Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

High-Dimensional Mixed Graphical Models

Jie Cheng, Tianxi Li, Elizaveta Levina & Ji Zhu

To cite this article: Jie Cheng, Tianxi Li, Elizaveta Levina & Ji Zhu (2017) High-Dimensional Mixed Graphical Models, Journal of Computational and Graphical Statistics, 26:2, 367-378, DOI: 10.1080/10618600.2016.1237362

To link to this article: https://doi.org/10.1080/10618600.2016.1237362

Accepted author version posted online: 22 Sep 2016.
Published online: 22 Sep 2016.

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2017, VOL. 26, NO. 2, 367–378
http://dx.doi.org/10.1080/10618600.2016.1237362

Taylor & Francis
Taylor & Francis Group

High-Dimensional Mixed Graphical Models

Jie Cheng[a], Tianxi Li[b], Elizaveta Levina[b], and Ji Zhu[b]

[a]Google, Inc., Mountain View, California; [b]Department of Statistics, University of Michigan, Ann Arbor, Michigan

ABSTRACT

While graphical models for continuous data (Gaussian graphical models) and discrete data (Ising models) have been extensively studied, there is little work on graphical models for datasets with both continuous and discrete variables (mixed data), which are common in many scientific applications. We propose a novel graphical model for mixed data, which is simple enough to be suitable for high-dimensional data, yet flexible enough to represent all possible graph structures. 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 lasso penalty for computational efficiency. We demonstrate the effectiveness of our method through an extensive simulation study and apply it to a music annotation dataset (CAL500), obtaining a sparse and interpretable graphical model relating the continuous features of the audio signal to binary variables such as genre, emotions, and usage associated with particular songs. While we focus on binary discrete variables for the main presentation, we also show that the proposed methodology can be easily extended to general discrete variables.

ARTICLE HISTORY
Received June 2014
Revised August 2016

KEYWORDS
Conditional Gaussian density; Graphical model; Group lasso; Mixed variables; Music annotation

1. Introduction

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 notable amount of attention over the past decade. In an undirected graphical model, nodes in the graph represent the variables, while an edge between a pair of variables indicates that they are dependent conditional on all other variables. The vast majority of the graphical models literature has been focusing on either the multivariate Gaussian model (see Meinshausen and Bühlmann 2006; Yuan and Lin 2007; d'Aspremont, Banerjee, and El Ghaoui 2008; Friedman, Hastie, and Tibshirani 2008; Rocha, Zhao, and Yu 2008; Rothman et al. 2008; Lam and Fan 2009; Peng et al. 2009; Ravikumar et al. 2009; Yuan 2010; Cai, Liu, and Luo 2011) or the Ising model for binary and discrete data (see Höfling and Tibshirani 2009; Ravikumar, Wainwright, and Lafferty 2010). The properties of these models are by now well understood and studied both in the classical and the high-dimensional settings. Both these models can only deal with variables of one kind—either all continuous variables in Gaussian models 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 article, we will focus on graphical models for this type of mixed data (mixed graphical models).

Sparse estimation of Gaussian graphical models using regularized maximum likelihood methods using ℓ_1 penalty on the precision matrix has received a lot of attention in recent years. Friedman, Hastie, and Tibshirani (2008) developed an efficient algorithm known as the graphical lasso, with excellent theoretical properties and fast implementations available, but its reliance on the assumption of normality can be restrictive in many real applications. Liu, Lafferty, and Wasserman (2009) relaxed this assumption to a Gaussian copula model they called nonparanormal. The authors assumed that there exist differentiable, monotone transformations $f = (f_1, f_2, \ldots, f_p)$ such that, $f(X) = (f_1(X_1), f_2(X_2), \ldots, f_p(X_p))$ is Gaussian with mean μ^f and precision matrix Ω^f. Then X_i and X_j are conditionally independent given the rest if and only if $\Omega^f_{ij} = 0$. The proposed algorithm estimates the marginal transformation functions nonparametrically and applies graphical lasso algorithm on the transformed data to estimate the underlying graphical structure. Liu et al. (2012) and Xue and Zou (2012) independently exploited the connection of nonparametric rank-based correlation estimators such as Spearman's rho (Spearman 1904) and Kendall's tau (Kendall 1938) with the nonparanormal covariance matrix to directly estimate Ω^f avoiding the plug-in estimator involving $\hat{f}_j$. Liu et al. (2012) established that their proposed estimator achieves optimal rates of convergence for graph recovery and parameter estimation, while Xue and Zou (2012) investigated the theoretical properties of rank-correlation-based algorithms including graphical lasso, Dantzig selector (Candes and Tao 2007), and CLIME (Cai, Liu, and Luo 2011). This line of work has clearly extended the scope of graphical models to

CONTACT Elizaveta Levina elevina@umich.edu
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JCGS.

general continuous data but they are not suitable for binary variables and therefore cannot handle mixed type of variables, which is the main setting of this article.

For binary data, most of the literature has focused on the Ising model (Ising 1925), originally proposed in statistical physics. For high-dimensional data, the Ising model becomes computationally challenging due to the intractability of the log partition function. Ravikumar, Wainwright, and Lafferty (2010) used an ℓ_1-penalized pseudo-likelihood method to estimate the edge set, in the spirit of the neighborhood selection method proposed by Meinshausen and Bühlmann (2006) for continuous data. More recently, Xue, Zou, and Cai (2012) proposed using a SCAD penalty (Fan and Li 2001) to estimate a sparse graphical model; they developed scalable and efficient algorithms for optimizing the nonconcave problem and proved theoretical performance guarantees superior to concave penalties such as the lasso.

For mixed data containing both continuous and discrete variables, the conditional Gaussian distribution (Lauritzen and Wermuth 1989; Lauritzen 1996) has become the foundation of most developments on this topic. In the original article, Lauritzen and Wermuth (1989) defined a general form of the conditional Gaussian density and characterized the connection between the model parameters and the conditional associations among the variables. The model is fitted via the maximum likelihood approach. The number of parameters in this model, however, grows exponentially with the number of variables, which renders it unsuitable for high-dimensional problems arising in many modern applications. Edwards (1990) generalized the conditional Gaussian distribution model to the hierarchical interaction model, which can account for all possible hierarchical/nested interactions between the discrete and continuous variables, and proposed methods of maximum likelihood estimation under these models using marginal mean and covariance calculations. They also showed some connection of hierarchical interaction models with MANOVA type of models. Much more recently, Lee and Hastie (2015) and Fellinghauer et al. (2013) have studied the mixed graphical model (simultaneously and independently of the present article), under a setting that could be viewed as a simplified special case of our proposal. Edwards, De Abreu, and Labouriau (2010) also proposed an extended algorithm based on the Chow–Liu algorithm (Chow and Liu 1968) for the multivariate discrete case to fit high-dimensional mixed graphical models. A more detailed discussion of these articles is postponed to Section 6.

In this article, 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 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 due to the overlaps between groups, 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.

