subseteq\Delta,|d|\leq 2} \lambda_d(z) + \sum_{d:d\subseteq\Delta,|d|\leq 1} \eta_d(z)^T y - \frac{1}{2}\sum_{d:d\subseteq\Delta,|d|\leq 1} y^T \Phi_d(z) y \\
&= \left(\lambda_0 + \sum_j \lambda_j z_j + \sum_{j>k} \lambda_{jk} z_j z_k\right) + y^T \left(\eta_0 + \sum_j \eta_j z_j\right) - \frac{1}{2} y^T \left(\Phi_0 + \sum_{j=1}^{q} \Phi_j z_j\right) y \\
&= \left(\lambda_0 + \sum_j \lambda_j z_j + \sum_{j>k} \lambda_{jk} z_j z_k\right) + \sum_{\gamma=1}^{p} \left(\eta_0^\gamma + \sum_j \eta_j^\gamma z_j\right) y_\gamma \\
&\quad -\frac{1}{2}\sum_{\gamma,\mu=1}^{p} \left(\Phi_0^{\gamma\mu} + \sum_{j=1}^{q} \Phi_j^{\gamma\mu} z_j\right) y_\gamma y_\mu \ , \qquad (2.3)
\end{aligned}$$

where $\{\mathrm{diag}(\Phi_j)\}_{j=1}^q = \{\Phi_j^{\gamma\gamma}; j = 1,\ldots,q,\ \gamma = 1,\ldots,p\}$ are all 0 and λ_0 is the normalizing constant,

$$\lambda_0 = \left\{ (2\pi)^{p/2} \sum_{z\in\{0,1\}^q} \det(K_z)^{-1/2} \exp\left(\sum_j \lambda_j z_j + \sum_{j>k} \lambda_{jk} z_j z_k + h_z^T K_z^{-1} h_z/2\right)\right\}^{-1} .$$

Note that the density is explicitly defined via the expanded terms in (2.2) but the canonical parameters (g_z, h_z, K_z) can be obtained immediately by summing up the corresponding terms. This model simplifies the full conditional Gaussian distribution (2.1) in two ways: first, it omits all interaction terms between the binary variables of order higher than two, and second, it models the conditional covariance matrix and the canonical mean vector of the Gaussian variables as a linear function of the binary variables instead of allowing their dependence on higher order interactions of the binary variables. These simplifications reduce the total number of parameters from $O(p^2 2^{(p+q)})$ in the full model to $O\left(\max(q^2, p^2 q)\right)$. On the other hand, this model is the simplest CG density among those allowing for varying conditional covariance $\mathrm{Var}(Y|Z)$ that can represent all possible graph structures, since it includes interactions between all the continuous and discrete variables and thus allows for a fully connected graph, an empty graph, and everything in between. The fact that it allows

both the conditional mean and the conditional covariance of Y given Z to depend on Z adds flexibility.

2.2.2 Parameter estimation

Given sample data $\{(\boldsymbol{z}_i, \boldsymbol{y}_i)\}_{i=1}^n$, directly maximizing the log-likelihood $\sum_{i=1}^n \log f(\boldsymbol{z}_i, \boldsymbol{y}_i)$ is impractical due to the normalizing constant λ_0. The conditional likelihood of one variable given the rest, however, is of much simpler form and easy to maximize. Hence, we focus on the conditional log-likelihood of each variable and fit separate regressions to estimate the parameters, much in the spirit of the neighborhood selection approach proposed by *Meinshausen and Bühlmann* (2006) for the Gaussian graphical model and by *Ravikumar et al.* (2010) for the Ising model. To specify the conditional distributions, let $Z_{-j} = (Z_1, \ldots, Z_{j-1}, Z_{j+1}, \ldots, Z_q)$ and $Y_{-\gamma} = (Y_1, \ldots, Y_{\gamma-1}, Y_{\gamma+1}, \ldots, Y_p)$. Then the conditional distribution of Z_j given (Z_{-j}, Y) is described by

$$\log \frac{P(Z_j = 1 | Z_{-j}, Y)}{P(Z_j = 0 | Z_{-j}, Y)} = \lambda_j + \sum_{k \neq j} \lambda_{jk} Z_k + \sum_{\gamma=1}^{p} \eta_j^\gamma Y_\gamma - \frac{1}{2} \sum_{\gamma,\mu=1}^{p} \Phi_j^{\gamma\mu} Y_\gamma Y_\mu \ . \tag{2.4}$$

Since the conditional log-odds in (2.4) is linear in parameters, maximizing this conditional log-likelihood can be done via fitting a logistic regression with (Z_{-j}, Y, Y^2) as predictors and Z_j as response.

For the continuous variables, the conditional distribution of Y_γ given $(Y_{-\gamma}, Z)$ is given by

$$Y_\gamma = \frac{1}{K_z^{\gamma\gamma}} \left(\eta_0^\gamma + \sum_j \eta_j^\gamma Z_j - \sum_{\mu \neq \gamma} \left(\Phi_0^{\gamma\mu} + \sum_j \Phi_j^{\gamma\mu} Z_j \right) Y_\mu \right) + e_\gamma,$$

where $e_\gamma \sim \mathcal{N}(0, (K_z^{\gamma\gamma})^{-1})$. With $\mathrm{diag}(\Phi_j) = 0$ as defined by (4.1), we have $K_z^{\gamma\gamma} =$

$\Phi_0^{\gamma\gamma}$, i.e., the conditional variance of Y_γ does not depend on Z. Rewrite

$$Y_\gamma = \tilde{\eta}_0^\gamma + \sum_j \tilde{\eta}_j^\gamma Z_j - \sum_{\mu \neq \gamma} \left(\tilde{\Phi}_0^{\gamma\mu} + \sum_j \tilde{\Phi}_j^{\gamma\mu} Z_j \right) Y_\mu + e_\gamma \ , \tag{2.5}$$

where the redefined parameters with "tilde" are proportional to the original ones up to the same constant for each regression. Again, the conditional mean of Y_γ is linear in parameters, which can be estimated via ordinary linear regression with predictors $(Y_{-\gamma}, Z, Y_{-\gamma}Z)$ and response Y_γ.

2.2.3 Regularization

Based on Theorem II.1, the following equivalences hold:

$$\begin{aligned}
Z_j \perp Z_k \mid \ X \backslash \{Z_j, Z_k\} \ &\iff \ \lambda_{jk} = 0, \\
Z_j \perp Y_\gamma \mid \ X \backslash \{Z_j, Y_\gamma\} \ &\iff \ \boldsymbol{\theta}_{j\gamma} = \left(\eta_j^\gamma, \{\Phi_j^{\gamma\mu} : \mu \in \Gamma \backslash \{\gamma\}\}\right) = 0, \\
Y_\gamma \perp Y_\mu \mid \ X \backslash \{Y_\gamma, Y_\mu\} \ &\iff \ \boldsymbol{\theta}_{\gamma\mu} = \left(\Phi_0^{\gamma\mu}, \{\Phi_j^{\gamma\mu} : j \in \Delta\}\right) = 0.
\end{aligned} \tag{2.6}$$

This means that each edge between pairs of (Z_j, Y_γ) and (Y_γ, Y_μ) depends on a parameter vector, denoted by $\boldsymbol{\theta}_{j\gamma}$ and $\boldsymbol{\theta}_{\gamma\mu}$, respectively. To encourage sparsity of the edge set under high-dimensional settings, a natural choice here would be to use a group lasso penalty, such as the $\ell_1 \backslash \ell_2$ penalty proposed by *Yuan and Lin* (2006) for group lasso. The groups are pre-determined by parameter vectors corresponding to each edge. Denoting the loss function for the logistic regression of Z_j by ℓ_j and the linear regression for Y_γ by ℓ_γ, we have

$$\begin{aligned}
\ell_j &= -\frac{1}{n} \sum_{i=1}^n \log(P(z_{ij} \mid (\boldsymbol{z}_{i,(-j)}, \boldsymbol{y}_i)), \\
\ell_\gamma &= \frac{1}{n} \sum_{i=1}^n (y_{i\gamma} - (\tilde{\eta}_0^\gamma + \sum_{j=1}^q \tilde{\eta}_j^\gamma z_{ij} - \sum_{\mu \neq \gamma} (\tilde{\Phi}_0^{\gamma\mu} + \sum_{j=1}^q \tilde{\Phi}_j^{\gamma\mu} z_{ij}) y_{i\mu}))^2.
\end{aligned}$$

We estimate the parameters by optimizing the following criteria separately, for $j = 1, \ldots, q$ and $\gamma = 1, \ldots, \gamma$

$$\textit{Logistic regression}\text{: min } \ell_j \quad + \quad \rho \left(\sum_{k \neq j} \|\lambda_{jk}\|_1 + \sum_{\gamma=1}^{p} \|\boldsymbol{\theta}_{j\gamma}\|_2 \right), \tag{2.7}$$

$$\textit{Linear regression}\text{: min } \ell_\gamma \quad + \quad \rho \left(\sum_{\mu \neq \gamma} \|\tilde{\boldsymbol{\theta}}_{\gamma\mu}\|_2 + \sum_{j=1}^{q} \|\tilde{\boldsymbol{\theta}}_{j\gamma}\|_2 \right), \tag{2.8}$$

where ρ is the tuning parameter. We use the same tuning parameter for all regressions to simplify tuning, but they can also be tuned separately if the computational cost is not prohibitive. Another reason to use a single tuning parameter is to simplify the treatment of overlapping groups of parameters from different regressions (see more on this below). Note that in linear regression, the parameters in (2.5) denoted with "tilde" are proportional to the original parameters. The original parameters can be recovered by multiplying the estimates by $(\hat{K}_z^{\gamma\gamma})^{-1}$, which can be estimated from the mean squared error of the linear regression.

Although the optimization problems (2.7) and (2.8) each have the form of regular group lasso regressions, they can not be jointly solved by existing group lasso algorithms, because the groups of parameters involved in each regression overlap. Specifically, in logistic regression, the parameter $\Phi_j^{\gamma\mu}$ is part of both $\boldsymbol{\theta}_{j\gamma}$ and $\boldsymbol{\theta}_{j\mu}$ and determines both the edges (Z_j, Y_γ) and (Z_j, Y_μ); thus $\boldsymbol{\theta}_{j\gamma}$ has one parameter overlapping with each of the other $\boldsymbol{\theta}_{j\mu}$'s. Similarly, in linear regression, $\Phi_j^{\gamma\mu}$ is part of both $\boldsymbol{\theta}_{j\gamma}$ and $\boldsymbol{\theta}_{\gamma\mu}$, and affects both the edges (Z_j, Y_γ) and (Y_γ, Y_μ). This overlapping pattern creates additional difficulties in using the group penalty to perform edge selection. The overlapping group lasso problem has received limited attention from a computational point of view. *Yuan et al.* (2011) recently proposed an algorithm for solving the overlapping group lasso problem, but it is computationally intensive for high-dimensional data. Instead of optimizing the overlapping group penalty directly, we look for a surrogate penalty with similar properties that is easier to optimize and

thus more suitable for a high-dimensional setting.

The weighted ℓ_1 penalty can provide an upper bound for the group penalty, since for any vector $\boldsymbol{b}$, $\|\boldsymbol{b}\|_2 \leq \|\boldsymbol{b}\|_1$. For the logistic regression (2.7), for example, we have

$$\sum_{k \neq j} \|\lambda_{jk}\|_1 + \sum_{\gamma=1}^{p} \|\boldsymbol{\theta}_{j\gamma}\|_2 \leq \sum_{k \neq j} |\lambda_{jk}| + \sum_{\gamma=1}^{p} |\eta_j^{\gamma}| + 2 \sum_{\gamma < \mu} |\Phi_j^{\gamma\mu}| \ .$$

The surrogate on the right penalizes the overlapping parameters twice as much as the other parameters, which makes intuitive sense since incorrectly estimating the overlapping parameters as non-zero will add two wrong edges, while incorrectly estimating unique parameters for each group as non-zero will only add one wrong edge.

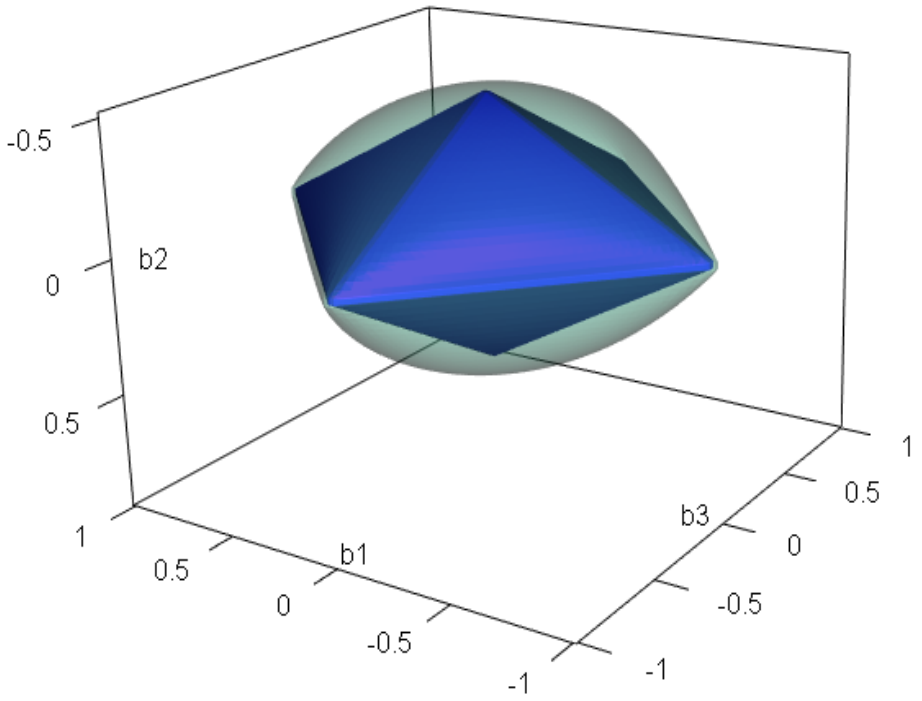

Figure 2.1: *Green (outside)*: $\{\boldsymbol{b} : \sqrt{b_1^2 + b_2^2} + \sqrt{b_3^2 + b_2^2} = 1\}$; *Blue (inside)*: $\{\boldsymbol{b} : |b_1| + |b_3| + 2|b_2| = 1\}$

To further illustrate why this upper bound provides a good approximation to the feasible region, consider the following toy example. Let the parameter vector be $\boldsymbol{b} = (b_1, b_2, b_3)$, with groups $\mathcal{G}_1 = (b_1, b_2)$ and $\mathcal{G}_2 = (b_2, b_3)$. The optimization problems

for the overlapping group lasso penalty and its ℓ_1 surrogate boil down to optimizing the same loss function on different feasible regions (for the same tuning parameter). Figure 2.1 compares the two feasible regions, $\mathcal{R}_1 = \{\boldsymbol{b} : \sqrt{b_1^2 + b_2^2} + \sqrt{b_3^2 + b_2^2} \leq 1\}$ and $\mathcal{R}_2 = \{\boldsymbol{b} : |b_1| + |b_3| + 2|b_2| \leq 1\}$. Since both the logistic loss and the least squares loss are smooth convex functions, the solutions will lie at singular points of the feasible regions $\mathcal{R}_1$ and $\mathcal{R}_2$. Note that $\mathcal{R}_2$ is not only a subset of $\mathcal{R}_1$ but it contains all four singular points of $\mathcal{R}_1$: $(\pm 1, 0, 0), (0, 0, \pm 1)$, which are also singular points in $\mathcal{R}_2$. Thus in this example, the optimum will be chosen from exactly the same set of singular points regardless of the penalty used. For higher dimensional $\boldsymbol{b}$, it still holds that $\mathcal{R}_2$ is a subset of $\mathcal{R}_1$, but it may not contain all of the singular points of $\mathcal{R}_1$ (this depends on the group structure). While that means that we may not have exactly the same optimal points, the approximation could still be good enough to identify the same non-zero groups, and this is what we found in practice.

With the group penalty replaced by the weighted ℓ_1 surrogate, we solve the following regression problems separately as an approximation to the original problems (2.7) and (2.8) to obtain the parameter estimates.

Logistic regression with ℓ_1 penalty: for $j = 1, \ldots, q$

$$\min \ \ell_j + \rho \left(\sum_{k \neq j} |\lambda_{jk}| + \sum_{\gamma=1}^{p} |\eta_j^\gamma| + 2 \sum_{\gamma < \mu} |\Phi_j^{\gamma\mu}| \right). \tag{2.9}$$

Linear regression with ℓ_1 penalty: for $\gamma = 1, \ldots, p$

$$\min \ \ell_\gamma + \rho \left(\sum_{j=1}^{q} |\tilde{\eta}_j^\gamma| + \sum_{\mu \neq \gamma} |\tilde{\Phi}_0^{\gamma\mu}| + 2 \sum_{j=1}^{q} \sum_{\mu \neq \gamma} |\tilde{\Phi}_j^{\gamma\mu}| \right). \tag{2.10}$$

Since we are estimating the parameters in separate regressions, all parameters determining edges will be estimated at least twice. This problem is common to all neighborhood selection approaches based on separate regressions, and is usually solved by taking either the largest or the smallest (in absolute value) of the estimates. For

the mixed Gaussian model, we found that taking the maximum of absolute values results in somewhat better model selection, and that is what we use throughout the paper as the final estimate. To fit both types of regressions with a weighted ℓ_1 penalty, we use the matlab package *glmnet* of *Friedman et al.* (2010).

2.2.4 Asymptotic Properties

Model selection consistency in high-dimensional regression problems is equivalent to estimating the non-zero parameters as non-zero. Correctly identifying the sign of non-zero parameters (in addition to correctly identifying them as non-zero) is referred to as sign consistency, while the usual ℓ_2 convergence of all the estimated parameters is referred to as norm consistency. Regression with ℓ_1 penalty has been shown to possess both sign consistency and norm consistency under certain regularity conditions, of which the most important one is the irrepresentable condition (see *Meinshausen and Bühlmann* (2006); *Zhao and Yu* (2006); *Van de Geer and Bühlmann* (2009) for linear regression and *Ravikumar et al.* (2010); *Van de Geer* (2008) for logistic regression). To show consistency of our method, we only need to require the main assumptions of the existing results to hold on a rescaled version of the original design matrix for each regression. Since we fit weighted ℓ_1-penalized regressions with fixed predetermined weights, we can treat the parameters multiplied by the weights as new parameters, and rescale the design matrix for each regression by dividing each column the corresponding weights. This converts the problem to a standard ℓ_1-penalized regression, and consistency is thus guaranteed by existing results cited above.

2.3 Numerical performance evaluation

This section includes simulation studies on evaluating model selection performance under different settings and comparing alternative choices of penalties. The results are summarized in ROC curves, where we plot the true positive count (TP) against

false positive count (FP) and the true positive rate (TPR) against the false positive rate (FPR), for both parameters and edges over a fine grid of tuning parameters. Let $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$ denote the true parameter vector and the fitted parameter vector respectively (without the intercept terms in the regressions), the parameter based quantities are defined as follows,

$$\begin{aligned} \text{TP} &= \#\{j : \hat{\theta}_j \neq 0 \text{ and } \theta_j \neq 0\}, \quad \text{FP} = \#\{j : \hat{\theta}_j \neq 0 \text{ and } \theta_j = 0\}, \\ \text{TPR} &= \frac{\text{TP}}{\#\{j : \theta_j \neq 0\}}, \qquad\qquad \text{FPR} = \frac{\text{FP}}{\#\{j : \theta_j = 0\}} \end{aligned}$$

The quantities based on the true edge set $\boldsymbol{E}$ and the estimated edge set $\hat{\boldsymbol{E}}$ are defined in a similar fashion, with parameter sets replaced by edge sets.

2.3.1 Model selection performance

In the first simulation, we fix the variable dimensions to be $p = 90$ continuous variables and $q = 10$ discrete variables, with the sample size $n = 100$. We vary the maximum node degree (i.e., the number of edges from the node) in the graph while maintaining the total number of edges fixed at 80. This results in graph structures with varying "uniformity", because given the total number of edges in a graph, the smaller the maximum node degree, the more uniform the graph is. A chain graph, for example, is the most extreme case of a uniform graph.

The data are generated as follows. First we generate the underlying graph structure given a maximum node degree: for the case where the maximum node degree is 2, we use a chain graph of the first 81 nodes; when the maximum node degree is 6, we generate a random graph from the Erdos-Renyi model. To enforce the maximum node degree constraint, we simply regenerate the adjacency matrix if there are any degrees larger than 6, which does not require many attempts since the graphs are very sparse under this setting. For the maximum node degree of 10, we assign 10 edges

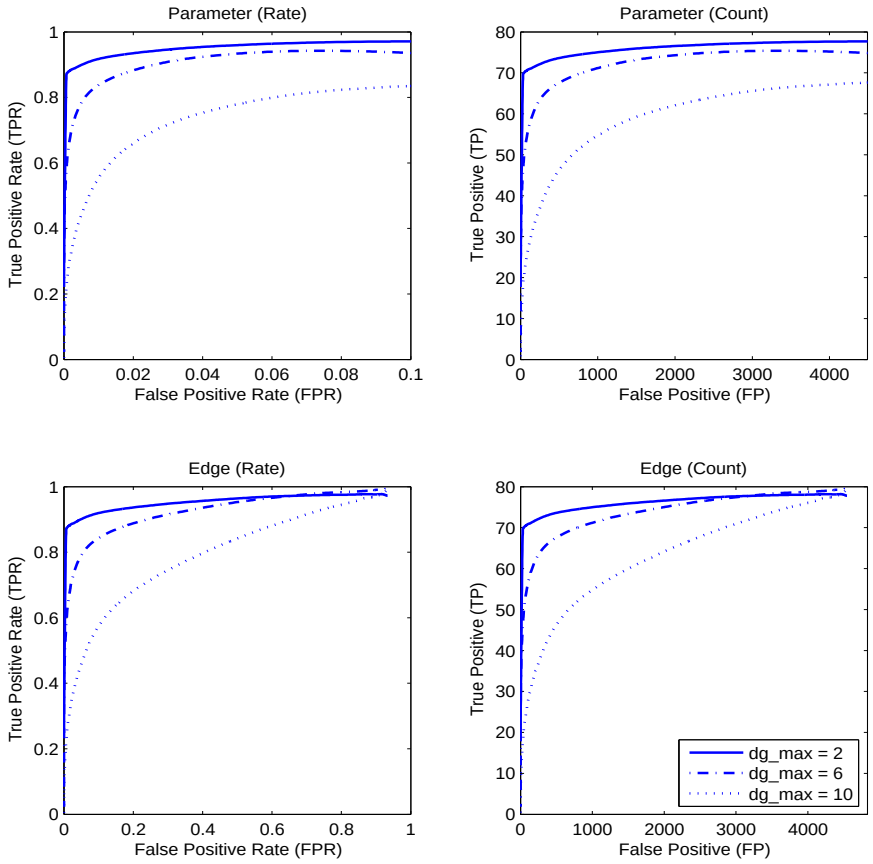

Figure 2.2: Upper left: percentage-based ROC curves for parameter identification; upper right: count-based ROC curves for parameter identification; lower left: percentage-based ROC curves for edge identification; lower right: count-based ROC curves for edge identification. The maximum node degree varies in $\{2, 6, 10\}$, and the total number of edges is fixed at 80. The variable dimensions are $p = 90$, $q = 10$; the sample size is $n = 100$; the curves are averaged over 20 replications.

to the first node and randomly generate the other 70 edges from the Erdos-Renyi model, thus creating a sparse graph with a hub. The first q nodes in the adjacency matrix are taken to correspond to the binary variables. Given the true graph, we set all parameters corresponding to absent edges to be 0. For the other (non-zero) parameters, we set the $\{\lambda_j, \lambda_{jk}, \eta_j\}$ to be 1 or -1 with equal probability, and the off-diagonal elements of $\{\Phi_0, \Phi_j\}$ to 2 or -2 with equal probability. The diagonal elements of Φ_0 are chosen so that $\Phi_0 + \sum_{j=1}^{q} \Phi_j z_j$ is positive definite for all possible

z's. We then generate the discrete variables $\boldsymbol{z}_i$'s based on 2^q probabilities given by P_z in (2.2). Since we use the exact discrete probabilities rather than MCMC methods to generate the binary data, the storage requirements prohibits taking a very large q in simulations; however, this does not affect real data, and the method works well with large q in Section 2.4. Finally, for each $\boldsymbol{z}_i$ we generate the continuous variables $\boldsymbol{y}_i$ from a multivariate Gaussian distribution with mean $\xi_{\boldsymbol{z}_i}$ and covariance $\Sigma_{\boldsymbol{z}_i}$ defined by (2.2).

The results in Figure 2.2 contain four ROC curves for both parameters and edges across a fine grid of the tuning parameters ρ, recorded both in percentage terms (FPR and TPR) and also in terms of counts (TP and FP), to get a better sense of the results on the real scale of the problem. The cut-off point for the parameter rate-based FPR is chosen at the point after which the curve does not change much, and the range of count-based FP is chosen to approximately match the range of FPR. The results show that as the maximum node degree increases, the model selection performance deteriorates, even though the total number of edges remains fixed.

Next, we vary the degree of sparsity in the true graphs. The variable dimensions are again fixed at $p = 90$, $q = 10$, and the sample size at $n = 100$. The number of edges is set to either 60, 80, or 100, while the degree of all nodes is at most 3. The true graph is again generated from the Erdos-Renyi model with the specified number of edges. The results are shown in Figure 3.1. Even though the underlying graphs are getting more dense as the number of edges increases, as long as the maximum degree is controlled, the percentage-based ROC curves for both edges and parameters stay nearly the same. The count-based ROC curves are not directly comparable in this case due to the varying number of parameters and edges of the true model.

2.3.2 Comparison with alternative penalized regressions

In this simulation, we compare our proposed method (denoted by weighted ℓ_1) with two other penalized regression approaches using the edge ROC curves. The

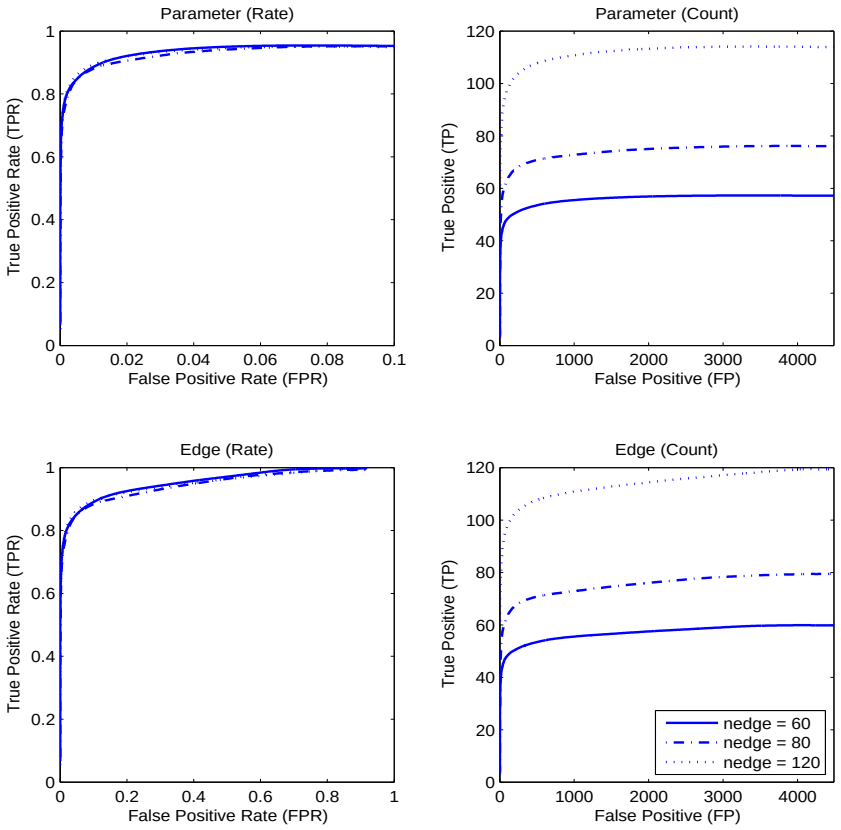

Figure 2.3: Upper left: percentage-based ROC curves for parameter identification; upper right: count-based ROC curves for parameter identification; lower left: percentage-based ROC curves for edge identification; lower right: count-based ROC curves for edge identification. The number of edges varies in $\{60, 80, 100\}$, and the maximum node degree is at most 3. The variable dimensions are $p = 90$, $q = 10$; the sample size is $n = 100$; the curves are averaged over 20 replications.

first alternative (denoted by simple ℓ_1), discussed by *Fellinghauer et al.* (2011), is to fit separate ℓ_1 regularized regressions by regressing each variable on the others, without including any of the interaction terms. The second alternative (denoted by L_1_regular) we consider is to replace the penalty in our method with the regular ℓ_1 penalty, ignoring the grouping patterns and the overlaps between groups. We consider two settings relevant for this comparison: true graphs with and without complete subgraphs.

2.3.2.1 Graphs without complete subgraphs

When the underlying graph does not contain any complete subgraphs, which can easily happen if it is very sparse, all the overlapping parameters $\Phi_j^{\gamma\mu}$ are zero. From (2.6), we can see that each edge is then represented by a unique parameter: the edge corresponding to (Z_j, Y_γ) is determined by η_j^γ; the edge for (Z_j, Z_k) is determined by λ_{jk}; and the edge for (Y_γ, Y_μ) is determined by $\Phi_0^{\gamma\mu}$. Then all the interaction terms in regressions (2.4) and (2.5) vanish. We follow the set-up of the first simulation, where the maximum node degree takes values in $\{2, 6, 10\}$, the total number of edges is fixed at 80, the dimensions are $p = 90, q = 10$, and $n = 100$. If a true graph we generate contains a complete subgraph, we discard it and generate a new one. Figure 2.4 shows that in each subplot, simple ℓ_1 and weighted ℓ_1 perform similarly, and they both outperform the regular ℓ_1. The results are to be expected, because for a true model with no interaction terms the simple ℓ_1, which excludes interaction terms automatically, should perform the best. Our weighted ℓ_1 approach penalizes the interaction terms twice as much as the other parameters, which allows it to achieve a similarly good performance. The generic regular ℓ_1 penalty fails to capture the group structure among the parameters and performs noticeably worse than the other two methods. We can conclude from this that if the underlying model is believed to be very sparse, simple ℓ_1 does well by not including the interaction terms; our method weighted ℓ_1 does equally well even with the ineffective interactions. The generic regular ℓ_1, which ignores the group structure, is the least accurate of the three alternatives.

2.3.2.2 Graphs with complete subgraphs

In this part, we study the case where the true graph contains fully connected dense subgraphs. Specifically, we set $p = 40$ and $q = 10$, make the first 20 nodes completely connected, and the other nodes are not connected to anything. This gives

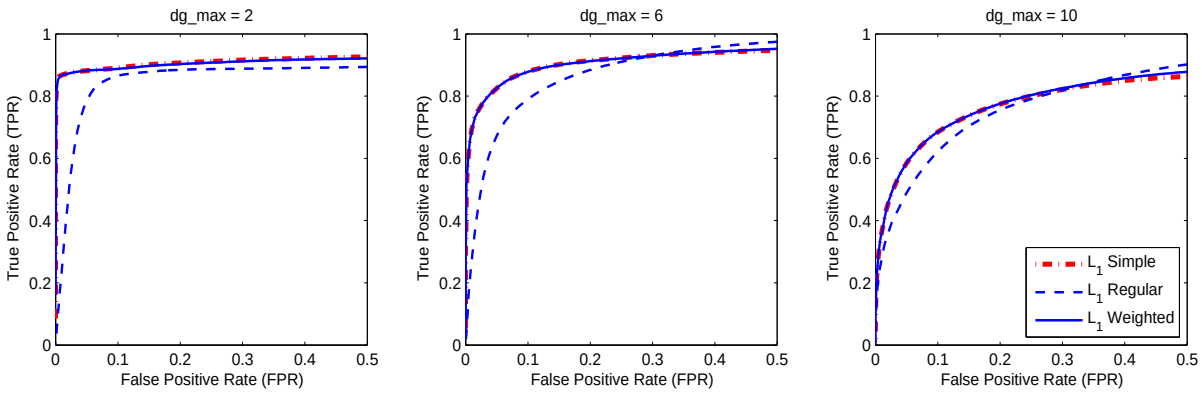

Figure 2.4: Edge-based ROC curves for sparse graphs without complete subgraphs. Blue solid: weighted ℓ_1; red dash-dot: simple ℓ_1; blue dash: regular ℓ_1. The maximum node degree is 2 (left), 6 (middle), or 10 (right), the total number of edges is 80. The variable dimensions are $p = 90$, $q = 10$; sample size is $n = 100$; the curves are averaged over 20 data replications.

approximately 190 edges and 650 non-zero parameters. To obtain comparable signal-to-noise ratios and comparable ROC curves in this setting, we adjust the sample size to $n = 200$ and decrease the non-zero parameter values by a factor of 10. For graphs with complete subgraphs, there are non-zero parameters $\Phi_j^{\gamma\mu}$ in overlapping groups, which implies that the true model includes some interaction terms in (2.4) and (2.5).