We start with a brief introduction to 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, $Y = (Y_\gamma)_{\gamma\in\Gamma}$ is a p-dimensional continuous sub-vector, and Δ and Γ are index sets for Z and Y, respectively. The conditional Gaussian density $f(x)$ is defined as

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

where $\{(g_z, h_z, K_z), g_z \in \mathbb{R}, h_z \in \mathbb{R}^p, K_z \in \mathbb{R}^+_{p\times p}\}$ are the canonical parameters of the distribution. The following equations connect the canonical parameters in (1) to the moments of Y and Z:

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

Also, $\mathcal{L}(Y|Z = z) = \mathcal{N}(\xi_z, \Sigma_z)$, so conditional on $Z = z$, each Y is normally distributed with the mean and variance determined by 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 1. Lauritzen and Wermuth (1989) represented the canonical parameters from (1) by the following expansions,

$$g_z = \sum_{d:d\subseteq\Delta} \lambda_d(z), \quad h_z = \sum_{d:d\subseteq\Delta} \eta_d(z), \quad K_z = \sum_{d:d\subseteq\Delta} \Phi_d(z), \tag{3}$$

where functions indexed by the index set 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

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

where $\eta_d^\gamma(z)$ is the γth 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 article is organized as follows. Section 2 introduces the simplified mixed graphical model that has just enough parameters to cover all possible graph structures and proposes an efficient estimation algorithm for the model. Section 3 uses several sets of simulation studies to evaluate the model selection performance and compare to some alternative methods for graph estimation. In Section 4, the proposed model is applied to a music annotation dataset *CAL500* with binary labels and continuous audio features. In Section 5, we describe the generalization of the model from binary to discrete variables. Finally, we conclude in Section 6 with a discussion.

2. Methodology

We propose a simplified but flexible version of the conditional Gaussian model for mixed data. The 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 overlapping groups of parameters. The natural solution to the problem 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 weighted ℓ_1 penalty.

2.1. The Simplified Mixed Graphical Model

Without loss of generality, we partition the random vector $X = (Z_1, Z_2, \ldots, Z_q, Y_1, Y_2, \ldots, Y_p)$ into the binary part with $\Delta = \{1, 2, \ldots, q\}$ and the continuous part with $\Gamma = \{1, 2, \ldots, p\}$. We propose to consider the conditional Gaussian distribution with the density function

$$\begin{aligned}\log f(z, y) &= \sum_{d:d\subseteq\Delta,|d|\le 2} \lambda_d(z) + \sum_{d:d\subseteq\Delta,|d|\le 1} \eta_d(z)^T y \\ &\quad - \frac{1}{2} \sum_{d:d\subseteq\Delta,|d|\le 1} y^T \Phi_d(z) y \\ &= \left(\lambda_0 + \sum_j \lambda_j z_j + \sum_{j>k} \lambda_{jk} z_j z_k\right) \\ &\quad + y^T \left(\eta_0 + \sum_j \eta_j z_j\right) - \frac{1}{2} y^T \left(\Phi_0 + \sum_{j=1}^{q} \Phi_j z_j\right) y \\ &= \left(\lambda_0 + \sum_j \lambda_j z_j + \sum_{j>k} \lambda_{jk} z_j z_k\right) \\ &\quad + \sum_{\gamma=1}^{p} \left(\eta_0^\gamma + \sum_j \eta_j^\gamma z_j\right) y_\gamma \\ &\quad - \frac{1}{2} \sum_{\gamma,\mu=1}^{p} \left(\Phi_0^{\gamma\mu} + \sum_{j=1}^{q} \Phi_j^{\gamma\mu} z_j\right) y_\gamma y_\mu \,, \end{aligned} \tag{4}$$

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,

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

Note that the density is explicitly defined via the expanded terms in (3) and 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 (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 arbitrary dependence on the binary variables. These simplifications reduce the total number of parameters from $\mathcal{O}(p^2 2^{(p+q)})$ in the full model to $\mathcal{O}(\max(q^2, p^2 q))$. This reduction is necessary especially in the high-dimensional setting, where there are limited number of samples and even if the true model involves higher order interactions, it may not be possible to estimate them well due to the bias-variance tradeoff. 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. 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, Wainwright, and Lafferty (2010) for the Ising model. To describe 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

$$\begin{aligned}\log \frac{P(Z_j = 1|Z_{-j}, Y)}{P(Z_j = 0|Z_{-j}, Y)} &= \lambda_j + \sum_{k\ne j} \lambda_{jk} Z_k + \sum_{\gamma=1}^{p} \eta_j^\gamma Y_\gamma \\ &\quad - \frac{1}{2} \sum_{\gamma,\mu=1}^{p} \Phi_j^{\gamma\mu} Y_\gamma Y_\mu. \end{aligned} \tag{5}$$

Since the conditional log-odds in (5) 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

$$\begin{aligned}Y_\gamma &= \frac{1}{K_z^{\gamma\gamma}} \left(\eta_0^\gamma + \sum_j \eta_j^\gamma Z_j - \sum_{\mu\ne\gamma} \left(\Phi_0^{\gamma\mu} + \sum_j \Phi_j^{\gamma\mu} Z_j\right) Y_\mu\right) \\ &\quad + e_\gamma,\end{aligned}$$

where $e_\gamma \sim \mathcal{N}(0, (K_z^{\gamma\gamma})^{-1})$. With $\mathrm{diag}(\Phi_j) = 0$ as defined by (4), we have $K_z^{\gamma\gamma} = \Phi_0^{\gamma\gamma}$, that is, the conditional variance of Y_γ does not depend on Z. Rewrite

$$Y_\gamma = \tilde{\eta}_0^\gamma + \sum_j \tilde{\eta}_j^\gamma Z_j - \sum_{\mu\ne\gamma} \left(\tilde{\Phi}_0^{\gamma\mu} + \sum_j \tilde{\Phi}_j^{\gamma\mu} Z_j\right) Y_\mu + e_\gamma, \tag{6}$$

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.

2.3. Regularization

Based on Theorem 1, the following equivalences hold:

$$\begin{aligned} &Z_j \perp Z_k \mid X\backslash\{Z_j, Z_k\} \Longleftrightarrow \lambda_{jk} = 0, \\ &Z_j \perp Y_\gamma \mid X\backslash\{Z_j, Y_\gamma\} \Longleftrightarrow \boldsymbol{\theta}_{j\gamma} = \left(\eta_j^\gamma, \{\Phi_j^{\gamma\mu} : \mu \in \Gamma\backslash\{\gamma\}\}\right) = 0, \\ &Y_\gamma \perp Y_\mu \mid X\backslash\{Y_\gamma, Y_\mu\} \Longleftrightarrow \boldsymbol{\theta}_{\gamma\mu} = \left(\Phi_0^{\gamma\mu}, \{\Phi_j^{\gamma\mu} : j \in \Delta\}\right) = 0. \end{aligned} \tag{7}$$

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, we add the $\ell_1\backslash\ell_2$ penalty, proposed by Yuan and Lin (2006) for group lasso, to the loss function in each regression. The groups are predetermined 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