Further, we consider two types of true models compatible with the graph structure above for regressions: in model I, both main and interaction effects in (2.4) and (2.5) are non-zero, and model II has main effects only and all the interaction effects are zero. Figure 2.5 shows the results. When interaction terms are present in the true model, the regular ℓ_1 penalty and our weighed penalty perform similarly, and better than the simple ℓ_1 penalty, as one would expect. When only main effects are present in the true model, all three methods show similar performance with the weighted and simple ℓ_1 approaches having a slight advantage for reasons similar to those in Section 2.3.2.1.

Taken together, the simulation results indicate that the weighted penalty approach does a good job balancing the need for penalizing interactions more than the main

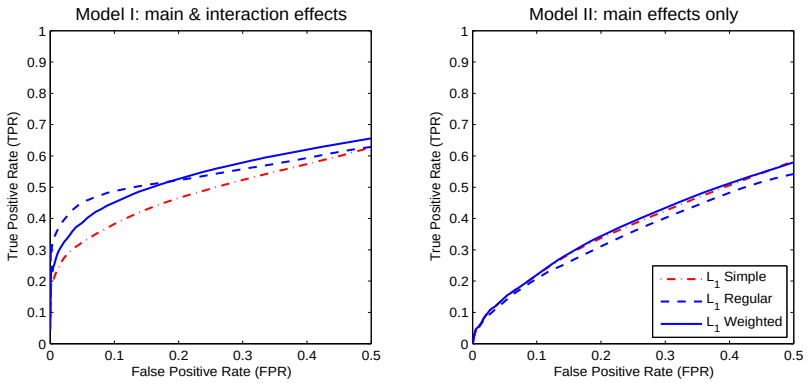

Figure 2.5: Edge-based ROC curves for graphs with complete subgraphs. Blue solid: L_1_weighted; red dash-dot: L_1_simple; blue dash: L_1_regular. Left: both main and interaction effects present, right: main effects only. The variable dimensions are $p = 40$, $q = 10$; sample size is $n = 200$; the curves are averaged over 20 data replications.

effects because of double-counting, and the need to keep them in the model. If we have no prior information on whether to expect complete subgraphs in the network or not, the weighted ℓ_1 penalty appears to be the safest choice.

2.4 Application to music annotation data

Music annotation has been studied by researchers in many areas, including audio signal processing, information retrieval, multi-label classification, and others. Music annotation data sets usually consist of two parts: "labels", typically assigned by human experts, contain the categorical semantic descriptions of the piece of music (emotions, genre, vocal type, etc.); and "features", continuous variables extracted from the time series of the audio signal itself using well developed signal processing methods. Representing these mixed variables by a Markov graph would allow us to understand how these different types of variables are associated with each other. For example, one can ask which rhythm and timbre features are associated with certain genres. We apply our method to the publicly available music data set CAL500 (*Turnbull et al.*, 2008), from the Mulan database (*Tsoumakas et al.*, 2011), to find

the conditional dependence patterns among the mixed variables.

The CAL500 dataset consists of 502 popular western music tracks (including both English language songs and instrumental music) composed within the last 55 years by 502 different artists. The collection covers a large range of acoustic variations and music genres, and the labeling of each song is supervised by at least three individuals. For each song, the label part includes a semantic vocabulary of 149 tags represented by a 149-dimensional binary vector indicating the presence of each annotation. Specifically, the labels are partitioned into the following six categories: emotions (36 total), genres (31), instruments (24), song characteristics (27), usages (15), and vocal types (16). The continuous features of the music are based on the short time Fourier transform (STFT) and are calculated for each short time window by sliding a half-overlapping, 23ms time window over the song's digital audio file. Detailed description of the feature extraction procedure can be found in *Tzanetakis and Cook* (2002). For each analysis window of 23ms, the following continuous features are extracted to represent the audio file: the *spectral centroid*, a measure of 'brightness' of the music texture with higher value indicating brighter music with more high frequencies; *spectral flux*, a measure of the amount of local spectral change; *zero crossings*, a measure of the noisiness of the signal; and the first MFCC coefficient (*Logan*, 2000) representing the amplitude of the music, which comes from a two-step transformation designed to capture the spectral structure. Every consecutive 512 of the 23ms short frames are then grouped into 1s long texture windows, based on which the following summary statistics for the four features defined above were calculated and used as the final continuous part of the data: overall mean, mean of the standard deviations of each texture window, standard deviation of the means of each texture window, and standard deviation of the standard deviations of each texture window.

In our analysis, we omitted labels which were assigned to less than 3% of the songs and kept only the first MFCC coefficient since it can be interpreted as the overall

amplitude of the audio signal, and other coefficients are not readily interpretable. Also, we standardized the continuous variables. This resulted in a dataset with $n = 502$ observations, $q = 118$ discrete variables, and $p = 16$ continuous variables.

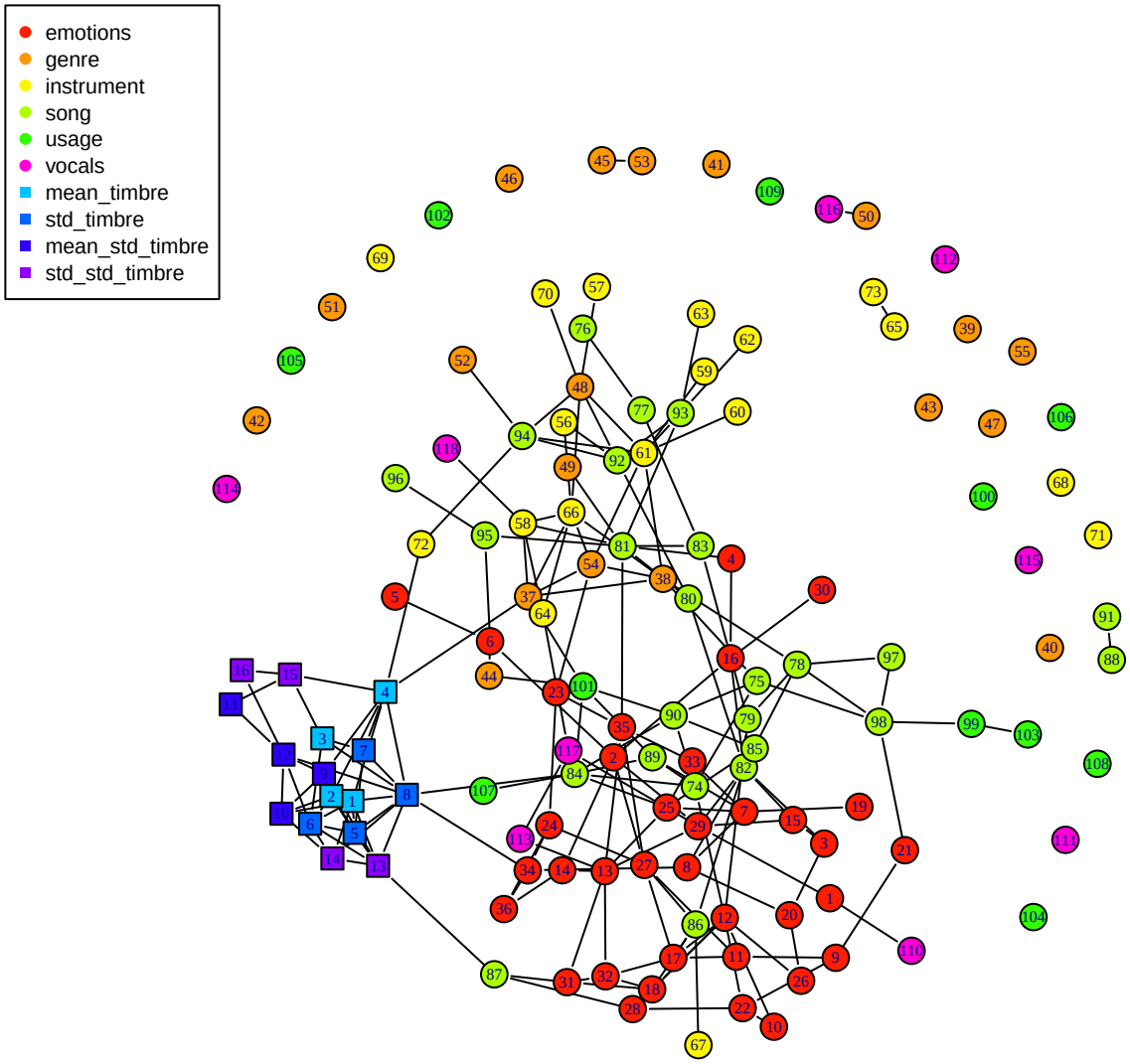

Figure 2.6: Estimated graphical model for CAL500 music data (edges with stability selection frequency of at least 0.9).

We applied our method coupled with stability selection (*Meinshausen and Bühlmann*, 2010) to identify the underlying Markov graph for the purposes of exploratory data analysis, which is the primary usage of graphical models. To perform stability selection, we run our algorithm 100 times on randomly drawn sub-samples of size $n/2$,

and kept only the edges that were selected at least 90 times. The results are shown in Figure 2.6. The continuous timbre features are represented by squares labeled 1-16 and the binary variables are represented by circles labeled 1-118. Each color represents a category of variables as shown in the legend. The graph has a number of interesting and intuitive connections. The continuous variables that represent the audio signal features are quite densely connected within themselves, which is to be expected. More interesting edges can be found between continuous features and expert labels. The average amplitude of the music (square 4) is connected with the genre "alternative rock" (circle 37) and the instrument "Synthesizer" (circle 72). The noisiness of the music (square 13) is associated with "negative feelings" (circle 87), which seems reasonable and has also been reported by *Blumstein et al.* (2012). We also find connections between short period amplitude variation (square 8) with popular likeable songs (circle 84) and not tender or soft emotions (circle 34). Moreover, there are some interesting patterns within the group of binary labels, which allow us to infer connections between different emotions, genres, instruments, usages and so on. For example, songs with positive feelings (circle 86) are connected to piano (circle 67), happy and not happy emotions (circle 17 and 18, highly negatively correlated), songs with high energy (circle 82) and optimistic emotions (circle 27). We also find edges connecting fast and not fast tempo music (circles 78 and 79) to classic rock (circle 38), songs with high energy (circle 82) and very danceable and not danceable songs (circles 97 and 98). Likeable or popular songs (circle 84) are associated with usages such as driving (circle 101) and reading (circle 107), which makes intuitive sense.

2.5 Extension to general discrete data

To extend our model to the general case where the discrete variables can take more than two values, we modify the previous model (4.1) into the following,

$$
\begin{aligned}
\log f(z,y) &= \sum_{d:d\subseteq\Delta,|d|\le 2} \lambda_d(z) + \sum_{d:d\subseteq\Delta,|d|\le 1} \eta_d(z)^T y - \frac{1}{2}\sum_{d:d\subseteq\Delta,|d|\le 1} y^T\Phi_d(z)y \ , \\
&= \left(\lambda_0 + \sum_{j=1}^{q}\lambda_j(z_j) + \sum_{j>k}\lambda_{jk}(z_j,z_k)\right) + \sum_{\gamma=1}^{p}\left(\eta_0^\gamma + \sum_{j=1}^{q}\eta_j^\gamma(z_j)\right) y_\gamma \ , \\
&\quad -\frac{1}{2}\sum_{\gamma,\mu=1}^{p}\left(\Phi_0^{\gamma\mu} + \sum_{j=1}^{q}\Phi_j^{\gamma\mu}(z_j)\right) y_\gamma y_\mu \ , \qquad (2.11)
\end{aligned}
$$

where each z_j takes integer values 1 to K_j; $\lambda_j(\cdot)$, $\eta_j^\gamma(\cdot)$, $\Phi_j^{\gamma\mu}(\cdot)$ are all discrete functions which take on K_j possible values and $\lambda_{jk}(\cdot,\cdot)$ is a discrete function with $K_j\times K_k$ values. For identifiability, we set $\lambda_j(1)=0$, $\eta_j^\gamma(1)=0$, $\Phi_j^{\gamma\mu}(1)=0$ and $\lambda_{jk}(1,\cdot)=\lambda_{jk}(\cdot,1)=0$. The correspondence between the parameters and the edges is then given by

$$
\begin{aligned}
Z_j \perp Z_k \mid X\backslash\{Z_j,Z_k\} &\iff \boldsymbol{\theta}_{jk} = (\lambda_{jk}(z_j,z_k)) = 0 \ , \\
Z_j \perp Y_\gamma \mid X\backslash\{Z_j,Y_\gamma\} &\iff \boldsymbol{\theta}_{j\gamma} = \left(\eta_j^\gamma(z_j), \{\Phi_j^{\gamma\mu}(z_j) : \mu\in\Gamma\backslash\{\gamma\}\}\right) = 0 \ , \\
Y_\gamma \perp Y_\mu \mid X\backslash\{Y_\gamma,Y_\mu\} &\iff \boldsymbol{\theta}_{\gamma\mu} = \left(\Phi_0^{\gamma\mu}, \{\Phi_j^{\gamma\mu}(z_j) : j\in\Delta\}\right) = 0 \ . \qquad (2.12)
\end{aligned}
$$

The generalized model can be fitted with separate regressions based on the conditional likelihood of each variable. The parameters in (2.12) still have group structure, which calls for using the group lasso penalty as in (2.7) and (2.8). The overlapping structure is more complex in this case, and we use the upper bound ℓ_1 approximation as in (2.9) and (2.10) to obtain the final estimates. Specifically, we minimize the following criteria separately:

Logistic regression with ℓ_1 penalty: for $j=1,\ldots,q$

$$
\min \ell_j + \rho\left(\sum_{k\neq j}\sum_{(z_j,z_k)}|\lambda_{jk}(z_j,z_k)| + \sum_{\gamma=1}^{p}\sum_{z_j=1}^{K_j}|\eta_j^\gamma(z_j)| + 2\sum_{\gamma<\mu}\sum_{z_j=1}^{K_j}|\Phi_j^{\gamma\mu}(z_j)|\right).
$$

Linear regression with ℓ_1 penalty: for $\gamma = 1, \ldots, p$

$$\min \ \ell_\gamma + \rho \left(\sum_{j=1}^{q} \sum_{z_j=1}^{K_j} |\tilde{\eta}_j^\gamma(z_j)| + \sum_{\mu \neq \gamma} |\tilde{\Phi}_0^{\gamma\mu}| + 2 \sum_{j=1}^{q} \sum_{\mu \neq \gamma} \sum_{z_j=1}^{K_j} |\tilde{\Phi}_j^{\gamma\mu}(z_j)| \right).$$

2.6 Discussion

We have proposed a new graphical model for mixed (continuous and discrete) data, which is particularly suitable for high-dimensional data. While the general conditional Gaussian model goes back to *Lauritzen and Wermuth* (1989), it is not appropriate for high-dimensional data, and there is little previous work on mixed graphical models that can scale to modern applications. While our model is substantially simpler than the original general model, it scales much better. Given that graphical models are primarily a tool for exploratory data analysis, this is a reasonable trade-off, and the ability to explore conditional dependence relationships between large numbers of discrete and continuous variables will hopefully be of use to practitioners in a range of application domains.

We have recently become aware of two new developments on this topic (which were derived in parallel with and independently of this manuscript). *Lee and Hastie* (2012) assume a more restricted version of the conditional Gaussian density by assuming constant conditional covariance for all the continuous variables. Their model can be viewed as a special case of ours in (4.1), where all the Φ_j are 0, which can be too restrictive for some applications. They considered both the maximum likelihood approach and maximum pseudo likelihood approach with a group lasso penalty for general discrete variables; when the discrete variables are all binary, this is equivalent to using the regular ℓ_1 penalty, which is one of the alternatives we compared in simulations in Section 2.3.2. *Lee and Hastie* (2012) did not focus on high-dimensional applications. Another recent paper, *Fellinghauer et al.* (2011), applied random forests and stability selection in fitting ℓ_1-regularized regressions of each variable on the rest,

without specifying any generative models for the mixed data. In our framework, this would amount to taking the separate regression approach to a simplified conditional Gaussian model as in *Lee and Hastie* (2012), except with regression fitted using random forests coupled with stability selection. As our simulations show, such an approach performs well when the true graph is very sparse and has no complete subgraphs; when the true graph has complete subgraphs, our method is expected to outperform a separate regression approach.

CHAPTER III

Sparse Ising Model with Covariates

3.1 Introduction

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.* (2010a,b). 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 and 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 chapter is organized as follows. In Section 3.2, we describe the conditional Ising model with covariates, and two estimation procedures for fitting it. Section 3.3 establishes asymptotic properties of the proposed estimation method. We evaluate the performance of our method on simulated data in Section 3.4, and apply it to a dataset on genomic instability in breast cancer samples in Section 3.5. Section 3.6 concludes with a summary and discussion.

3.2 Conditional Ising Model with Covariates

3.2.1 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

$$\theta_{jk} = 0 \Longleftrightarrow y_j \perp y_k \parallel \boldsymbol{y}_{\setminus(j,k)}, \quad \forall j \neq k, \tag{3.1}$$

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 $\mathcal{D}_n = \{(\boldsymbol{x}^1, \boldsymbol{y}^1), \ldots, (\boldsymbol{x}^n, \boldsymbol{y}^n)\}$ 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

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{\theta}(\boldsymbol{x}))} \exp\left(\sum_{j=1}^{q} \boldsymbol{\theta}_{jj}(\boldsymbol{x}) y_j + \sum_{(j,k):1\leq k<j\leq q} \boldsymbol{\theta}_{jk}(\boldsymbol{x}) y_j y_k\right). \tag{3.2}$$

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,

$$\log\left(\frac{P(y_j=1|\boldsymbol{y}_{\setminus j},\boldsymbol{x})}{1-P(y_j=1|\boldsymbol{y}_{\setminus j},\boldsymbol{x})}\right)=\boldsymbol{\theta}_{jj}(\boldsymbol{x})+\sum_{k:k\neq j}\boldsymbol{\theta}_{jk}(\boldsymbol{x})y_k, \tag{3.3}$$

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

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

Similarly to (3.1), 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

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

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

$$P_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})=\frac{1}{Z(\boldsymbol{\theta}(\boldsymbol{x}))}\exp\left(\sum_{j=1}^{q}(\theta_{jj0}+\boldsymbol{\theta}_{jj}^T\boldsymbol{x})y_j+\sum_{k>j}(\theta_{jk0}+\boldsymbol{\theta}_{jk}^T\boldsymbol{x})y_jy_k\right). \tag{3.4}$$

Instead of (3.3), we now have the log-odds that depend on the covariates, through

$$\log\left(\frac{P(y_j=1|\boldsymbol{y}_{\setminus j},\boldsymbol{x})}{1-P(y_j=1|\boldsymbol{y}_{\setminus j},\boldsymbol{x})}\right)=\theta_{jj0}+\boldsymbol{\theta}_{jj}^T\boldsymbol{x}+\sum_{k:k\neq j}(\theta_{jk0}+\boldsymbol{\theta}_{jk}^T\boldsymbol{x})y_k. \tag{3.5}$$

The choice of linear parametrization for $\boldsymbol{\theta}_{jk}(\boldsymbol{x})$ has several advantages. First, (3.5) 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 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 $\boldsymbol{x}$ 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.

3.2.2 Fitting the Model

The probability model $P_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})$ in (4.1) 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, (3.5) 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 $\boldsymbol{y}_{\setminus j}^i$ 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

$$\ell_j(\boldsymbol{\theta}; \mathcal{D}_n) = -\frac{1}{n}\sum_{i=1}^{n} \log P(y_j^i|\boldsymbol{x}^i, \boldsymbol{y}_{\setminus j}^i) = -\frac{1}{n}\sum_{i=1}^{n} \left(\log(1+e^{\eta_j^i}) - y_j^i \eta_j^i\right), \tag{3.6}$$

where

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

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 (3.6).

3.2.2.1 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 $\hat{\theta}^0_{jk\ell}$ denote the initial estimate from the separate regularized logistic regressions. Then for any $1 \le j < k \le q$ and any $l = 0, \ldots, p$, we can use one of the two symmetrizing approaches:

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

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

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 3.4.

3.2.2.2 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,

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

where $\boldsymbol{\theta}_{\setminus 0} = \boldsymbol{\theta}\setminus\{\theta_{110}, \theta_{220}, \ldots, \theta_{qq0}\}$. 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.

3.3 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 $\boldsymbol{x}$, both our assumptions and conclusions are different.

In this analysis, we treat the covariates $\boldsymbol{x}_i$'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

$$\begin{aligned} \boldsymbol{I}_j^* &= \mathbb{E}_{\theta^*}(\nabla^2 \log P_{\boldsymbol{\theta}}(y_j|\boldsymbol{x}, \boldsymbol{y}_{\setminus j})) \qquad (3.7) \\ &= \mathbb{E}_{\theta^*}\left(p_j(1-p_j)(\boldsymbol{x} \otimes \boldsymbol{y}_{\setminus j})(\boldsymbol{x} \otimes \boldsymbol{y}_{\setminus j})^T\right) \quad \text{(Information matrix)} \\ \boldsymbol{U}_j^* &= \mathbb{E}_{\theta^*}\left((\boldsymbol{x} \otimes \boldsymbol{y}_{\setminus j})(\boldsymbol{x} \otimes \boldsymbol{y}_{\setminus j})^T\right) \qquad (3.8) \end{aligned}$$

where

$$\begin{aligned} p_j &= p_j(\boldsymbol{x}, \boldsymbol{y}_{\setminus j}) = P_{\boldsymbol{\theta}^*}(y_j = 1|\boldsymbol{x}, \boldsymbol{y}_{\setminus j}) \ , \\ \boldsymbol{x} \otimes \boldsymbol{y}_{\setminus j} &= (1, x_1, \ldots, x_p)^T \otimes (y_1, \ldots, y_{j-1}, 1, y_{j+1}, \ldots, y_q)^T \setminus \{1\} \ . \end{aligned}$$

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 compliment set of $\mathcal{S}_j$. Moreover, for any matrix A, let $\|A\|_\infty = \max_i \sum_j |A_{ij}|$ be the matrix L_∞ 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

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

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

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

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

$$\min_{\boldsymbol{\theta}_j} \quad -\ell_j(\boldsymbol{\theta}_j; \mathcal{D}_n) + \lambda_n \|\boldsymbol{\theta}_j\|_1. \tag{3.9}$$

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

$$P(\|\boldsymbol{x}\|_\infty \geq M) \leq \exp(-M^\delta), \qquad \textit{for all} \;\; M \geq M_0 > 0, \tag{3.10}$$

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

$$M_n \geq (C\lambda_n^2 n)^{\frac{1}{1+\delta}} \;, \tag{3.11}$$

$$\lambda_n \geq C M_n \sqrt{\frac{\log p + \log q}{n}} \;, \tag{3.12}$$

$$n \geq C M_n^2 d^3 (\log p + \log q) \;, \tag{3.13}$$

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 III.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 $\boldsymbol{x}$ 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.

3.4 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 and 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 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 $P_{\boldsymbol{\theta}}(y_j^i = 1 | \boldsymbol{y}_{\setminus j}^i, \boldsymbol{x}^i)$ 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 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.

3.4.1 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 3.1.

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

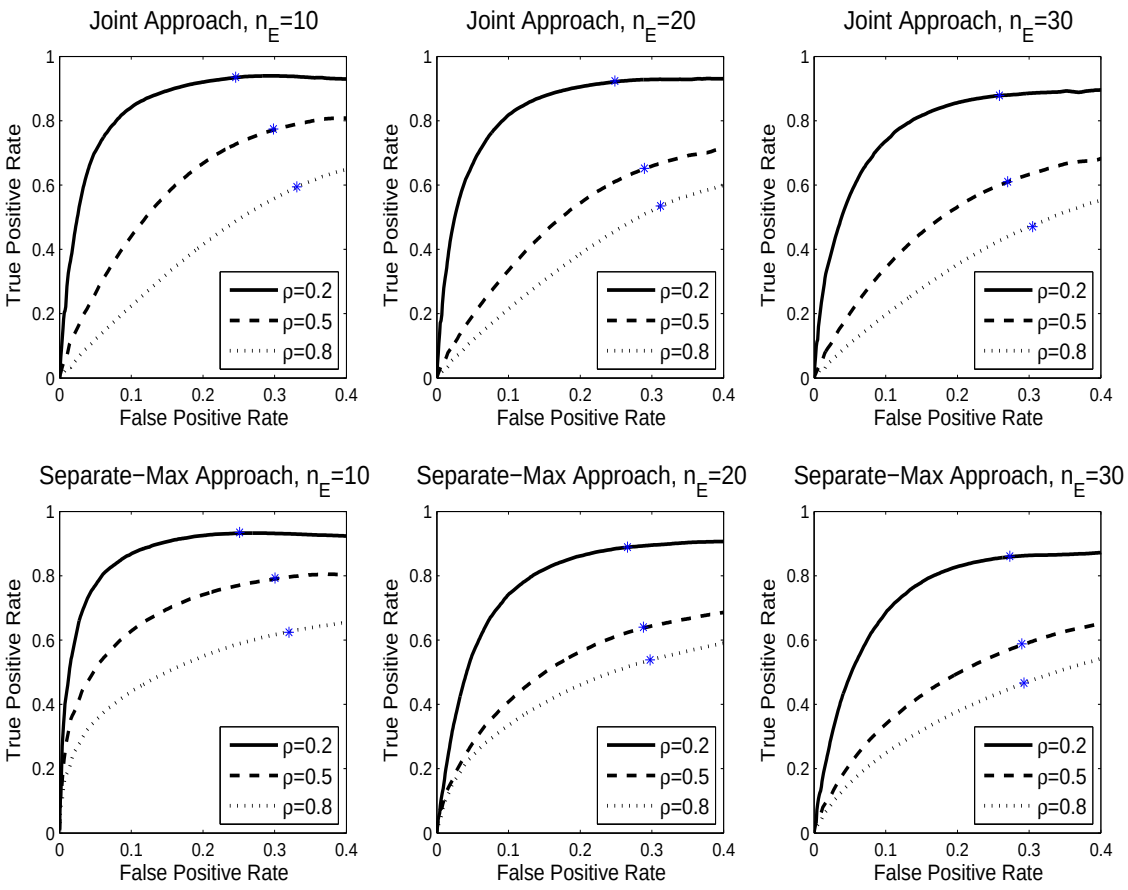

Figure 3.1: ROC curves for varying levels of sparsity, as measured by the number of edges (n_E) and expected proportion of non-zero covariates (ρ). The star on each curve corresponds to an optimal value of λ selected on an independent validation set.

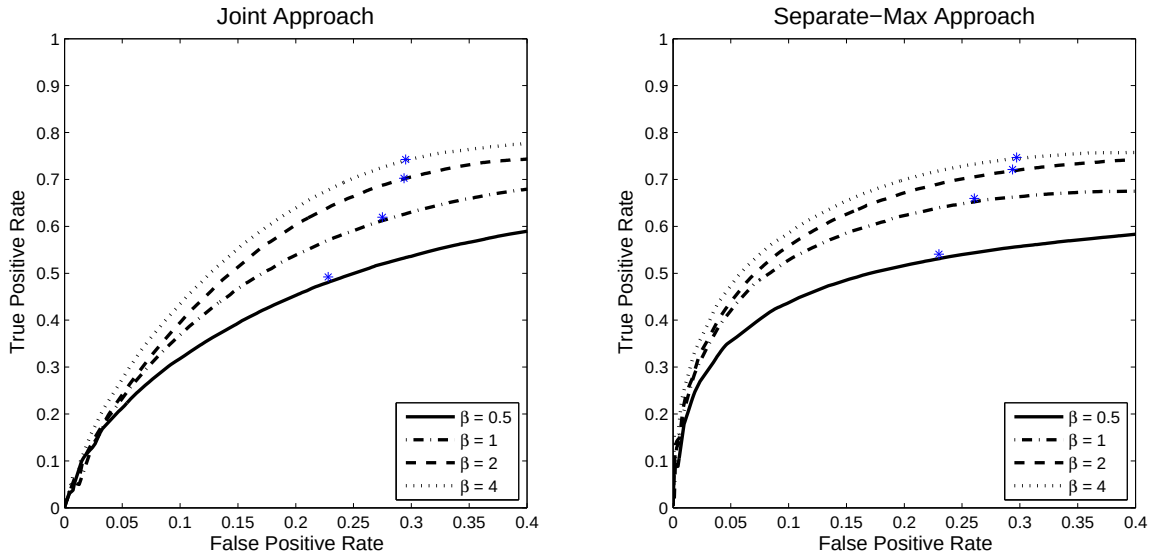

Figure 3.2: ROC curves for varying levels of signal strength, as measured by the parameter β. The star on each curve corresponds to an optimal value of λ selected on an independent validation set.

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" λ, selected by maximizing the conditional log-likelihood on an independent validation dataset of the same size as the training data.

3.4.2 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 3.2.

As the signal strength β increases, both the separate and the joint methods show improved 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.

3.4.3 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_{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 3.3. 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.

3.5 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

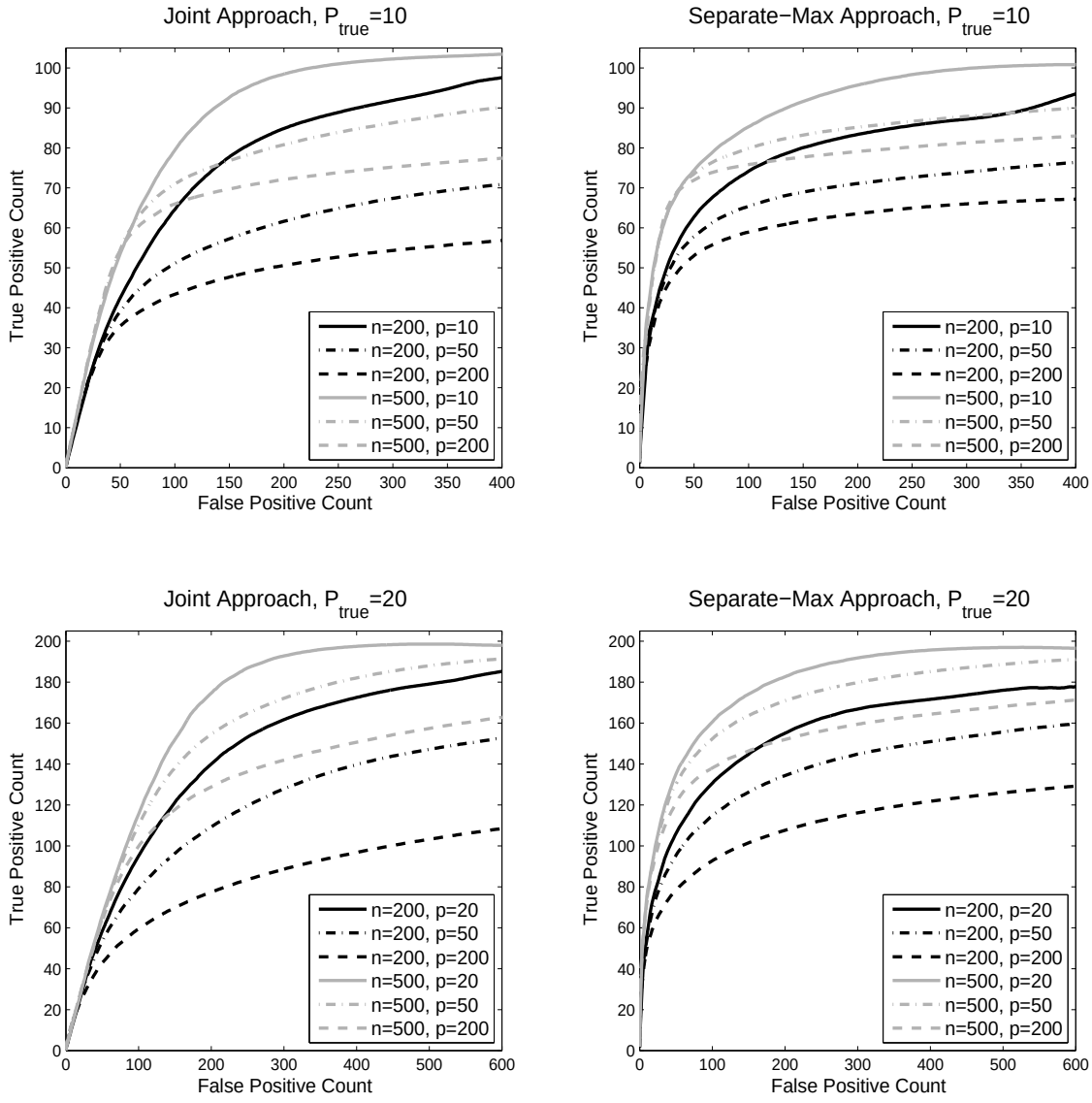

Figure 3.3: ROC curves for varying dimension, number of noise covariates, and sample size.

.