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

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

$$\text{Logistic regression: } \min\ \ell_j + \rho\left(\kappa\sum_{k\neq j}|\lambda_{jk}| + \sum_{\gamma=1}^{p}\|\boldsymbol{\theta}_{j\gamma}\|_2\right), \tag{8}$$

$$\text{Linear regression: } \min\ \ell_\gamma + \rho\left(\sum_{\mu\neq\gamma}\|\tilde{\boldsymbol{\theta}}_{\gamma\mu}\|_2 + \sum_{j=1}^{q}\|\tilde{\boldsymbol{\theta}}_{j\gamma}\|_2\right), \tag{9}$$

where ρ and κ are tuning parameters. Using two tuning parameters, ρ and κ, allows us to penalize individual parameters and groups of parameters differently, essentially allowing the edges between binary variables to be penalized differently from other edges. While in principle both parameters can be tuned, in simulations we got good and very stable results over the range $0.1 \leq \kappa \leq 0.5$, and thus in simulations we set $\kappa = 0.1$. Note that we use the same tuning parameter ρ for both linear and logistic regressions. One reason to use a single tuning parameter ρ is to simplify the treatment of overlapping groups of parameters from different regressions (see more on this below). We did conduct simulation experiments using two different tuning parameters, ρ_1 for linear regression and ρ_2 for logistic regression, and the results (not shown in this article) are similar to simply using one tuning parameter ρ for both linear and logistic regressions. In principle, one could also tune each regression separately, but the computational cost is prohibitive; further the estimation variance can be extraordinarily high when using a large number of tuning parameters, see, for example, Meinshausen and Bühlmann (2006) and Ravikumar, Wainwright, and Lafferty (2010) for similar neighborhood selection settings where a single tuning parameter was used for different regressions. Finally, note that in linear regression, the parameters in (6) 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 (8) and (9) appear to be group lasso regressions, they cannot be solved by regular group lasso algorithms, because the groups of parameters involved in each regression overlap. Specifically, in logistic regression, parameter $\Phi_j^{\gamma\mu}$ is part of both $\boldsymbol{\theta}_{j\gamma}$ and $\boldsymbol{\theta}_{j\mu}$ and affects 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 was theoretically investigated by Jenatton, Audibert, and Bach (2011) (see also Jacob, Obozinski, and Vert 2009) but has received limited attention from a computational point of view. Yuan, Liu, and Ye (2013) recently proposed an algorithm for solving the overlapping group lasso problem but the speed is still an issue that limits its capability of handling high-dimensional data. Therefore, we took the approach of finding a surrogate for the overlapping group lasso penalty to make the problem both efficient to solve and suitable for high-dimensional settings without losing much accuracy.

Instead of the overlapping group lasso penalty, we propose to use its upper bound as a surrogate, which is essentially a weighted ℓ_1 penalty. The upper bound results from the fact that for any vector $\boldsymbol{b}$, $\|\boldsymbol{b}\|_2 \leq \|\boldsymbol{b}\|_1$. Take the logistic regression (8), for example, $\sum_{k\neq j}|\lambda_{jk}| + \sum_{\gamma=1}^{p}\|\boldsymbol{\theta}_{j\gamma}\|_2 \leq \sum_{k\neq j}|\lambda_{jk}| + \sum_{\gamma=1}^{p}|\eta_j^\gamma| + 2\sum_{\gamma<\mu}|\Phi_j^{\gamma\mu}|$. The surrogate on the right penalizes the overlapped parameters twice as much as the other parameters, which makes intuitive sense since incorrectly identifying the overlapped parameters as nonzero will result in two wrong edges, while the incorrect unique parameters for each group will only cause one wrong edge.

To illustrate the upper bound geometrically, we show a toy example. Suppose the parameter vector is $\boldsymbol{b} = (b_1, b_2, b_3)$, and two groups are $\mathcal{G}_1 = (b_1, b_2)$ and $\mathcal{G}_2 = (b_2, b_3)$. The optimization problem for the overlapping group lasso penalty and its ℓ_1 surrogate boils down to optimizing the same loss function over different feasible regions (for the same tuning parameter). Figure 1 compares the 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-square loss are smooth

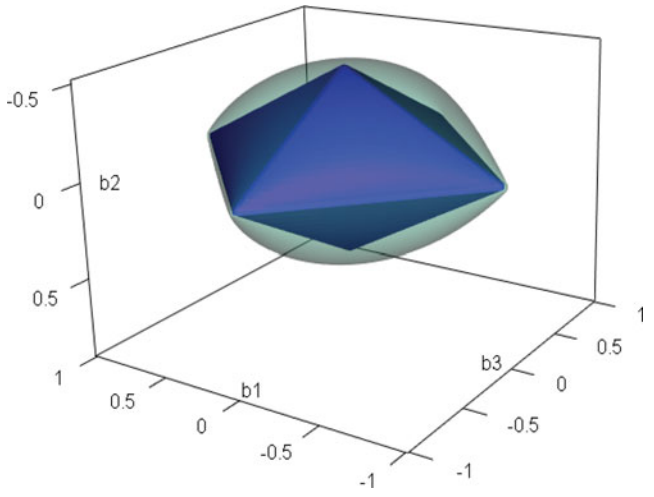


Figure 1. Green (outside): $\{\boldsymbol{b} : \sqrt{b_1^2 + b_2^2} + \sqrt{b_3^2 + b_2^2} = 1\}$; Blue (inside): $\{\boldsymbol{b} : |b_1| + |b_3| + 2|b_2| = 1\}$.

convex functions, their optima are likely to occur at singular points of the feasible region. Note that $\mathcal{R}_2$ is not only a subset of $\mathcal{R}_1$ but it contains all four singular points of $\mathcal{R}_1$: $(\pm 1, 0, 0)$, $(0, 0, \pm 1)$. Thus for this example, it is guaranteed that all the optimal points of singular points on $\mathcal{R}_1$ for the overlapping group lasso penalty are also optimal points for its surrogate. The effectiveness of this approximation will be further demonstrated by a simulation study in Section 3.

With the penalty being replaced by the weighted ℓ_1 surrogate, we solve the following regression problems separately as an approximation to the original problems (8) and (9) to obtain the parameter estimates.

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

$$\min \ \ell_j + \rho \left(\kappa \sum_{k \neq j} |\lambda_{jk}| + \sum_{\gamma=1}^{p} |\eta_j^{\gamma}| + 2 \sum_{\gamma < \mu} |\Phi_j^{\gamma\mu}| \right). \quad (10)$$

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

$$\min \ \ell_\gamma + \rho \left(\sum_{j=1}^{q} |\tilde{\eta}_j^{\gamma}| + \sum_{\mu \neq \gamma} |\tilde{\Phi}_0^{\gamma\mu}| + 2 \sum_{j=1}^{q} \sum_{\mu \neq \gamma} |\tilde{\Phi}_j^{\gamma\mu}| \right). \quad (11)$$

Since we are estimating parameters in separate regressions, all parameters determining edges will be estimated at least twice. This situation is common in 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. Here, we chose taking the maximum of absolute values as the final estimate, based on simulations studies (not shown), which resulted in the maximum giving better model selection results than the minimum. To fit both types of regressions with a weighted ℓ_1 penalty, we used the Matlab package *glmnet* of Friedman, Hastie, and Tibshirani (2010).

3. Numerical Performance Evaluation

In this section, we first demonstrate the effectiveness of the weighted lasso approximation to the overlapping group lasso. Then we show simulation results regarding model selection performance under different settings and comparison with other graph selection methods (Fellinghauer et al. 2013; Lee and Hastie 2015). The results for graph selection are summarized in ROC curves, where we plot the true positive rate (TPR) against the false positive rate (FPR), for both parameters and edges across a fine grid of tuning parameters. All ROC curves are obtained based on 100 replications with local smoothing. 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 of interest are defined as

$$\text{TP} = \#\{j : \hat{\theta}_j \neq 0 \text{ and } \theta_j \neq 0\},$$

$$\text{FP} = \#\{j : \hat{\theta}_j \neq 0 \text{ and } \theta_j = 0\},$$

$$\text{TPR} = \frac{\text{TP}}{\#\{j : \theta_j \neq 0\}}, \qquad \text{FPR} = \frac{\text{FP}}{\#\{j : \theta_j = 0\}}.$$

The quantities based on the true edge set $\boldsymbol{E}$ and the estimated edge set $\hat{\boldsymbol{E}}$ can be defined in a similar fashion.