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 $\boldsymbol{y}_{143\times 430}$, where y_j^i indicates the deletion status of the j^{th} cytoband in the i^{th} sample. Let $\boldsymbol{x}^i$ denote the covariate vector containing the three clinical phenotypes of the i^{th} sample, and $\boldsymbol{x}_l$ 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-max fitting method. We then apply stability selection

(*Meinshausen and 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 λ from a fixed grid of values, we record the frequency of $\hat{\theta}_{jkl}$ being non-zero respectively for each covariate $\boldsymbol{x}_l$, $l = 0, 1, 2, 3$ on all pairs of (j, k), $1 \leq j < k \leq 430$, and denote it by $f_{jkl}(\lambda)$. Note that $\boldsymbol{x}_0$ corresponds to the main effect interaction between a pair of $\boldsymbol{y}_j$'s and does not involve any covariates. Then we use $f^*_{jkl} = \max_\lambda f_{jkl}(\lambda)$ as a measure of importance of covariate $\boldsymbol{x}_l$ for the edge (j, k). Finally, for each covariate $\boldsymbol{x}_j$, 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 $\boldsymbol{x}_j$ 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 3.1). 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 $d^m_{j,l} = \#\{k : \hat{\theta}_{jkl} \neq 0\}$. A ranking of nodes can then be produced for each covariate l and each replication m, with $r^m_{j,l}$ being the corresponding rank. Finally, we compute the median rank across all stability selection subsamples $r_{j,l} = \text{median}\{r^m_{j,l}, m = 1, \ldots, 100\}$, and order nodes by rank for each covariate. The results are listed in Table 3.2. Interestingly, cytoband 8p11.22 was ranked close to the top for

Table 3.1: 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	Tumor stage		
1q31.1	22q11.21	0.76	3q26.1	11p13	0.55	Gene1	Gene2	Freq
2q32.1	6q14.1	0.76	4q35.2	22q12.3	0.55	16q23.3	17p13.1	0.61
9q21.11	16q21	0.76	15q22.33	17p11.2	0.55	12p11.23	16q12.2	0.59
9q31.3	14q24.3	0.76	3p22.1	6p21.31	0.54	3q13.13	Xq23	0.57
10q25.3	12p13.31	0.76	4q28.2	7q21.13	0.54	7p21.3	12p11.23	0.56
4q35.1	15q22.2	0.75	5q13.1	6q22.33	0.54	9q34.13	15q21.1	0.55
3p21.31	17p11.2	0.74	5q23.2	8p21.2	0.54	11q24.2	13q32.3	0.55
6p21.32	13q31.2	0.74	16q22.1	17q21.31	0.54	8q21.13	13q33.1	0.54
10q11.21	12p13.32	0.74	4q28.3	9p21.3	0.53	2p21	12p13.31	0.53
9q33.1	14q12	0.73	4q35.1	9p21.3	0.53	10q26.3	17p11.2	0.53
12p13.31	17q11.2	0.73	4q35.2	16q22.1	0.53	7p21.3	12p12.1	0.51
1p34.2	3p22.1	0.72	2q31.3	4q13.2	0.52	3q13.13	7p21.3	0.5
5q33.1	11p15.4	0.72	3p26.1	14q13.1	0.52	9q34.13	15q22.1	0.5
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			

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 3.1, 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; *Devilce 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.

3.6 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 only the strength of the links is af-

Table 3.2: 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

fected. 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.

3.7 Appendix

3.7.1 Proof of Theorem III.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 $\boldsymbol{I}^n$ and $\boldsymbol{U}^n$, the sample versions of of $\boldsymbol{I}^*$ and $\boldsymbol{U}^*$ defined in (3.7) (Proposition III.2). Then we show that if **A1** and **A2** hold for the population versions $\boldsymbol{I}^*$ and $\boldsymbol{U}^*$, they also hold for $\boldsymbol{I}^n$ and $\boldsymbol{U}^n$ with high probability (Proposition III.7). The sample quantities $\boldsymbol{I}^n$ and $\boldsymbol{U}^n$ are defined as

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

Proposition III.2. *If* $\boldsymbol{A1}$ *and* $\boldsymbol{A2}$ *are satisfied by* $\boldsymbol{I}^n$ *and* $\boldsymbol{U}^n$*, assume moreover that*

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

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

Proof of Proposition III.2. The proof requires several steps. The uniqueness part follows directly from the following lemma:

Lemma III.3. (Shared sparsity and uniqueness of $\hat{\boldsymbol{\theta}}$, *Ravikumar et al.* (2010)). *Define the sign vector* $\boldsymbol{t}$ *for* $\boldsymbol{\theta}$ *to satisfy the following properties,*

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

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

We now proceed to prove the rest of Proposition III.2. For $\hat{\boldsymbol{\theta}}$ to be a solution of (3.9), the sub-gradient at $\hat{\boldsymbol{\theta}}$ must be 0, i.e.,

$$\nabla\ell(\hat{\boldsymbol{\theta}}, \mathcal{D}_n) + \lambda_n\hat{\boldsymbol{t}} = 0 \ . \tag{3.14}$$

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

$$W^n = -\nabla\ell\left(\boldsymbol{\theta}^*, \mathcal{D}_n\right) = \frac{1}{n}\sum_{i=1}^n(\boldsymbol{x}^i \otimes \boldsymbol{y}^i_{\setminus j})(y^i_j - p^i_j(\boldsymbol{\theta}^*)) \ .$$

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

$$\boldsymbol{I}^n \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \right) = W^n - \lambda_n \hat{\boldsymbol{t}} + R^n \ . \tag{3.15}$$

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

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

$$\min_{(\hat{\boldsymbol{\theta}}_{\mathcal{S}}, 0)} \ell(\hat{\boldsymbol{\theta}}, \mathcal{D}_n) + \lambda_n \|\hat{\boldsymbol{\theta}}_{\mathcal{S}}\|_1 \ , \tag{3.16}$$

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 $\hat{\boldsymbol{t}}_{\mathcal{S}}$ for $\hat{\boldsymbol{\theta}}_{\mathcal{S}}$ similarly defined as in Lemma III.3, and $\hat{\boldsymbol{t}}_{\mathcal{S}^C} = -\frac{1}{\lambda_n} \nabla_{\mathcal{S}^C} \ell(\hat{\boldsymbol{\theta}}_{\mathcal{S}}, \mathcal{D}_n)$ as obtained in (3.14). Now we need to show that with high probability,

$$\|\hat{\boldsymbol{t}}_j\|_\infty \ < \ 1, \qquad \text{for} \ \ j \in \mathcal{S}^C \tag{3.17}$$

$$\hat{\boldsymbol{t}}_j \ = \ sign(\boldsymbol{\theta}_j^*), \ \ \text{for} \ \ j \in \mathcal{S} \text{ and } \|\theta_j^*\| \geq \frac{10 \lambda_n \sqrt{d}}{\Delta_{\min}} \tag{3.18}$$

The following three lemmas form the proof.

Lemma III.4. (Control the remainder term W^n). *For $\alpha \in (0, 1]$, assume $\|\boldsymbol{x}\|_\infty \leq M_n$ a.s, then,*

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

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

Proof of Lemma III.4. We can write $W^n = \frac{1}{n} \sum_{i=1}^n (\boldsymbol{x}^i \otimes \boldsymbol{y}^i_{\setminus j})(y_j^i - p_j^i(\boldsymbol{\theta}^*)) = \sum_{i=1}^n Z_i$,

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 III.5. (ℓ_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 III.5. Let $G(u_{\mathcal{S}}) = \ell(\boldsymbol{\theta}_{\mathcal{S}}^* + u_{\mathcal{S}}, \mathcal{D}_n) - \ell(\boldsymbol{\theta}_{\mathcal{S}}^*, \mathcal{D}_n) + \lambda_n(\|\boldsymbol{\theta}_{\mathcal{S}}^* + u_{\mathcal{S}}\|_1 - \|\boldsymbol{\theta}_{\mathcal{S}}^*\|_1)$ be a function $G : \mathbb{R}^d \to \mathbb{R}$. It is easy to see that $G(u_{\mathcal{S}})$ is convex and it achieves its minimum at $\hat{u}_{\mathcal{S}} = \hat{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_{\mathcal{S}}^*$. Moreover, $G(0) = 0$. Thus if we can show that $G(u_{\mathcal{S}})$ is positive on the set $\|u_{\mathcal{S}}\|_2 = B$, then we will have $\hat{u}_{\mathcal{S}} \leq B$ due to convexity of $G(u_{\mathcal{S}})$. Note that

$$G(u_{\mathcal{S}}) = -W_{\mathcal{S}}^{nT}u_{\mathcal{S}} + u_{\mathcal{S}}^T\nabla^2\ell(\boldsymbol{\theta}_{\mathcal{S}}^* + \alpha u_{\mathcal{S}})u_{\mathcal{S}} + \lambda_n(\|\boldsymbol{\theta}_{\mathcal{S}}^* + u_{\mathcal{S}}\|_1 - \|\boldsymbol{\theta}_{\mathcal{S}}^*\|_1) .$$

Further,

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

Combining all of the above, we have

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

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 III.6. (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 III.6. Recall that

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

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

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

where $Z_k^i = x_l^i y_m^i$, for some (l, m). By **A1** and Lemma III.5, we have

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

$\square$

Putting all the lemmas together, we are ready to prove Proposition III.2.

Proof of Proposition III.2. Set $\lambda_n = \frac{8M_n(2-\alpha)}{\alpha}\sqrt{\frac{\log p + \log q}{n}}$. By Lemma III.4, 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 $\lambda_n d \leq \frac{\Delta_{\min}^2}{100 M_n \Delta_{\max}} \frac{\alpha}{2-\alpha}$, thus the conditions of Lemmas III.5 and III.6 hold.

By rewriting (3.15) and utilizing the fact that $\hat{\boldsymbol{\theta}}_{\mathcal{S}^C} = \boldsymbol{\theta}^*_{\mathcal{S}^C} = 0$, we have

$$\begin{aligned} \boldsymbol{I}^n_{\mathcal{S}^C\mathcal{S}}(\hat{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}^*_{\mathcal{S}}) &= W^n_{\mathcal{S}^C} - \lambda_n \hat{\boldsymbol{t}}_{\mathcal{S}^C} + R^n_{\mathcal{S}^C} \ , \qquad (3.19) \\ \boldsymbol{I}^n_{\mathcal{S}\mathcal{S}}(\hat{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}^*_{\mathcal{S}}) &= W^n_{\mathcal{S}} - \lambda_n \hat{\boldsymbol{t}}_{\mathcal{S}} + R^n_{\mathcal{S}} \ . \qquad (3.20) \end{aligned}$$

Since $\boldsymbol{I}^n_{\mathcal{S}\mathcal{S}}$ is invertible by assumption, combining (3.19) and (3.20) gives

$$\boldsymbol{I}^n_{\mathcal{S}^C\mathcal{S}}(\boldsymbol{I}^n_{\mathcal{S}\mathcal{S}})^{-1}(W^n_{\mathcal{S}} - \lambda_n \hat{\boldsymbol{t}}_{\mathcal{S}} + R^n_{\mathcal{S}}) = W^n_{\mathcal{S}^C} - \lambda_n \hat{\boldsymbol{t}}_{\mathcal{S}^C} + R^n_{\mathcal{S}^C} \ . \qquad (3.21)$$

To show (3.17), we reorganize (3.21) and use results from Lemmas III.4 and III.6:

$$\begin{aligned} \lambda_n \|\hat{\boldsymbol{t}}_{\mathcal{S}^C}\|_\infty &= \|\boldsymbol{I}^n_{\mathcal{S}^C\mathcal{S}}(\boldsymbol{I}^n_{\mathcal{S}\mathcal{S}})^{-1}(W^n_{\mathcal{S}} - \lambda_n \hat{\boldsymbol{t}}_{\mathcal{S}} + R^n_{\mathcal{S}}) - W^n_{\mathcal{S}^C} - R^n_{\mathcal{S}^C}\|_\infty \\ &\leq \|\boldsymbol{I}^n_{\mathcal{S}^C\mathcal{S}}(\boldsymbol{I}^n_{\mathcal{S}\mathcal{S}})^{-1}\|_\infty(\|W^n\|_\infty + \lambda_n + \|R^n\|_\infty) + \|W^n\|_\infty + \|R^n\|_\infty \\ &\leq \lambda_n(1 - \frac{\alpha}{2}) \ . \end{aligned}$$

To show (3.18), it suffices to show that $\|\hat{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}^*_{\mathcal{S}}\|_\infty \leq \frac{\boldsymbol{\theta}^*_{\min}}{2}$. By Lemma III.5,

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

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 III.2. □

Proposition III.7. *If $\boldsymbol{I}^*$ and $\boldsymbol{U}^*$ satisfy **A1** and **A2**, and $M_n = \sup\|\boldsymbol{x}\|_\infty < \infty$ a.s.,*

the following hold for any $\delta > 0$. A and B are some positive constants.

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

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

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

We omit the proof of Proposition III.7, which is very similar to Lemmas 5 and 6 in *Ravikumar et al.* (2010).

Proof of Theorem III.1. With Propositions III.2 and III.7, the proof of Theorem III.1 is straightforward. Given that A1 and A2 are satisfied by $\boldsymbol{I}^*$ and $\boldsymbol{U}^*$ and that conditions (3.12) and (3.13) hold, on the set $\mathcal{A} = \{\boldsymbol{x} : M_n = \sup\|\boldsymbol{x}\| < \infty\}$ the assumptions in Proposition III.7 are satisfied. Thus with probability at least $1 - \exp(-\frac{C\lambda_n^2 n}{M_n^2})$, the conditions of Proposition III.2 hold, and therefore the results in Theorem III.1 hold. Finally, let $\mathcal{T}$ stand for the set where the results of Theorem III.1 hold. Then by (3.10) and (3.11), we have

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

□

CHAPTER IV

Multi-label Classification via Ising Models

4.1 Introduction

Despite many interesting applications and connections to several other well-studied statistical problems, the statistical literature on multi-label classification is somewhat sparse. Most of the existing methods share the common approach to break down the multi-label classification into separate binary classification problems and builds a binary classifier for each label independently. *Joachims* (1998) uses a set of binary SVM classifiers and shows that using SVM as binary classifiers achieves higher accuracy than others competing methods. However he does not discuss multi-label training models or specific testing criteria. Certain modification of such separate binary SVM classifiers include *Elisseeff and Weston* (2002), *Godbole and Sarawagi* (2004), etc. *Boutell et al.* (2004) applied similar decomposed classifiers to a image scene classification problem. (*Zhang and Zhou*, 2007) proposed ML-KNN where the k-nearest neighboring instances are identified first, and based on the label sets of these neighboring instances, maximum a posteriori (MAP) principle is applied to determine the label set. The perks of constructing separate classifiers is that most of the well-established single-label classification techniques can be employed readily; however this approach fails to take into account the correlation among labels, which can be a drawback as co-occurrence of the labels is an intrinsic feature of the problem.

Another commonly taken approach is to view the problem as a single classification problem by treating each possible combination of the labels as a new label l^*; such a transformation is named *label powerset*(LP). The pruned problem transformation (PPT) method proposed by *Read* (2008) extends LP by pruning away label sets that occur less times than a small user-defined threshold and optionally replaces their information by introducing disjoint subsets of these label sets that do exist more times than the threshold. *Tsoumakas and Vlahavas* (2007) proposed random k-labelsets(RAkEL) method which constructs an ensemble of LP classifiers. Each LP classifier is trained using a different small random subset of $\mathcal{L}$. RAkEL manages to take into account the label correlation, if not explicitly. However, the major concern of this transformational approach is that the data can be very sparse in the sense that for many combinations of the labels, there might be very few instances in the training data or even none, which results in fitting these classes difficult or even impossible for most of the single-label classifiers. Also, the size of transformed label space increases exponentially comparing to the original label set size q. For example, if the original label set has 8 labels, they are transformed into 256 new labels which might be too much to handle even for the very best multi-class classification techniques.

A third majority of approaches *Ranking the labels* is uniquely motivated by the multi-label classification problem itself and attracts much of interests in the text categorization applications. Instead of trying to construct a classifier that outputs multiple labels directly, this approach focuses on the relevance level of the labels to the instance and attempts to find a score function that assigns higher value to the more relevant labels and outputs a ranked list of the labels. *Schapire and Singer* (2000) proposes a Boosting algorithm called AdaBoost.MR(commonly known as BoosTexter) which modifies the boosting algorithm to fit a ranking function for all document-label pairs. *Joachims* (2002) designed a SVM-based method (RankSVM) to perform a classification task on label-instance pairs and finally produce a linear ranking function

that aims at minimizing the mismatched pairs of labels for all samples. *Crammer and Singer* (2003) proposes a family of one-against-the-rest online ranking algorithms that create a weight vector for each label and compute the ranking between a document and a label using the inner product of the feature vector and the weight vector. These approaches have straightforward objectives that makes intuitive sense and the optimization procedure is usually standard without too much computational challenge. However, in terms of prediction, the ranking approaches have to choose a threshold on the score function in order to output the labels whose score are above it and in most of the cases, the choice of this threshold is heuristic.