3.1. Weighted Lasso Approximation to the Overlapping Group Lasso

We first briefly discuss the weighted lasso approximation to the overlapping group lasso penalty in the setting of linear regression, since this approximation itself is not limited to graphical models. One of the most frequent uses of group lasso in regression is to encourage a hierarchical variable selection path (Zhao, Rocha, and Yu 2009), to ensure main effects are selected before the corresponding interactions are included. Zhao, Rocha, and Yu (2009) designed an example to investigate this property, and here we use the same setting to compare the proposed weighted lasso approximation as well as the regular lasso to the overlapping group penalty. Following Zhao, Rocha, and Yu (2009) exactly, we have variables $x_1, \ldots, x_{10}$ and all of their pairwise products $x_i x_j$, $1 \leq i < j \leq 10$ in the model, resulting in 55 variables. The variables are generated from a standard normal distribution; the coefficients of the first four variables are 7, 2, 1, 1, and all main effects and interaction effects involving the other six are zero. The response follows the model $Y = X\beta + \epsilon$, where $\epsilon \sim N(0, 3.7I)$, and the sample size $n = 121$. We consider three settings of Zhao, Rocha, and Yu (2009) for the interaction effects between $x_1, \ldots, x_4$, shown in Table 1, corresponding to weak, moderate, and strong interaction effects; we omit their two other cases to save space (the results are similar). The groups we use include $\{x_i, x_i x_j, j \neq i\}$, $i = 1, \ldots, 10$ and singleton groups with only one interaction term $\{x_i x_j\}$, $i \neq j$. So there are 10 groups of size 10 and 45 groups of size 1.

Table 1. Three configurations of interaction effects (weak, moderate, and strong) from Zhao, Rocha, and Yu (2009).

	x_1x_2	x_1x_3	x_1x_4	x_2x_3	x_2x_4	x_3x_4
Weak	1	0	0	0.5	0.4	0.1
Moderate	5	0	0	4	2	0
Strong	7	7	7	2	2	1

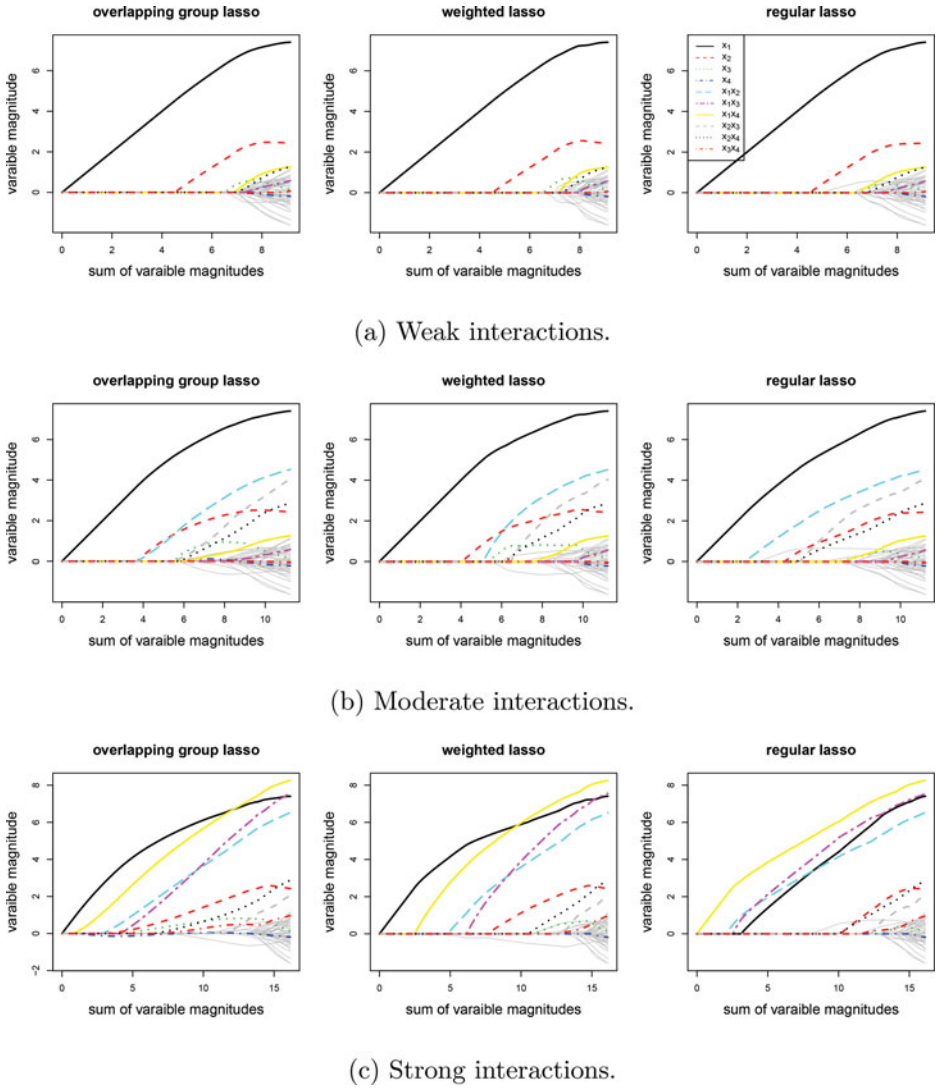


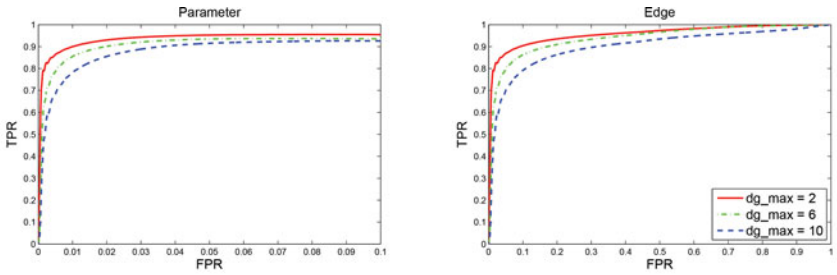
Figure 2. Variable selection paths of overlapping group lasso, weighted lasso, and regular lasso. The x-axis is $\sum_i |\beta_i|$ and the y-axis is the value of the coefficients.

Figure 2 shows the variable selection path for overlapping group lasso, weighted lasso, and regular lasso for the three settings. We focus on the four active variables and their pairwise interactions, shown in bold colored curves in Figure 2; the remaining 45 coefficient paths are shown in thin gray curves. For weak interactions, the three methods give nearly identical results. This is expected since the interaction effects are too weak to make any difference and without the interaction effects, the three penalties are the same. As the interaction effects become stronger, the difference becomes clear. Note that the nature of the overlapping group lasso may let a subset of variables enter the model simultaneously, but this is typically not true for weighted lasso. Thus, it is reasonable to treat the lasso approximation as correct as long as all the variables in such a subset enter the model before any others. For moderate effects, the weighted lasso gives the same variable selection as the overlapping group lasso as well. For strong effects, weighted lasso makes mistakes on the two weakest main effects, x_3 and x_4 (the green curve) and x_4 (the blue dotted curve), which remain close to zero along the entire group lasso path. The regular lasso differs from the overlapping group lasso on x_3 for moderate interactions, and makes even more mistakes for strong interactions, missing even the strongest effect x_1 (black curve).

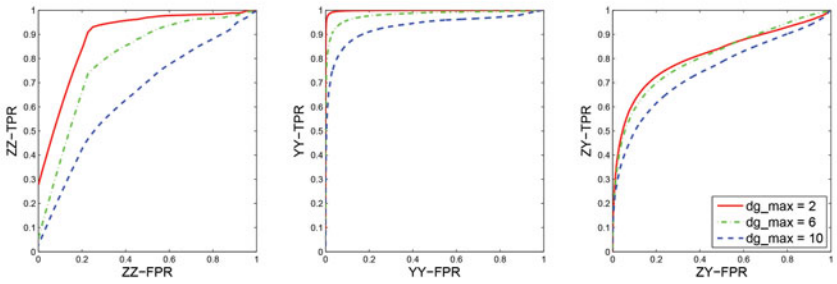
Overall, in this example the weighted lasso approximates the overlapping group lasso well, and much better than regular lasso, except when the interaction effects are weak. In general settings, the quality of the weighted lasso approximation to the overlapping group lasso can depend on many factors, such as the pattern and degree of overlap between groups, the signal-to-noise ratio, the degree of correlation between predictors, etc. Further investigation of this topic in the context of regression is beyond the scope of this article and is left for future work.

3.2. Parameter Estimation and Edge Identification

To start, we investigate the impact of heterogeneity of node degrees (i.e., the number of edges connected to a node) of the underlying graph on the performance of the proposed method, as it is generally a challenge for graphical models. We set the first $q = 10$ variables to be binary and the remaining $p = 90$ variables to be continuous, with the sample size $n = 100$. We first vary the maximum node degree by setting it to be 2, 6, and 10 in the graph while maintaining the total number of edges fixed



(a) ROC curves for parameters and edges with varying maximum node degree.



(b) ROC curves for each edge category with varying maximum node degree.

Figure 3. Model selection results.

at 80. The smaller the maximum node degree is, with fixed total number of edges, the more homogenous the degree distribution.

We generate the graph using the Erdös–Renyi model, and simply use rejection sampling to enforce the constraints, that is, we keep generating the graph until the maximum node degree meets the requirement. The edges of the graph are of three types: edges connecting binary variables (*ZZ*), edges connecting continuous variables (*YY*), and edges connecting binary and continuous variables (*ZY*). To be able to compute the true positive rates for each category, we further require the graph have at least one edge in each category.

Once the graph is fixed, we set all parameters corresponding to absent edges to 0. For the nonzero parameters, we set $\{\lambda_j, \lambda_{jk}, \eta_j\}$ to be positive or negative with equal probability and the absolute value of each nonzero η_j is drawn from the uniform distribution on the interval $(0.9a, 1.1a)$ and each nonzero λ_j or λ_{jk} is from $(0.9c, 1.1c)$. For $\{\Phi_0, \Phi_j\}$, we set the off-diagonal elements to be positive or negative with equal probability, and the absolute value of each nonzero parameter is also drawn from a uniform distribution, on the interval $(0.9b, 1.1b)$. The parameters a, b, and c control the overall magnitude of the nonzero parameters and therefore the effective signal-to-noise ratio; we set $a = c = 1$ and $b = 2$. For the purpose of investigating the effect of the maximum node degree, varying the values of a, b, and c does not result in a qualitative difference in the results. 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). Since we use the exact probability rather than Markov chain Monte Carlo (MCMC) methods to generate the binary variable, the memory requirements for the distribution P_z makes it difficult to generate a large number of binary variables in simulations. However, this is not a problem for real data where the variables are already observed, and in fact our data analysis later in the article demonstrates the method works well with large q. Finally, for each $\boldsymbol{z}_i$, we generate the continuous part $\boldsymbol{y}_i$ from a multivariate Gaussian distribution with mean ξ_{z_i} and covariance Σ_{z_i} defined by (2).

Figure 3(a) shows the impact of maximum node degree. The more homogenous the node degrees are, the easier the model selection task. This is because without prior information, the proposed method treats all nodes equally and uses the same tuning parameter for all regressions. We also report ROC curves for each of the three edge types (*ZZ, YY, ZY*), shown in Figure 3(b). The pattern in each category is qualitatively consistent with the overall pattern in Figure 3(a), and it appears that accuracy on the edges between binary variables (*ZZ*) suffers the most from increased degree heterogeneity.

3.3. Comparison With Other Graphical Model Methods

Here, we compare the proposed method with several other penalized regression approaches. Fellinghauer et al. (2013) proposed to fit separate ℓ_1 regularized regressions by regressing each variable on the others, without including any interaction terms. This is a special case of our model. Lee and Hastie (2015) fit the same model as Fellinghauer et al. (2013) (no interaction terms) by maximizing a joint pseudo-likelihood instead of fitting separate regressions, and also applies calibration to adjust the penalty weights. We also include a comparison to our model (with interaction terms) penalized by the regular lasso penalty instead of the weighted lasso penalty, effectively replacing the weight of 2 in front of the interaction terms with 1 while keeping everything else the same. We implemented our method and the method of Fellinghauer et al. (2013) using the glmnet package in Matlab. The method of Lee and Hastie (2015) is based on the matlab code provided by the authors. The computational cost of our method is about the same as that of Fellinghauer et al.

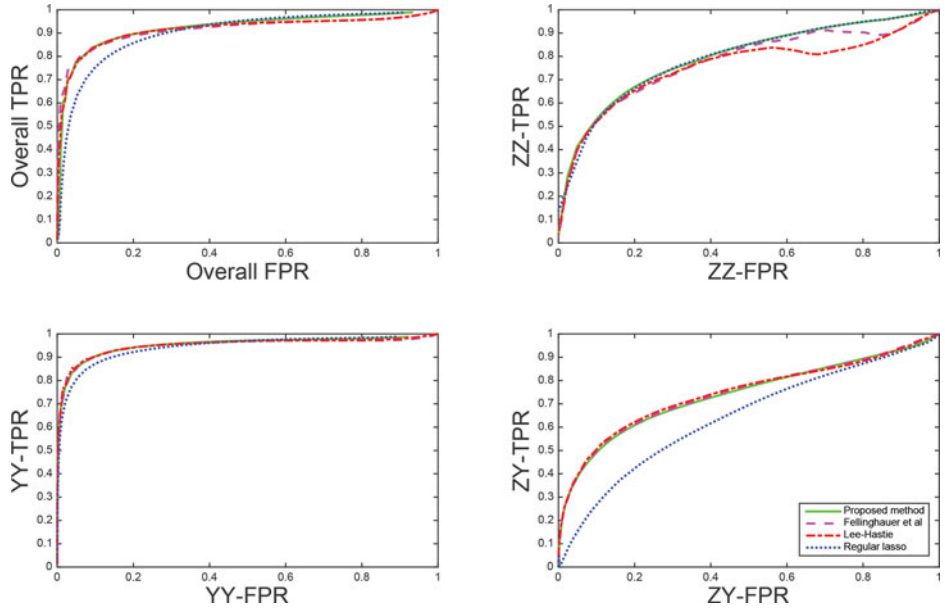


Figure 4. Edge-based ROC curves for three graphical model methods when there are only main effects in the true model.

(2013) and hundreds of times lower than that of Lee and Hastie (2015).

We consider two simulation settings. In the first setting, we set all $\Phi_j^{\gamma\mu}$, $j = 1, \ldots, q$; $\gamma, \mu = 1, \ldots, p$ parameters to zero. Thus, the true model is exactly what Fellinghauer et al. (2013) and Lee and Hastie (2015) assume (no interaction terms in regressions (5) and (6)), and each edge is 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 $\Phi_0^{\gamma\mu}$. We follow the set-up of the simulation in Section 3.2, setting the maximum node degree to 6, the total number of edges to 125, the number of variables to $p = 90$ continuous and $q = 10$ categorical, the sample size to $n = 100$, and $a = c = 1$ and $b = 2$.