Apart from the three most common approaches mentioned above, there also exists some other notable works which propose generative models explicitly. *McCallum* (1999) proposed a mixture model trained by EM, selecting the most probable set of labels from the power set of all possible subsets and used heuristics to overcome the associated computational complexity. However, his generative model is based on learning word frequencies in documents and is thus restricted to text applications. *Ueda and Saito* (2003) also proposed a probabilistic generative model that uses a different mixture approach. The advantage of their approach is they explicitly model the label correlations and require no threshold for determining the category label for each data point. However, this paper shares the same drawback that it only applies to text categorization problem and their prediction rule is not optimal. Two other related works are done by *Niagm et al.* (1999) and *Zhu et al.* (2005) based on the idea of *Maximum Entropy*. Nigam et al. first introduce maximum entropy techniques to model the conditional distribution of labels given the feature vectors and seek for the solution by improved iterative scaling algorithm. However, they still break the problem into single-labeled problems and fit the conditional probability of each label given the data individually. *Zhu et al.* (2005) et al. modified this approach by exploring the correlation between different labels with maximum entropy techniques

and derive a classification algorithm which predicts the label set directly.

In the presented work, we formulate the multi-label classification task into a multivariate binary response regression problem and propose several methods to tackle data sets with both small dimension and large dimension of labels. To better assist the mathematical formulation of the problem, we can represent the label set $\mathcal{Y}_i$ as q-dimensional binary vector $\boldsymbol{y}_i$, with the jth element being the indicator of whether label l_j is relevant for the ith sample. As the foundation, we propose a covariate dependent binary Markov network, (also known as the Ising model), to explicitly model the conditional distribution $P(\boldsymbol{y}|\boldsymbol{x})$ and predict the label set with the maximum likelihood estimates of the fitted model. This model is efficient to fit however is far less efficient to compute the prediction for large dimension of labels as the number of possible label set increases exponentially with q. To overcome this prediction computational difficulty, we propose two alternative methods based on this model to work for large dimension of labels (recommended for $q \geq 20$). One is a two-step regression approach and the other is an ensemble method. We then apply our methods on five benchmark data sets and show their performances. Compared to most of existing methods, our approach takes into account of the correlation among labels and is capable of handling large dimensional data. The final classification boundary for each individual label is in general non-linear, which allows us to build more flexible models. We also point out that for multi-label classification tasks, different prediction rules based on the joint likelihood should be used for minimizing different classification error measures.

The remainder of the report is organized as follows. In section 4.2, we describe the foundational model as building blocks. Section 4.3 addresses the prediction and model tuning issue. Section 4.4 proposes two alternatives to tackle the large dimensional data. We apply our method to the several bench mark data sets and present the results in section 4.5. Section 4.6 concludes with discussions and future work.

4.2 Covariate Dependent Multivariate Binary Markov Model

In a multi-label classification, the training data are denoted as $\{(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$, where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iP}) \in \mathbb{R}^p$ is the feature vector for the ith sample and $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iq}) \in \{0,1\}^q$ is the binary response vector. $y_{iq} = 1$ indicates that the q-th label is associated with the i-th data point and $y_{iq} = 0$ otherwise. Our goal is to explicitly model the conditional distribution $P(\mathbf{y}|\mathbf{x})$ and come up with optimal prediction rules given different evaluation measures. As mentioned in the introduction previously, the reason we adopt a joint probabilistic model is that we would like to take into account of the correlations among the labels and use the information to produce better prediction. A simple yet flexible model we choose to use here, is based on the binary Markov network model, also known as Ising model. For a binary random vector $\boldsymbol{y} = (y_1, \ldots, y_q)$, an Ising model characterize its distribution via the following

$$P(\boldsymbol{y}) = \frac{1}{Z(\boldsymbol{\theta}, \boldsymbol{\gamma})} \exp\left(\sum_{j=1}^{q} \theta_j y_j + \sum_{j>k} \theta_{jk} y_j y_k\right),$$

where the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q, \theta_{12}, \ldots, \theta_{1q}, \theta_{23}, \ldots, \theta_{q-1q})$ fully characterizes the model and $Z(\boldsymbol{\theta})$ is the partition function which ensures the probability sum up to 1. An Ising model can be thought as a second order approximation of all possible distributions of a binary random vector and thus represents a rich class of distributions. Most importantly, an Ising model is known for its convenient interpretation of conditional dependency among the variables; specifically, y_j and y_k are conditionally independent given the other variables if and only if $\theta_{jk} = 0$.

In presence of covariates $\boldsymbol{x}$, a natural way to extend the model is to assume the conditional distribution $P(\boldsymbol{y}|\boldsymbol{x})$ follows an Ising model associated with $\boldsymbol{x}$. We propose the extension as follows

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{\theta}(\boldsymbol{x}))} \exp\left(\sum_{j=1}^{q} (\boldsymbol{\theta}_{jj}^T \boldsymbol{x}) y_j + \sum_{j>k} \theta_{jk} y_j y_k\right). \tag{4.1}$$

For simplicity of the notation, we assume that the first element of covariates $\boldsymbol{x}$ is always 1 so we do not need an extra intercept term.This model has a total number of parameters as $\mathcal{O}(\max(pq, q^2))$. To assist our understanding of the interpretations of the parameters, we consider the conditional log-odds of each of the response variable

$$\log\left(\frac{P(y_j = 1|\boldsymbol{y}_{-j}, \boldsymbol{x})}{1 - P(y_j = 1|\boldsymbol{y}_{-j}, \boldsymbol{x})}\right) = \boldsymbol{\theta}_{jj}^T \boldsymbol{x} + \sum_{k:k\neq j} \theta_{jk} y_k. \tag{4.2}$$

The interpretation of each parameter becomes straightforward in this setting: θ_{jjl} directly measure the effect of covariate x_l on the log-odds of y_j and θ_{jk} measures the effect of y_k on the same conditional log-odds ratio. This formulation shares the same assumption as a logistic regression model, where the response is one of the binary variable y_j, and the predictors are the feature vector $\boldsymbol{x}$ and remaining response vector $\boldsymbol{y}_{-j}$.

The joint likelihood $P(\boldsymbol{y}|\boldsymbol{x})$ involves a normalizing constant $Z(\boldsymbol{\theta})$ which is the sum of 2^q terms and makes it difficult to maximizing this likelihood directly. Instead, we have a nice linear logistic model in (4.2) based on the conditional likelihood $P(y_j|\boldsymbol{y}_{-j}, \boldsymbol{x})$; so we maximize pseudo likelihood of the data to obtain the parameter estimates. Further more, we introduce ℓ_1 penalty on the parameters to control the sparsity of the estimates so that our model can be used for high-dimensional data scenario which is very common in multi-label data set. To be more specific, the estimates of the parameters is obtained by solving the following optimization problem

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta})}{\operatorname{argmin}} -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{q} \log(P(y_{ij}|\boldsymbol{y}_{i(-j)}, \boldsymbol{x}_i)) + \lambda\|\boldsymbol{\theta}\|_1,$$

where $\lambda > 0$ is the tuning parameter that controls the degree of penalization. We adopt the *coordinate shooting algorithm* (*Fu*, 1998) to solve a reweighed single-parameter lasso problem to update one parameter at a time. We then use 5-fold cross validation on the chosen classification error is employed to choose the optimal λ.

To end this section, we would like to point out a connection of our model with the maximum entropy method. *Niagm et al.* (1999) first introduced maximum entropy distributions for text categorization problems, which was later extended by *Zhu et al.* (2005). They explore the correlations among labels and attempt to derive $P(\mathbf{y}|\mathbf{x})$ based on maximum entropy criterion. Specifically, then they seek a solution to the following constrained optimization problem

$$\hat{P} = \underset{P}{\operatorname{argmax}}\, \mathbb{E}_P(\log(P(\boldsymbol{y}|\boldsymbol{x}))),$$

subject to

$$\begin{aligned}
\langle y_j \rangle_P &= \langle y_j \rangle_{\tilde{P}}, \qquad \text{for all } 1 \le j \le q; \\
\langle y_j x_l \rangle_P &= \langle y_j x_l \rangle_{\tilde{P}}, \qquad \text{for all } 1 \le j \le q \text{ and } 1 \le l \le p; \\
\langle y_j y_k \rangle_P &= \langle y_j y_k \rangle_{\tilde{P}}, \qquad \text{for all } 1 \le j < k \le q; \\
\sum_{\boldsymbol{y}} P(\boldsymbol{y}|\boldsymbol{x}) &= 1;
\end{aligned}$$

where $\tilde{P}$ is the empirical distribution of the training data. The constrained optimization problem can be solved using Lagrange multiplier algorithms. The final distribution has the same form as our covariate depenent Ising model (4.1). However, since the motivation behind the model assumption is different the results are not the same. The maximum entropy methods are actually fitting the exact log-likehood with an additional ℓ_2 penalty on the parameters.

4.3 Prediction and Tuning

4.3.1 Classification Error Measures and Optimal Prediction Rules

The previous section builds up to the model fitting part, yet the final goal is to predict the most probable combination of labels based on the estimated probability $P_{\hat{\boldsymbol{\theta}}}(\boldsymbol{y}|\boldsymbol{x})$. In this section, we discuss that for different misclassification error measure,

we should adopt different likelihood-based estimates. For a data point $\boldsymbol{x}$, Let $\hat{\boldsymbol{y}}(\boldsymbol{x}) = (\hat{y}_1(\boldsymbol{x}), \ldots, \hat{y}_q(\boldsymbol{x}))$ be the predicted label vector and $\boldsymbol{y}(\boldsymbol{x}) = (y_1(\boldsymbol{x}), \ldots, y_q(\boldsymbol{x}))$ be the true label set. The following two classification errors are usually considered in the literature.

$$\begin{aligned} \text{Hamming-Loss} &= \frac{1}{q}\sum_{j=1}^{q} \mathbf{I}(\hat{y}_j(\boldsymbol{x}) \neq y_j(\boldsymbol{x})), \\ \text{Vector Error} &= \mathbf{I}(\hat{\boldsymbol{y}}(\boldsymbol{x}) \neq \boldsymbol{y}(\boldsymbol{x})). \end{aligned}$$

Hamming loss is the average number of incorrectly predicted labels, where each labels are computed separately and the errors are accumulated. Vector error in general is a much more strict evaluation measure as it requires the predicted set of labels to be an exact match of the true set of labels for loss being 0, otherwise the loss is 1. Depending on which of the two quantities fits better of the application, the optimal prediction rule differs. The following proposition illustrates this point.

Proposition IV.1. *Given a joint distribution $P(\boldsymbol{y}, \boldsymbol{x})$, where $\boldsymbol{y} \in \{0,1\}^q$ and $\boldsymbol{x} \in \mathbb{R}^q$, define*

$$\begin{aligned} \hat{\boldsymbol{y}}_{marginal}(\boldsymbol{x}) &= (\tilde{y}_1(\boldsymbol{x}), \ldots, \tilde{y}_q(\boldsymbol{x})), \quad \textit{where } \tilde{y}_j(\boldsymbol{x}) = \operatorname*{argmax}_{y_q} P(y_j|\boldsymbol{x}), \\ \hat{\boldsymbol{y}}_{joint}(\boldsymbol{x}) &= \operatorname*{argmax}_{y} P(\boldsymbol{y}|\boldsymbol{x}), \end{aligned}$$

then the following holds

$$\hat{\boldsymbol{y}}_{marginal}(\boldsymbol{x}) = \operatorname*{argmin}_{\hat{\boldsymbol{y}}(\boldsymbol{x})} \frac{1}{q}\sum_{j=1}^{q} P(\hat{y}_j(x) \neq y_j(\boldsymbol{x})) \ , \quad \hat{\boldsymbol{y}}_{joint}(\boldsymbol{x}) = \operatorname*{argmin}_{\hat{\boldsymbol{y}}(\boldsymbol{x})} P(\hat{\boldsymbol{y}}(\boldsymbol{x}) \neq \boldsymbol{y}(\boldsymbol{x})).$$

Proposition IV.1 states a convenient fact that if the goal is to minimize the expected hamming loss, we should use maximum marginal probability of each y_q to predict while if the goal is to minimize vector classification error, maximum joint probability estimates are preferred. Though the point is not complicated to derive, it

has been overlooked by most of the existing methods where they all use the maximum joint probability prediction as default even though they aim to minimize hamming loss. Note that all of above are valid for general distribution $P(\boldsymbol{y}|\boldsymbol{x})$ without assuming Ising model.

In terms of computation, both marginal and joint prediction would involve 2^q calculations of the probability for each combination of the labels and the maginal prediction would involve one more round of summation. This is computationally very challenging when the size of label set q is even of moderate size, say 15. Thus we discuss two alternative approaches based on model (4.1) in the next section to deal with data set with large q .

4.3.2 Choice of Tuning Parameter λ

As mentioned in section 4.2, we are choosing the tuning parameters based on cross-validation. However, it will involve the choice of a proper error/accuracy measure for the multi-label prediction. The Hamming loss and the vector error preferred different prediction rules as mentioned above, thus would not serve as a fair criterion. Other than the two classification errors, there are a few more accuracy measures that are often used in the multi-label context.

$$\begin{aligned}
\text{Precision} &= \frac{\sum_{j=1}^{q} \mathbf{I}(y_j = 1,\ \hat{y}_j = 1)}{\sum_{j=1}^{q} \mathbf{I}(\hat{y}_j = 1)} \quad \text{(proportion of correct labels among the predicted ones)},\\
\text{Recall} &= \frac{\sum_{j=1}^{q} \mathbf{I}(y_j = 1,\ \hat{y}_j = 1)}{\sum_{j=1}^{q} \mathbf{I}(\hat{y}_j = 1)} \quad \text{(proportion of correct labels among the true ones)},\\
\text{F1} &= \frac{\sum_{j=1}^{q} \mathbf{I}(y_j = 1,\ \hat{y}_j = 1)}{\sum_{j=1}^{q} \mathbf{I}(y_j = 1) + \mathbf{I}(\hat{y}_j = 1)} \quad \text{(in between Precision and Recall)},\\
\text{Accuracy} &= \frac{\sum_{j=1}^{q} \mathbf{I}(y_j = 1,\ \hat{y}_j = 1)}{\sum_{j=1}^{q} \mathbf{I}(\hat{y}_j = 1 \text{ or } y_j = 1)} \quad \text{(smaller than both Precision and Recall)}.
\end{aligned}$$

Maximizing the precision encourage methods that output a small number of labels while maximizing the recall will oppositely encourage method to output too many

labels. Therefore, we use F1 measure as our cross-validation criterion to balance the precision and recall simultaneously. Specifically, we conduct a 5-fold cross validation, compute the F1 measure on each of the testing data and choose the λ that maximize the average F1 measure.

We shall mention that in multi-label classification literature, some other evaluation measures such as one -error, ranking loss, average precision, coverage are used. These measures are all based on the predicted ranking of the labels. They are not suitable if the output is already the subset of labels as in our approach; therefore we do not use these measures in our later analysis.

4.4 Two Alternative Approaches for Large-Sized Label Set

As mentioned in the end of subsection 4.3.1, the computational cost for prediction based on the joint probability model $P(\boldsymbol{y}|\boldsymbol{x})$ grows exponentially with the size of the label set q. Therefore, this section proposes two alternative approaches motivated by the original model (4.1) to accommodate the computational inefficiency. The first approach is a two-step approach modeling the conditional log-odds of each label and the final output are separate binary linear classifiers; however the model training involves interaction between the labels thus takes into account of the label correlation. The second approach is an ensemble method that creates a compromise between the prediction computation and the joint probability model where we randomly select subsets of the labels to fit model (4.1) and then aggregates of many sub-models to produce the final prediction. Details of the two methods are discussed as follows.

4.4.1 Two-step Logistic Model

One way to avoid joint probability modeling is to build a separate logistic model for y_j depending only on the feature vector $\boldsymbol{x}$, in which case the prediction computation is linear in q rather than exponential. Meanwhile, we hope not to lose the mutual

association among the labels which might help improve the prediction accuracy. Motivated by the conditional log-odds in (4.2), the prediction can be done through the conditional probability if we have $\boldsymbol{x}$ and all the other y_k's; in reality however, we do not have the value of other labels as they themselves are part of the prediction task, therefore, we propose to replace the y_k with a surrogate based on $\boldsymbol{x}$ only, i.e, some score function $f_k(\boldsymbol{x})$. Ideally, $f_k(\boldsymbol{x})$ should be viewed as an approximation to y_k, indicating how likely y_k is to be 1. One natural candidate of such surrogate would be the marginal log-odds of y_k given $\boldsymbol{x}$. To be specific, we propose the following two-step model.