Figure 4 shows that both Fellinghauer et al. (2013) and Lee and Hastie (2015) perform well, as is to be expected, since both Fellinghauer et al. (2013) and Lee and Hastie (2015) assumed the true model. Our method performs as well as Fellinghauer et al. (2013) and Lee and Hastie (2015), meaning that it is able to learn that the interaction terms are irrelevant and recover the true model with main effects only. The regular lasso penalty, on the other hand, is inferior in this case, doing similarly on estimating ZZ edges but worse on estimating YY and ZY edges. This likely happens because it penalized the interaction terms less than our method with the weighted lasso, and thus is not able to eliminate them as effectively.

In the second simulation setting, we allow for nonzero $\Phi_j^{\gamma\mu}$ parameters, keeping the dimensions $p = 90$ and $q = 10$ and the sample size $n = 100$ fixed. Note that, this corresponds to a graph that is sparse overall but has dense small locally dense subgraphs ($\Phi_j^{\gamma\mu} \neq 0$ indicates a YZY clique). We first randomly generate 40 edges in the same way as in the first setting. Then we set $\{z_1, \ldots, z_4\}$, $\{z_8, \ldots, z_{10}, y_1, \ldots, y_6\}$, and $\{y_{11}, \ldots, y_{20}\}$ to be the three complete subgraphs, and there are no other edges in the graph. The resulting graph has 127 edges, which is similar as before. We also set the corresponding main and interaction effects in (5) and (6) to be nonzero. Here we set $a = 0.1$, $b = 0.2$, $c = 0.6$ to obtain a reasonable signal-to-noise ratio and make sure the problem is neither impossible nor trivial.

Figure 5 shows the results. Since in this case Fellinghauer et al. (2013) and Lee and Hastie (2015) assumed a wrong model, the comparison of the ROC curves for parameter identification is automatically biased in our favor; instead, we show the ROC curves for edge identification only. As expected, when interaction terms are present in the true model, our method outperforms both Fellinghauer et al. (2013) and Lee and Hastie (2015) on the overall ROC curve. Decomposing the overall ROC curves into the three subtypes, we sees that there is no difference between the three methods on the YY edges. This may be because these edges involve only the continuous variables, and the Gaussian graphical model is often easier to estimate than the Ising model. For edges involving discrete variables, which are more difficult to identify, our method performs much better than Fellinghauer et al. (2013) and Lee and Hastie (2015) on ZY edges and somewhat better on the ZZ edges. The regular lasso, which, like our method, fits the true model here, gives results similar to our method on ZZ and YY edges, and worse results on the ZY edges, resulting in a somewhat worse overall ROC curve.

Overall, we observe that if the underlying model does not contain any interaction terms, both Fellinghauer et al. (2013) and Lee and Hastie (2015) perform well by not including them, and our model combined with the weighted lasso penalty does equally well by estimating these interactions to be 0. When the true model does contain interaction terms, our method performs much better than Fellinghauer et al. (2013) and Lee and Hastie (2015) in terms of edge identification. If we use our model with the regular lasso penalty instead of the weighted penalty, it performs a little worse when interaction terms are present but can be much worse when there are no interactions.

4. Application to Music Annotation

Music annotation uses techniques from several disciplines, including audio signal processing, information retrieval,

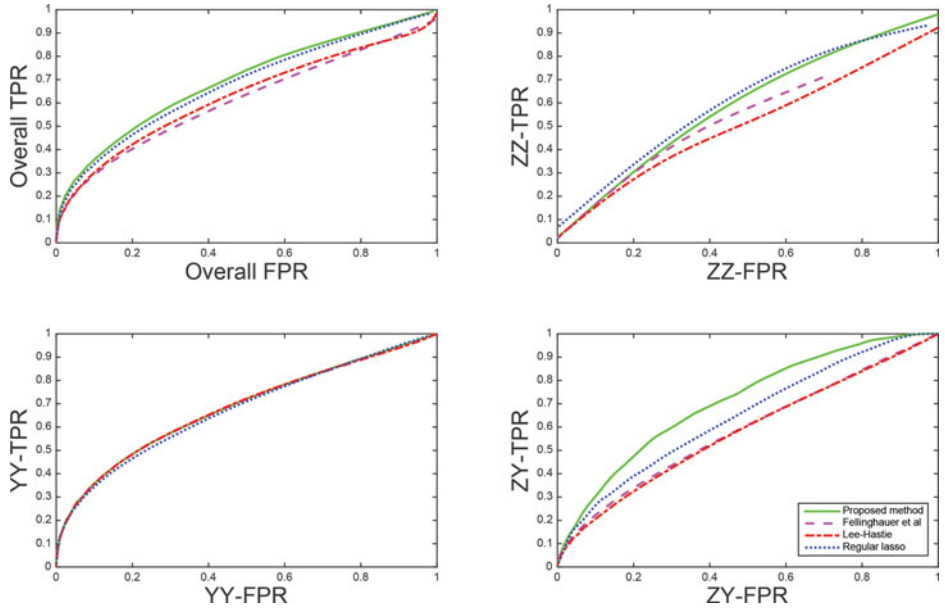


Figure 5. Edge-based ROC curves for three graphical model methods when there are both main effects and interactions in the true model.

multi-label classification, and others. Music annotation datasets usually consist of two parts: "labels," typically assigned by human experts, contain the categorical semantic description 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 graphical model 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 from the audio signal are associated with particular music genres, or emotions perceived to be conveyed by the music. We apply our method to the publicly available music annotation dataset *CAL500* (Turnbull et al. 2008) from the *http://mulan.sourceforge.net/index.html* Mulan database (Tsoumakas et al. 2011) to find the conditional dependence patterns among these mixed variables.

CAL500 dataset consists of 502 popular music tracks (including songs with English lyrics 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 obtained from 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. These 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 are based on the short time Fourier transform (STFT) and are calculated for each short time window by sliding a half-overlapping, 23 msec 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 23 msec, the following continuous features are extracted to represent the audio file: *zero crossings*, a measure of the noisiness of the signal; *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; 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 23 msec short frames are then grouped into 1 sec long texture windows, based on which the following summary statistics for the four features defined above were calculated and used as the final continuous variables: 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 that were assigned to less than 3% of the songs. Also, we standardized the continuous variables. This resulted in a dataset with $n = 502$ observations, $q = 118$ discrete variables, and $p = 16$ continuous variables.

We applied our method coupled with stability selection (Meinshausen and Bühlmann 2010) to identify the underlying graph for the purpose of exploratory data analysis, which is the primary usage of graphical models. Stability selection was implemented by running the algorithm on 100 randomly drawn sub-samples of size $n/2$ for a grid of (ρ, κ) values, and only keeping the edges that were selected at least 99% of the time for a given value of (ρ, κ). The results are shown in Figure 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. 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, the genre "likable or popular songs" (circle 84) is associated with "catchy and memorable" (circle 74), "would like to recommend" (circle 89), the usage "reading" (circle 107), and the emotion "pleasant and comfortable" (circle 25). Whether or not a song is "very danceable" (circle 97, 98) is connected to "fast tempo" (circle 78), the usage "at a party" (circle 99), and the emotion "light and playful" (circle 21). We also find

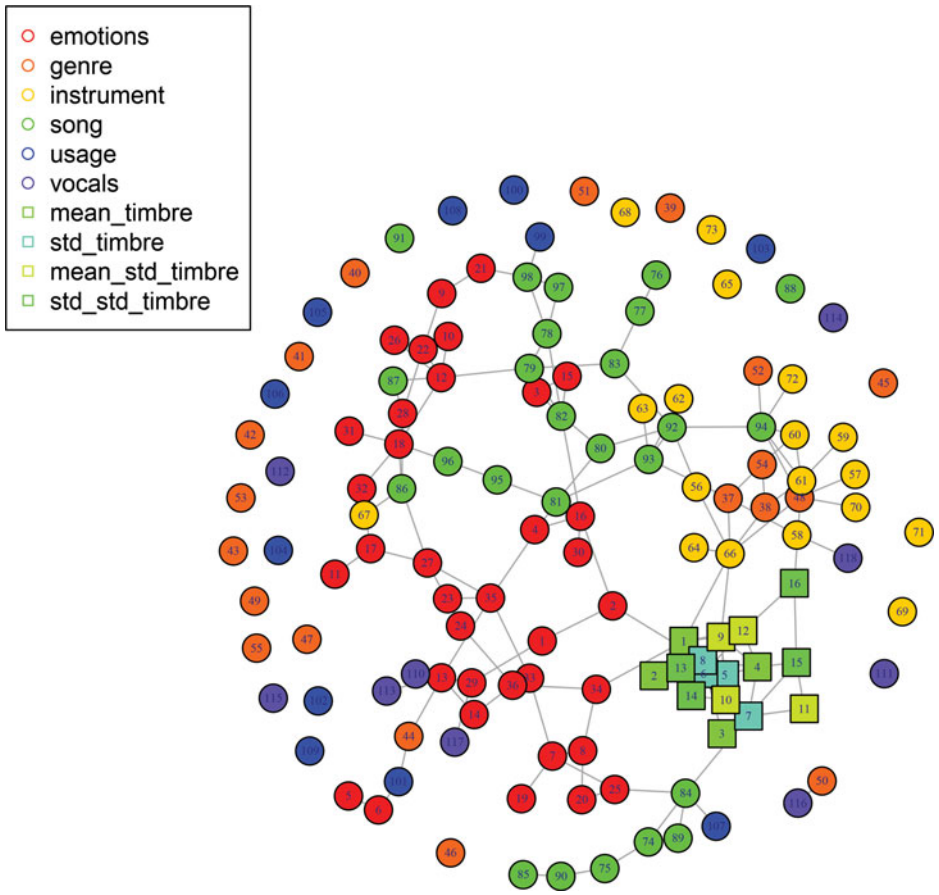


Figure 6. Estimated graphical model for CAL500 music data (edges with stability selection frequency of at least 99).

connections between the instrument "piano" (circle 67), and "positive feelings" (circle 86), "not sad" (circle 32), and "happy" (circle 17). The continuous variables that represent the audio signal features are quite densely connected within themselves, which is expected. Edges connecting continuous and binary variables may also be interesting. For instance, the average noisiness of the music (square 1) is connected with emotions that are not tender or soft (circle 34) and emotions that are not angry or aggressive (circle 2) as well as with "male lead vocals" (circle 66).

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) into the following,

$$\begin{aligned}\log f(z,y) &= \sum_{d:d\subseteq\Delta,|d|\le 2}\lambda_d(z) + \sum_{d:d\subseteq\Delta,|d|\le 1}\eta_d(z)^T y \\ &\quad - \frac{1}{2}\sum_{d:d\subseteq\Delta,|d|\le 1} y^T\Phi_d(z)y, \\ &= \left(\lambda_0 + \sum_{j=1}^{q}\lambda_j(z_j) + \sum_{j>k}\lambda_{jk}(z_j,z_k)\right) \\ &\quad + \sum_{\gamma=1}^{p}\left(\eta_0^\gamma + \sum_{j=1}^{q}\eta_j^\gamma(z_j)\right)y_\gamma, \\ &\quad - \frac{1}{2}\sum_{\gamma,\mu=1}^{p}\left(\Phi_0^{\gamma\mu} + \sum_{j=1}^{q}\Phi_j^{\gamma\mu}(z_j)\right)y_\gamma y_\mu,\end{aligned} \tag{12}$$

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 that 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

$$\begin{aligned}&Z_j \perp Z_k \mid X\backslash\{Z_j,Z_k\} \iff \boldsymbol{\theta}_{jk} = \big(\lambda_{jk}(z_j,z_k)\big) = 0, \\ &Z_j \perp Y_\gamma \mid X\backslash\{Z_j,Y_\gamma\} \iff \boldsymbol{\theta}_{j\gamma} \\ &\quad = \left(\eta_j^\gamma(z_j), \{\Phi_j^{\gamma\mu}(z_j) : \mu\in\Gamma\backslash\{\gamma\}\}\right) = 0, \\ &Y_\gamma \perp Y_\mu \mid X\backslash\{Y_\gamma,Y_\mu\} \iff \boldsymbol{\theta}_{\gamma\mu} \\ &\quad = \left(\Phi_0^{\gamma\mu}, \{\Phi_j^{\gamma\mu}(z_j) : j\in\Delta\}\right) = 0.\end{aligned} \tag{13}$$

The generalized model can be fitted with separate regressions based on the conditional likelihood of each variable. The parameters in (13) still have a group structure, which calls for using the

group lasso penalty as in (8) and (9). The structure of overlaps is more complex in this case, and we use the upper bound ℓ_1 approximation as in (10) and (11) to obtain the final estimates. Specifically, we minimize the following criteria separately:

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

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

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

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

Yuan and Lin (2006) proposed further adjusting the weights in the group lasso penalty for categorical variables to reflect its number of levels, which can carry over to our proposed weighted lasso approximation.

6. Discussion

In this article, we have proposed a new graphical model for mixed (continuous and discrete) data, which is particularly suitable for high-dimensional settings. As discussed in the introduction, while the general conditional Gaussian model is well known and 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. Two recent new developments on this topic, Fellinghauer et al. (2013) and Lee and Hastie (2015), were derived in parallel with and independently of this article. Both Fellinghauer et al. (2013) and Lee and Hastie (2015) assume a more restricted version of the conditional Gaussian density by assuming constant conditional covariance for all the continuous variables and is thus a special case of our model (4), where all the Φ_j are 0. This can be too restrictive for some applications, since our model is the most parsimonious conditional Gaussian density that allows for varying conditional covariances, and we showed that when interaction terms are present in the model, our method does indeed perform much better. Fellinghauer et al. (2013) considered fitting ℓ_1-regularized regressions of each variable on the rest, while Lee and Hastie (2015) considered the maximum pseudo-likelihood approach. We chose to fit separate regressions, rather than maximize the joint pseudo-likelihood. One reason is that the number of parameters in our model is $\mathcal{O}(\max(q^2, p^2 q))$, making maximizing the joint pseudo-likelihood computationally more expensive than in the simpler setting of Lee and Hastie (2015) with $\mathcal{O}(q^2 + p^2)$ parameters; even in the simpler setting, maximizing the joint pseudo-likelihood is hundreds of times slower. Another reason is that we did not observe much difference between the separate regression approach and the joint pseudo-likelihood approach in the simpler setting of Fellinghauer et al. (2013) and Lee and Hastie (2015).

As we already know from the literature on the Gaussian graphical model and the Ising model, estimating conditional independence relationships between binary variables is in general more challenging. We observe it in this context as well, with interactions between two continuous variables being estimated better than interactions between two binary variables, or between a binary and a continuous variable. Establishing theoretical performance guarantees for the overlapping group penalty is outside the scope of the present article but presents an interesting challenge for the future, as the mixed variable setting is substantially more complicated than either the Gaussian or the pure binary setting. However, with the weighted lasso penalty approximation the separate regressions we fit reduce to standard settings for either the lasso linear or the logistic regression, where model selection results under appropriate conditions have already been established. We did not state these results in the article since joint conditions for the continuous and the binary variables are awkward and require a lot of notation and space to write out, the results themselves are standard, and this article is not focused on theory; nonetheless, this connection with the sparse regression literature guarantees reasonable behavior of our method provided the conditions are satisfied.