Step I: Fit logistic regression on $y_j \sim \boldsymbol{x}$ to obtain $\hat{\boldsymbol{\beta}}_j$ for all $j = 1, \ldots, q$

$$\log\left(\frac{P(y_k = 1|\boldsymbol{x})}{1 - P(y_k = 1|\boldsymbol{x})}\right) = {\boldsymbol{\beta}_k}^T \boldsymbol{x}.$$

Step II: Fit logistic regression on $y_j \sim (\boldsymbol{x}, \{f_k(\boldsymbol{x}, \hat{\boldsymbol{\beta}}_k), k \neq j\})$ to obtain $(\hat{\boldsymbol{\theta}}_{jj}, \hat{\boldsymbol{\theta}}_{jk})$, where $f_k(\boldsymbol{x}, \hat{\boldsymbol{\beta}}_k) = \text{logistic}(\hat{\boldsymbol{\beta}}_k^T \boldsymbol{x})$,

$$\log\left(\frac{P(y_j = 1|\boldsymbol{x})}{1 - P(y_j = 1|\boldsymbol{x})}\right) = {\boldsymbol{\theta}_{jj}}^T \boldsymbol{x} + \sum_{k:k\neq j} \theta_{jk} f_k(\boldsymbol{x}, \hat{\boldsymbol{\beta}}_j). \tag{4.3}$$

The final classification rule for each y_j is given in (4.3) which depends on $\boldsymbol{x}$ only; therefore prediction can be done separately. Note that each classifier is not linear in $\boldsymbol{x}$ due to the logistic transformation of the first step model, which increases the flexibility of the boundary.

4.4.2 Ensemble Method

While prediction on large dimension labels become prohibitive because of the computational cost, we would still like to predict the labels jointly to gain accuracy. A compromise between the two leads us to a solution where the problem is broken into smaller subsets and we fit joint models on the subsets separately. The results

can be aggregated from a number of these smaller models. To be more specific, we randomly sample the subsets of labels with size k, where k is much smaller than q. We fit the model (4.1) to the subset of labels and repeat this experiment for many times. The final prediction takes the majority votes for each label from sub-models in which that label is being selected. The following describes the ensemble method in mathematical details.

1. Repeat the following for $m = 1, \ldots, M$:

   (a) Randomly sample (without replacement) S_m from the entire label set $\{1, \ldots, q\}$, where $\|S_m\|_0 = k$.

   (b) Fit model (4.1) on $\{(\boldsymbol{x}_i, \boldsymbol{y}_{iS_m}) : i = 1, \ldots, n\}$, denote the fitted model by $P_m(\boldsymbol{y}_{S_m}|\boldsymbol{x})$.

2. Given any input $\boldsymbol{x}$, let $\hat{\boldsymbol{y}}_{S_m} = \text{argmax}_{\boldsymbol{y}_{S_m}} P_m(\boldsymbol{y}_{S_m}|\boldsymbol{x})$, and $\hat{y}_{j,m}$ is the prediction for y_j for all $j \in S_m$, we have

$$\hat{y}_j = \frac{1}{M} \sum_{m=1}^{M} \hat{y}_{j,m} \mathbf{1}(j \in S_m).$$

4.4.3 Approximation Methods

There has been many literature on approximating algorithms for finding the maximum likelihood estimation of binary Markov random fields, most of which originated from the problem of noisy figure reconstruction. For comparison purpose, we also apply two approximation methods in our data analysis, respectively ICM (Iterative Conditional Modes) (*Besag*, 1986) and MinCut (Minimum Cut of a Graph) (*Greig et al.*, 1989). The generative model we fit is the same as (4.1), with only the prediction approximated to gain computational efficiency.

4.5 Numerical Examples

4.5.1 Simulation Study: Prediction

In this section we conduct simulation study to further illustrate the prediction rules in proposition IV.1. The purpose of this study is to show that when different classification error measures are considered, maximum marginal probability prediction and maximum joint probability prediction should be differentiated accordingly. We show that under certain settings, the difference of prediction accuracy can be significant. Most of the researchers only consider optimizing Hamming loss or using Hamming loss as evaluation measure since zero-one error is a very strict measure and it is usually high. Hence we will mainly focus on Hamming loss in this section.

4.5.1.1 Simulation 1: Ising model without covariates

This simulation considers only $\boldsymbol{y} \in \{0,1\}^q$ from an Ising model without covariates $\boldsymbol{x}$. We specify one such sufficient condition on Chain graphical structure that the marginal prediction has a significant advantage over the joint prediction in minimizing the Hamming loss. A Chain graphical model indicates that y_j is conditionally independent with all other y_k's except y_{j-1} and y_{j+1}.

Lemma IV.2. *Assume* $\boldsymbol{y} \sim Ising(\boldsymbol{\Theta}, \boldsymbol{\theta})$*, if the following holds,*

1. $\theta_1 = -\theta_{12} = \|\theta_1^-\|_1 > 0$, $\quad \theta_2 > \theta_1$*, where,* $\boldsymbol{\theta}_1 = (\theta_{12}, \ldots, \theta_{1Q})$.
2. $\theta_{qq'} < 0, \quad$ *for* (q, q')*:* $|q - q'| = 1$.
3. *all the other* θ_q*'s and* $\theta_{qq'}$*'s are 0.*

Then the maximum joint probability prediction has a higher expected hamming loss than the maximum marginal probability prediction.

An easy extension of this lemma to more general graphical structures is by setting all the other θ_{jk}'s to be negative, then the same results will hold as well.

We simulate 9 sets of Ising models satisfying the above conditions in lemma IV.2, calculate the theoretical hamming loss and compare the difference of loss between joint prediction and marginal prediction. We vary the dimension $q \in \{3, 6, 9\}$ and the magnitude of parameters which are uniformly generated in $[1, 5]$, $[6, 10]$, and$[11, 15]$ respectively. In every setting, 50 Ising models are generated and we compute the exact theoretical value of the Hamming loss both on the joint prediction and the marginal prediction. Figure 4.1 shows the difference of Hamming loss between two prediction methods over 50 repetitions in boxplots. Table 4.1 records the mean(std) of the Hamming loss.

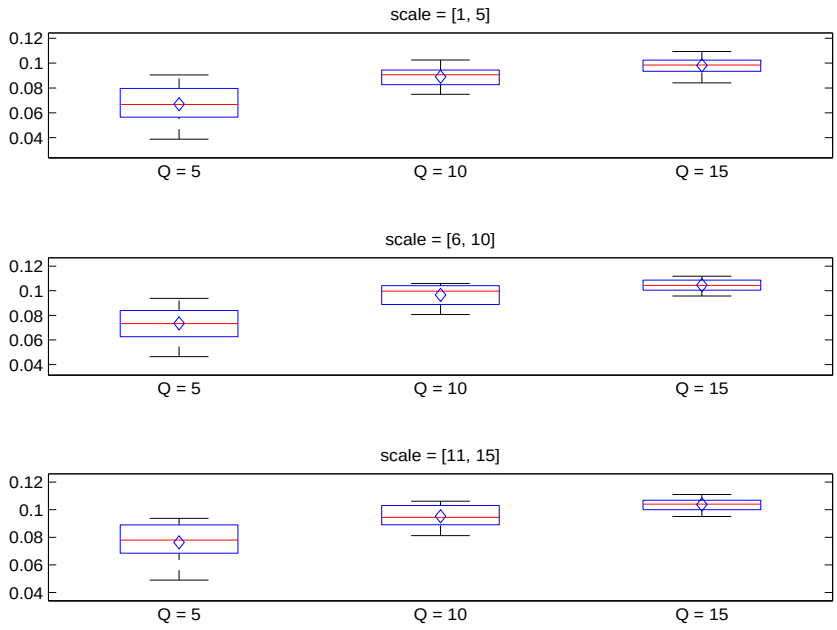

Figure 4.1: Hamming-loss($\hat{y}_{joint}$)-Hamming-loss($\hat{y}_{marginal}$)

From the box plots 4.1, it is clear that the difference of the Hamming loss is always positive which means the marginal prediction is always better than the joint prediction under the setting of Lemma IV.2. As the dimension q increases, the advantage became more significant (higher differences) and more consistent (smaller dispersion). The difference does not seem to be affected by the parameters scale that much. From Table

		hloss($\boldsymbol{y}_{joint}$)	hloss($\boldsymbol{y}_{marginal}$)	Δhloss
scale = [1,5]	Q = 5	0.3609(0.0357)	0.2939 (0.0236)	0.0670 (0.0142)
	Q = 10	0.3854 (0.0165)	0.2964 0.0102)	0.0890 (0.0075)
	Q = 15	0.3943 (0.0121)	0.2962 (0.0085)	0.0981 (0.0058)
scale = [6,10]	Q = 5	0.3515 (0.0347)	0.2780 (0.0210)	0.0735 (0.0137)
	Q = 10	0.3779 (0.0182)	0.2812 0.0103)	0.0967 (0.0079)
	Q = 15	0.3826 (0.0110)	0.2781 (0.0062)	0.1045 (0.0048)
scale = [11,15]	Q = 5	0.3584 (0.0342)	0.2821 (0.0207)	0.0763 (0.0135)
	Q = 10	0.3746 (0.0177)	0.2793 (0.0100)	0.0953 (0.0077)
	Q = 15	0.3808 (0.0102)	0.2770 (0.0058)	0.1038 (0.0044)

Table 4.1: average value(std) of Hamming-Loss for different prediction

4.1, we can see the accuracy gain using marginal prediction comparing to the original Hamming loss using joint prediction is quite large in terms of relative percentage, for example, when $q = 15$ and $scale = [11, 15]$, the reduction in Hamming loss by marginal prediction is 7.63%, which is around 20% of the original 35.84%.

4.5.1.2 Simulation 2: Ising model with covariates

When covariates $\boldsymbol{x}$ are considered, explicit calculation of the theoretical Hamming loss become intractable because we can not integrate the partition function $Z(\boldsymbol{\theta}, \boldsymbol{x})$. Still, we can numerically evaluate the Hamming loss by generating a large number of data points from the joint distribution $P(\boldsymbol{y}, \boldsymbol{x})$ and calculating the sample average of Hamming loss. The data are generated as follows: Firstly, we generate features $\boldsymbol{x}$ from a multi-variate normal distribution with mean 0 and covariance matrix I; for each given $\boldsymbol{x}$, the distribution of $\boldsymbol{y}$ is Ising model (4.1) with chain structure, where the θ_{jj1}'s, θ_{jk}'s are generated under the same condition as in Lemma IV.2 and scales in [1, 5]; for the other θ_{jjl}'s where $l > 1$, we set them to be 0.5. We set $q \in \{3, 9\}$ and $p \in \{20, 200\}$ respectively. For each combination of p and q, 20 replication are generated and the results are shown in Figure 4.2.

The results in Figure 4.2 shows the difference in Hamming-loss under each setting. Only for $p = 20$ and $q = 3$, the median of the difference was around 0 with the lower

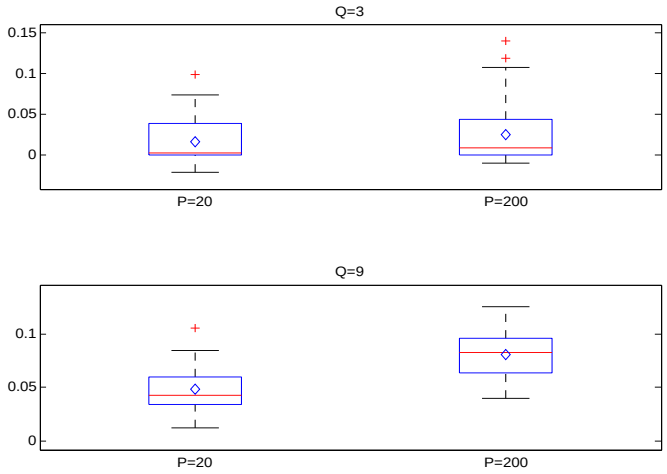

Figure 4.2: Hamming-loss($\hat{y}_{joint}|x$)-Hamming-loss($\hat{y}_{marginal}|x$)

half very close to 0; as either p and q grows larger, all of the differences are shifting to the larger side. For $q = 9$ specifically, the difference is strictly all above 0 which partially coincides with the results when there is no covariates that the advantage of marginal probability prediction is more significant for larger dimension of q.

4.5.2 Real Data Analysis

In this section, we apply the proposed methods on four multi-label data sets and compare the results for different prediction rules as well as the alternative methods when the label set is of large dimension. The data sets are downloaded from a publicly available online library MuLan(a java library for multi-label learning). We start with a brief description of each one of the data sets and Table 4.2 gives the summary statistics for each data set.

emotions data set This data set was created using a collection of 700 songs equally distributed between 7 different genres. For each song a period of 30 seconds after the initial 30 seconds was extracted and used for labeling and feature extraction. Two types of features were considered for this study which are rhythmic features and timbre features. The labels consisted of 6 main emotion groups, including amazed-

surprised, happy-pleased, relaxing-calm, etc. Each song were rated by three male experts. The songs which had exactly identical labels for all three annotators were included in the final data set. This led to a final data set consisting of 593 songs. More details regarding the data set could be found at *Trohidis et al.* (2008).

scene data set This data set consists nearly 2400 images of Corel stock photo library and personal images. There are 6 pre-specified labels including beach, sunset, field, etc.; the images were originally chosen so that each of the 6 class contained around 400 images. The images were later re-labeled with multiple labels by three human observers. After re-labeling, approximately 7.4% of the images belonged to multiple classes. The features of each image are spatial color moments in LUV space. For more details of the data set, see *Boutell et al.* (2004)

yeast data set This data set (*Elisseeff and Weston*, 2002) is formed by micro-array expression data and phylogenetic profiles with a total of 2417 genes. The data were generated from spotted arrays using samples collected at various time points under different experimental conditions. Each gene is associated with a set of functional classes whose maximum size can be potentially 190. The whole set of classes if indeed structured is a tree whose leaves are functional categories. Here we try to predict all the first level nodes of the tree which has a total number of 14.

CAL500 data set This data set consists a set of 500 Western popular songs from 500 unique artists, each of which was reviewed by a minimum of three individuals. The labels are semantic annotations containing categories such as emotional content, genre, instrumentation, and vocal characterizations. A vocabulary of 174 musically relevant semantic keywords are finally created as the label set. Each song is represented as feature vector calculated by analyzing a short-time segment of the audio signal. More details regarding the description of the data set can be found in *Turnbull et al.* (2008)

For *emotions* and *scene* data sets, we fit the original model (4.1) as well as the two

name	**domain**	n	p	q	**cardinality**	**density**	**distinct**
emotions	music	593	72	6	1.869	0.311	27
scene	image	2407	294	6	1.074	0.179	15
yeast	biology	2417	103	14	4.237	0.303	198
CAL500	music	502	68	174	26.044	0.15	502

Table 4.2: Summary statistics of the data sets. **cardinality**: average number of labels per sample; **density**: average proportion of labels per sample; **distinct**: number of distinct labels combinations in the data set.

approximation algorithms for joint prediction; for *yeast* data set, we fit the original model as well as the two alternative methods; for *CAL500* data set, joint prediction is infeasible due to the large size of the label set, therefore we implement only the alternative methods mentioned in section 4. For each of the test data set, we report the six error/accuracy measures mentioned in section 3, i.e, Hamming loss, vector error, precision, recall, F1 and accuracy, with the optimal tuning parameters chosen by 5-fold cross validation.

data set	method	1-HamLoss	1-VecErr	Precision	Recall	F1	Accuracy
emotions	baseline	0.8741	0.4507	0.7923	0.7622	0.7500	0.6793
	joint-prob	0.7883	0.3338	0.6360	0.7111	0.6495	0.5701
	mrg-prob	0.8027	0.2933	0.6638	0.6126	0.6074	0.5303
scene	baseline	0.9000	0.5476	0.6311	0.6595	0.6344	0.6124
	joint-prob	0.8974	0.6647	0.7262	0.6979	0.7068	0.6962
	mrg-prob	0.8914	0.5214	0.5706	0.5539	0.5577	0.5486
yeast	baseline	0.8002	0.1469	0.7048	0.5844	0.6111	0.5014
	joint-prob	0.7864	0.2044	0.6589	0.6271	0.6219	0.5171
	mrg-prob	0.7951	0.1192	0.7297	0.5274	0.5822	0.4681
	two-step	0.7935	0.1444	0.6853	0.5923	0.6077	0.4982
	ensemble	0.7993	0.1680	0.6980	0.6043	0.6218	0.5135
CAL500	ensemble	0.8504	0	0.5130	0.3090	0.3743	0.2369

Table 4.3: Results of proposed methods on testing data sets.

4.6 Conclusion

In this project, we propose to use a binary Markov network, i.e, Ising model with covariates, to explicitly model the conditional distribution $P(\boldsymbol{y}|\boldsymbol{x})$ for a multi-label classification problem. The pseudo-likelihood based method is adopted to develop a computationally efficient estimation procedure. We also investigate the choice of evaluation measures in connection to different prediction rules. For prediction with the joint probability, however, computation cost can become huge when the size of the label set is even moderately large. Therefore, we consider two alternative approaches motivated from the original Ising model with covariates: one approach is to fit separate logistic models to each label by approximating the interaction with other labels with a score function fit beforehand; the other approach is an ensemble method that aggregates a collection of smaller models built on randomly selected subsets of the labels. We apply all of the methods on four benchmark multilabel data sets and compare their prediction performance.