Acknowledgments

The authors thank the associate editor and two referees for their careful reading of the article and many helpful suggestions. This work was performed when the first author was a PhD student at the University of Michigan. E. Levina's research was partially supported by NSF grants DMS-1106772, DMS-1159005, DMS-1521551; J. Zhu's research was partially supported by NSF grant DMS-1407698 and NIH grant R01GM096194.

References

Cai, T., Liu, W., and Luo, X. (2011), "A Constrained ℓ_1 Minimization Approach to Sparse Precision Matrix Estimation," *Journal of the American Statistical Association*, 106, 594–607. [367]

Candes, E., and Tao, T. (2007), "The Dantzig Selector: Statistical Estimation When p is Much Larger than n," *Annals of Statistics*, 35, 2313–2351. [367]

Chow, C., and Liu, C. (1968), "Approximating Discrete Probability Distributions With Dependence Trees," *IEEE Transactions on Information Theory*, 14, 462–467. [368]

d'Aspremont, A., Banerjee, O., and El Ghaoui, L. (2008), "First-Order Methods for Sparse Covariance Selection," *SIAM Journal on Matrix Analysis and its Applications*, 30, 56–66. [367]

Edwards, D. (1990), "Hierarchical Interaction Models," *Journal of the Royal Statistical Society*, Series B, 52, 3–20. [368]

Edwards, D., De Abreu, G., and Labouriau, R. (2010), "Selecting High-Dimensional Mixed Graphical Models using Minimal AIC or BIC Forests," *BMC Bioinformatics*, 11, 18. [368]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [368]

Fellinghauer, B., Bühlmann, P., Ryffel, M., Von Rhein, M., and Reinhardt, J. D. (2013), "Stable Graphical Model Estimation With Random Forests for Discrete, Continuous, and Mixed Variables," *Computational Statistics & Data Analysis*, 64, 132–152. [368,371,373,374,377]

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse Inverse Covariance Estimation With the Graphical Lasso," *Biostatistics*, 9, 432–441. [367]

——— (2010), "Regularized Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33, 1–22. [371]

Höfling, H., and Tibshirani, R. (2009), "Estimation of Sparse Binary Pairwise Markov Networks using Pseudo-Likelihoods," *Journal of Machine Learning Research*, 10, 883–906. [367]

Ising, E. (1925), "Beitrag zur Theorie der Ferromagnetismus," *Zeitschrift fur Physik*, 31, 253–258. [368]

Jacob, L., Obozinski, G., and Vert, J.-P. (2009), "Group Lasso with Overlap and Graph Lasso," in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 433–440. [370]

Jenatton, R., Audibert, J.-Y., and Bach, F. (2011), "Structured Variable Selection with Sparsity-Inducing Norms," *The Journal of Machine Learning Research*, 12, 2777–2824. [370]

Kendall, M. (1938), "A New Measure of Rank Correlation," *Biomatrika*, 30, 81–89. [367]

Lam, C., and Fan, J. (2009), "Sparsistency and Rates of Convergence in Large Covariance Matrices Estimation," *Annals of Statistics*, 37, 4254–4278. [367]

Lauritzen, S. (1996), *Graphical Models*, Oxford: Oxford University Press. [368]

Lauritzen, S. L., and Wermuth, N. (1989), "Mixed Graphical Association Models," *Annals of Statistics*, 17, 31–57. [368,377]

Lee, J. D., and Hastie, T. J. (2015), "Learning the Structure of Mixed Graphical Models," *Journal of the Computational and Graphical Statistics*, 24, 230–253. [368,371,373,374,377]

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012), "High Dimensional Semiparametric Gaussian Copula Graphical Models," *Annals of Statistics*, 40, 2293–2326. [367]

Liu, H., Lafferty, J., and Wasserman, L. (2009), "The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs," *Journal of Machine Learning Research*, 10, 2295–2328. [367]

Logan, B. (2000), "Mel Frequency Cepstral Coefficients for Music Modeling," in *Proceedings of International Symposium on Music Information Retrieval (ISMIR)*. [375]

Meinshausen, N., and Bühlmann, P. (2006), "High Dimensional Graphs and Variable Selection With the Lasso," *Annals of Statistics*, 34, 1436–1462. [367,368,369,370]

——— (2010), "Stability Selection," *Journal of the Royal Statistical Society*, Series B, 72, 417–473. [375]

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009), "Partial Correlation Estimation by Joint Sparse Regression Model," *Journal of the American Statistics Association*, 104, 735–746. [367]

Ravikumar, P., Wainwright, M., and Lafferty, J. (2010), "High-Dimensional Ising Model Selection using ℓ_1-Regularized Logistic Regression," *Annals of Statistics*, 38, 1287–1319. [367,368,369,370]

Ravikumar, P., Wainwright, M., Raskutti, G., and Yu, B. (2009), "Model Selection in Gaussian Graphical Models: High-Dimensional Consistency of ℓ_1-Regularized MLE," *Advances in Neural Information Processing Systems*, 21, 1329–1336. [367]

Rocha, G. V., Zhao, P., and Yu, B. (2008), "A Path Following Algorithm for Sparse Pseudo-Likelihood Inverse Covariance Estimation," Technical Report 759, Department of Statistics, UC Berkeley. [367]

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008), "Sparse Permutation Invariant Covariance Estimation," *Electronic Journal of Statistics*, 2, 494–515. [367]

Spearman, C. (1904), "The Proof and Measurement of Association between Two Things," *American Journal of Psychology*, 15, 72–101. [367]

Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J., and Vlahavas, I. (2011), "Mulan: A Java Library for Multi-Label Learning," *Journal of Machine Learning Research*, 12, 2411–2414. [375]

Turnbull, D., Barrington, L., Torres, D., and Lanckriet, G. (2008), "Semantic Annotation and Retrieval of Music and Sound Effects," *IEEE Transactions on Audio, Speech and Language Processing*, 16, 467–476. [375]

Tzanetakis, G., and Cook, P. (2002), "Musical Genre Classification of Audio Signals," *IEEE Transactions on Audio, Speech and Language Processing*, 10, 293–302. [375]

Xue, L., and Zou, H. (2012), "Regularized Rank-Based Estimation of High-Dimensional Nonparanormal Graphical Models," *Annals of Statistics*, 40, 2541–2571. [367]

Xue, L., Zou, H., and Cai, T. (2012), "Nonconcave Penalized Composite Conditional Likelihood Estimation of Sparse Ising Models," *Annals of Statistics*, 40, 1403–1429. [368]

Yuan, L., Liu, J., and Ye, J. (2013), "Efficient Methods for Overlapping Group Lasso," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35, 2104–2116. [370]

Yuan, M. (2010), "Sparse Inverse Covariance Matrix Estimation via Linear Programming," *Journal of Machine Learning Research*, 11, 2261–2286. [367]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society*, Series B, 68, 49–67. [368,370,377]

——— (2007), "Model Selection and Estimation in the Gaussian Graphical Model," *Biometrika*, 94, 19–35. [367]

Zhao, P., Rocha, G., and Yu, B. (2009), "The Composite Absolute Penalties Family for Grouped and Hierarchical Variable Selection," *The Annals of Statistics*, 37, 3468–3497. [371]