4.7 Appendix

proof of Proposition IV.1 To prove the proposition, we need to following lemma.

Lemma IV.3. *Given* $\boldsymbol{y} = (y_1, \ldots, y_Q) \sim Pr(\boldsymbol{y})$, $\boldsymbol{y} \in \{0,1\}^Q$, *define*

$$\begin{aligned} \hat{\boldsymbol{y}}_{marginal} &= (\tilde{y}_1, \ldots, \tilde{y}_Q), \quad \textit{where } \tilde{y}_q = argmax_{y_q} Pr(y_q) \\ \hat{\boldsymbol{y}}_{joint} &= argmax_{y} Pr(\boldsymbol{y}) \end{aligned}$$

the following holds

$$\hat{\boldsymbol{y}}_{marginal} = argmin_{\hat{\boldsymbol{y}}} \frac{1}{Q} \sum_{q=1}^{Q} Pr(\hat{y}_q \neq y_q) \quad , \qquad \hat{\boldsymbol{y}}_{joint} = argmin_{\hat{\boldsymbol{y}}} \; Pr(\hat{\boldsymbol{y}} \neq \boldsymbol{y})$$

Proof. Given any prediction $\hat{\boldsymbol{y}}$,

$$\begin{aligned}
\frac{1}{Q}\sum_{j=1}^{Q} Pr(\hat{y}_q \neq y_q) &= \frac{1}{Q}\sum_{q=1}^{Q} Pr(y_j = 1 - \hat{y}_q) \\
&\geq \frac{1}{Q}\sum_{q=1}^{Q} \min\{Pr(y_q = 1), Pr(y_j = 0)\} \\
&= \frac{1}{Q}\sum_{q=1}^{Q} Pr(\hat{y}_{q,marginal} \neq y_q)
\end{aligned}$$

Similarly,

$$\begin{aligned}
Pr(\hat{\boldsymbol{y}} \neq \boldsymbol{y}) &= 1 - Pr(\boldsymbol{y} = \hat{\boldsymbol{y}}) \\
&\geq 1 - \max_{\boldsymbol{y}} Pr(\boldsymbol{y}) = Pr(\hat{\boldsymbol{y}}_{joint} \neq \boldsymbol{y})
\end{aligned}$$

□

Lemma IV.3 basically tells that if the goal is to minimize hamming loss, the marginal prediction $\hat{\boldsymbol{y}}_{marginal}$ based on maximum marginal probability is better than the joint prediction $\hat{\boldsymbol{y}}_{joint}$, which is actually the label set with the maximum joint probability; however if the goal is to minimize zero-one classification error, then $\hat{\boldsymbol{y}}_{joint}$ is preferred to $\hat{\boldsymbol{y}}_{marginal}$. With this lemma, the results can be easily extended to models with covariates $\boldsymbol{x}$.

Proof. We only need to show the first part and the second part follows similarly. Given any prediction function $\hat{\boldsymbol{y}}(x)$,

$$\begin{aligned}
\frac{1}{Q}\sum_{q=1}^{Q} Pr(\hat{y}_q(\boldsymbol{x}) \neq y_q(\boldsymbol{x})) &= \frac{1}{Q}\sum_{q=1}^{q} \int Pr(\hat{y}_q(\boldsymbol{x}) \neq y_q(\boldsymbol{x}) \mid \boldsymbol{x})\, dP(\boldsymbol{x}) \\
&\geq \frac{1}{Q}\sum_{q=1}^{q} \int Pr(\hat{y}_{q,marginal}(\boldsymbol{x}) \neq y_q(\boldsymbol{x}) \mid \boldsymbol{x})\, dP(\boldsymbol{x}) \qquad \text{(By Lemma 1.)} \\
&= \frac{1}{Q}\sum_{q=1}^{Q} Pr(\hat{y}_{q,marginal}(\boldsymbol{x}) \neq y_q(\boldsymbol{x}))
\end{aligned}$$

□

The sufficient condition in section 4.3.1 is based on the following fact.

Fact. Assuming

$$Pr(y) = \frac{1}{Z(\Theta,\theta)} exp(\sum_{q=1}^{Q} \theta_q y_q + \sum_{q'>q} \theta_{qq'} y_q y_{q'}) = \frac{1}{Z(\boldsymbol{\Theta},\boldsymbol{\theta})} exp(y^T A y)$$

where

$$A = \begin{bmatrix} \theta_1 & \frac{\theta_{12}}{2} & \cdots & \frac{\theta_{1q}}{2} \\ \frac{\theta_{21}}{2} & \theta_2 & \cdots & \frac{\theta_{2q}}{2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\theta_{q1}}{2} & \frac{\theta_{q2}}{2} & \cdots & \theta_q \end{bmatrix}.$$

denote $\theta_j = (\theta_{q1}, \ldots, \theta_{qq-1}, \theta_{qq+1}, \ldots, \theta_{qQ})$; θ_q^+ is the positive part of θ element-wise, θ_q^- defined similarly. the following result holds:

1. if $\theta_q \geq \|\theta_q^-\|_1$, $Pr(y_q = 1) \geq 0.5$.

2. if $\theta_q \leq -\|\theta_q^+\|_1$, $Pr(y_q = 0) \geq 0.5$.

the above inequality holds strictly unless all the $\theta_{qq'}$'s are 0.

Proof. let A_q be the $(q-1)\times(q-1)$ matrix by deleting both qth row and qth column of A, then

$$\begin{aligned} \frac{Pr(y_q = 1)}{Pr(y_q = 0)} &= \frac{\sum_{y_{-q}} exp\{(\theta_q + \theta_q^T y_{-q}) + y_{-q}^T A_q y_{-q}\}}{\sum_{y_{-q}} exp(y_{-q}^T A_q y_{-q})} \\ &= \sum_{y_{-q}} w(y_{-q}) exp(\theta_q + \theta_q^T y_{-q}) \\ (1) \quad &\leq exp(\theta_q + \|\theta_q^+\|_1) \\ (2) \quad &\geq exp(\theta_q - \|\theta_q^-\|_1) \end{aligned}$$

$w(y_{-q}) = \frac{exp(y_{-q}^T A_q y_{-q})}{\sum_{y_{-q}} exp(y_{-q}^T A_j y_{-q})}$ and $\sum_{y_{-q}} w(y_{-q}) = 1$. From above, it is straight forward to reach the final conclusion. □

BIBLIOGRAPHY

BIBLIOGRAPHY

Adelaide, J., et al. (1998), Chromosome region 8p11-p21: Refined mapping and molecular alterations in breast cancer, *Genes, Chromosomes & Cancer*, *22*, 186–199.

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

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

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

Bergamaschi, A., Y. Kim, P. Wang, T. Sørlie, T. Hernandez-Boussard, P. Lonning, R. Tibshirani, A. Børresen-Dale, and J. Pollack (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*(11), 1033–1040.

Besag, J. E. (1986), On the statistical analysis of dirty pictures (with discussion), *Journal of the Royal Statistical Society, Series B*, *48*, 259 – 302.

Blumstein, D., G. Bryant, and P. Kaye (2012), The sound of arounsal in music is context-dependent, *Biology Letters*, *8*(5), 744–747.

Boutell, M., J. Luo, X. Shen, and C. Brown (2004), Learning multi-label scene classification, *Pattern Recognition*, *37*, 1757–1771.

Cai, T., H. Li, W. Liu, and J. Xie (2011a), Covariate adjusted precision matrix estimation with an application in genetical genomics, *Biometrika*, pp. 1–19.

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

Conte, N., E. Charafe-Jauffret, J. Adelaide, C. Ginestier, J. Geneix, D. Isnardon, J. Jacquemier, and D. Birnbaum (2002), Carcinogenesis and translational controls: Tacc1 is down-regulated in human cancers and associates with mrna regulators, *Oncogene*, *21*(36), 5619–5630.

Crammer, K., and Y. Singer (2003), A family of additive onlint algorithms for category ranking, *Journal of Machine Learning Research*, *3*, 1025–1058.

Dempster, A. (1972), Covariance selection, *Biometrics*, *28*, 157–175.

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

Elisseeff, A., and J. Weston (2002), A kernel method for multi-labelled classification, *Advances in Neural Information Processing Systems*, *14*.

Fellinghauer, B., P. Bühlmann, M. Ryffel, M. Rhein, and J. Reinhardt (2011), Stable graphical model estimation with random forests for discrete, continuous, and mixed variables, arXiv: 1109.0152.

Friedman, J., T. Hastie, and R. Tibshirani (2008), Sparse inverse covariance estimation with the graphical lasso, *Biostatistics*, *9*(3), 432–441.

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

Fu, W. (1998), Penalized regressions: the bridge versus the lasso, *Journal of Computational and Graphical Statistics*, *7*(3), 397–416.

Garcia, I., P. Dietrich, M. Aapro, G. Vauthier, L. Vadas, and E. Engel (1989), Genetic alterations of c-myc, c-erbb-2, and c-ha-ras protooncogenes and clinical associations in human breast carcinomas, *Cancer Research*, *49*(23), 6675–6679.

Geisler, S., A. Børresen-Dale, H. Johnsen, T. Aas, J. Geisler, L. Akslen, G. Anker, and P. Lonning (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.

Godbole, S., and S. Sarawagi (2004), Discriminative methods ofr multi-labeled classification, in *Proceedings of the 8th Pacific-Asia Conference on Knoledge Discovery and Data Mining (PAKDD 2004)*, pp. 22–30.

Greig, D. M., B. T. Porteous, and A. H. Seheult (1989), Exact maximum a posteriori estimation for binary images, *Journal of the Royal Statistical Society, Series B*, *51*(2), 271 – 279.

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

Guo, J., E. Levina, G. Michailidis, and J. Zhu (2010b), Estimating heterogeneous graphical models for discrete data with an application to roll call voting, manuscript.

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

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

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

Huang, T., P. Yeh, M. Martin, R. Straub, T. Gilliam, C. Caldwell, and J. Skibba (1995), Genetic alterations of microsatellites on chromosome 18 in human breast carcinoma, *Diagnostic Molecular Pathology*, *4*(1), 66–72.

Ising, E. (1925), Beitrag zur theorie der ferromagnetismus, *Zeitschrift für Physik*, *31*, 253–258.

Joachims, T. (1998), Text cateorization with support vector machines: Learning with many relevant features, in *Tenth European Conference on Machine Learning (ECML)*.

Joachims, T. (2002), Optimizing search engines using clickthrough data.

Lam, C., and J. Fan (2009), Sparsistency and rates of convergence in large covariance matrices estimation, *Annals of Statistics*, *37*(6B), 4254–4278.

Lassus, H., R. Salovaara, L. Altonen, and R. Butzow (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*(1), 35–42.

Lauritzen, S. (1996), *Graphical Models*, Oxford University Press.

Lauritzen, S. L., and N. Wermuth (1989), Mixed graphical association models, *Annals of Statistics*, *17*(1), 31–57.

Lee, J., and T. Hastie (2012), Learning mixed graphical models, arXiv: 1205.5012.

Liu, H., X. Chen, J. Lafferty, and L. Wasserman (2010), Graph-valued regression, *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, *23*.

Logan, B. (2000), Mel frequency cepstral coefficients for music modeling.

MacKay, J., P. Elder, D. Porteous, C. Steel, R. Hawkins, J. Going, and U. Chetty (1988), Partial deletion of chromosome 11p in breast cancer correlates with size of primary tumour and oestrogen receptor level, *British Journal of Cancer*, *6*(58), 710–714.

Manning, C., and H. Schutze (1999), *Foundations of Statistical Natural Language Processing*, MIT press.

Mazumder, R., and T. Hastie (2012), Exact covariance thresholding into connected components for large-scale graphical lasso, *JMLR*, to appear.

McCallum, A. (1999), Multi-label text classification with a mixture model trained by em, in *Proceedings of the AAAI '99 Workshop on Text Learning*.

Meinshausen, N., and P. Bühlmann (2006), High dimensional graphs and variable selection with the Lasso, *Ann. Statist.*, *34*(3), 1436–1462.

Meinshausen, N., and P. Bühlmann (2010), Stability selection, *Journal of the Royal Statistical Soceity*, *72*(4), 417–473.

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

Mitelman, F., F. Merterns, and B. Johansson (1997), A breakpoint map of recurrent chromosomal rearrangements in human neoplasia, *Nature Genetics*, *15*, 417–474.

Negrini, M., S. Sabbioni, L. Possati, S. Rattan, A. Corallini, G. Barbanti-Brodano, and C. Croce (1994), Suppression of tumorigenicity of breast cancer cells by microcell-mediated chromosome transfer: Studies on chromosomes 6 and 11, *Cancer Research*, *54*, 1331–1336.

Niagm, K., J. Lafferty, and A. McCallum (1999), Using maximum entropy for text classification, in *IJCAI Workshop on Machine Learning for Information Filtering*, pp. 61–67.

Peng, J., P. Wang, N. Zhou, and J. Zhu (2009), Partial correlation estimation by joint sparse regression model, *Journal of the American Statistics Association*, *104*(486), 735–746.

Ravikumar, P., M. Wainwright, G. Raskutti, and B. Yu (2008), Model selection in gaussian graphical models: High-dimensional consistency of l1-regularized mle, *Advances in Neural Information Processing Systems(NIPS)*, *21*.

Ravikumar, P., M. Wainwright, and J. Lafferty (2010), High-dimensional ising model selection using l1-regularized logistic regression, *Annals of Statistics*, *38*(3), 1287–1319.

Read, J. (2008), A pruned problem transformation method for multi-label classification, in *Proceedings of New Zealand Computer Science Research Student Conference*, pp. 143–150.

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

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

Schapire, R., and Y. Singer (2000), Boostexter: A boosting-based system for text categorization, *Machine Learning*, *39*, 135–168.

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

Tibshirani, R. (1996), Regression shrinkage and selection via the lasso, *J. Roy. Statist. Soc., Ser. B*, *58*, 267–288.

Trohidis, K., G. Tsoumakas, G. Kalliris, and I. Vlahavas (2008), Multilabel classification of music into emotions, in *Proceedings of International Conference on Music Information Retrieval (ISMIR)*, pp. 325–330.

Tsoumakas, G., and I. Vlahavas (2007), Random k-labelsets: An ensemble method for multi-label classification, in *Proceedings of the 18th European Conference on Machine Learning*, pp. 406–417.

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

Turnbull, D., L. Barrington, D. Torres, and G. Lanckriet (2008), Semantic annotation and retrieval of music and sound effects, *IEEE Transactions on Audio, Speech and Language Processing*, *16*(2), 467–476.

Tzanetakis, G., and P. Cook (2002), Musical genre glassification of audio signals, *IEEE transactions on audio, speech and language processing*, *10*(5), 293–302.

Ueda, N., and K. Saito (2003), Parametric mixture models for multi-labeled text, *Advances in Neural Information Processing Systems*, *15*, 721–728.

Van de Geer, S. (2008), High-dimensional generalized linear models and the lasso, *Annals of Statistics*, *36*(2), 614–645.

Van de Geer, S., and P. Bühlmann (2009), On the conditionsl used to prove oracle results for the lasso, *Electronic Journal of Statistics*, *3*, 1360 – 1392.

Wang, P., D. Chao, and L. Hsu (2011), Learning networks from high dimensional binary data: An application to genomic instability data, *Biometrics*, *67*(1), 164–173.

Witten, D. M., J. H. Friedman, and N. Simon (2011), New insights and faster computations for the graphical lasso, *Journal of Computational and Graphical Statistics*, *20*(4), 892–900.

Woods, J. (1978), Markov image modeling, *IEEE Transactions on Automatic Control*, *23*, 846–850.

Yang, Z., K. Streicher, M. Ray, J. Abrams, and S. Etheir (2006), Multiple interacting oncogenes on the 8p11-p12 amplicon in human breast cancer, *Cancer Research*, *66*, 11,632–11,634.

Yin, J., and H. Li (2011), A sparse conditional gaussian graphical model for analysis of genetical genomics data, *Annals of Applied Statistics*, *5*(4), 2630–2650.

Yuan, L., J. Liu, and J. Ye (2011), Efficient methods for overlapping group lasso, in *The Twenty-Fifth Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 352–360.

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

Yuan, M., and Y. Lin (2006), Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society, Series B*, *68*(1), 49–67.

Yuan, M., and Y. Lin (2007), Model selection and estimation in the Gaussian graphical model, *Biometrika*, *94*(1), 19–35.

Zhang, M., and Z. Zhou (2007), Ml-knn: A lazy learning approach to multi-label learning, *Pattern Recognition*, *40*, 2038–2048.

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

Zhao, P., and B. Yu (2006), On model selection consistency of lasso, *Journal of Machine Learning Research*, *7*, 2541 – 2567.

Zhu, S., X. Ji, W. Xu, and Y. Gong (2005), Multi-labelled classification using maximum entropy method, in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in Information Retrieval*.