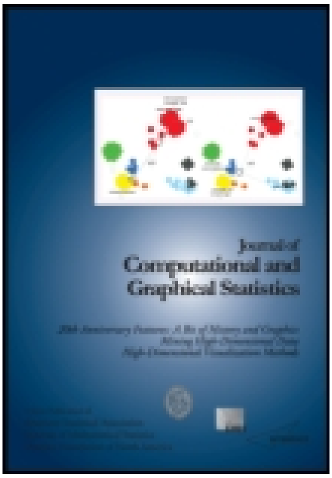

Journal of Computational and Graphical Statistics

Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/ucgs20

Graphical Models for Ordinal Data

Jian Guo, Elizaveta Levina, George Michailidis & Ji Zhu

Accepted author version posted online: 13 Mar 2014.Published online: 31 Mar 2015.

To cite this article: Jian Guo, Elizaveta Levina, George Michailidis & Ji Zhu (2015) Graphical Models for Ordinal Data, Journal of Computational and Graphical Statistics, 24:1, 183-204, DOI: 10.1080/10618600.2014.889023

To link to this article: http://dx.doi.org/10.1080/10618600.2014.889023

Graphical Models for Ordinal Data

Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu

This article considers a graphical model for ordinal variables, where it is assumed that the data are generated by discretizing the marginal distributions of a latent multivariate Gaussian distribution. The relationships between these ordinal variables are then described by the underlying Gaussian graphical model and can be inferred by estimating the corresponding concentration matrix. Direct estimation of the model is computationally expensive, but an approximate EM-like algorithm is developed to provide an accurate estimate of the parameters at a fraction of the computational cost. Numerical evidence based on simulation studies shows the strong performance of the algorithm, which is also illustrated on datasets on movie ratings and an educational survey.

Key Words: Lasso; Ordinal variable; Probit model.

1. INTRODUCTION

Graphical models have been successful in identifying directed and undirected structures from high-dimensional data. In a graphical model, the nodes of the network correspond to random variables and the edges represent their corresponding associations (Lauritzen 1996). Two canonical classes of graphical models are the Gaussian one, where the dependence structure is fully specified by the inverse covariance matrix and the Markov one, where the dependence structure is captured by the interaction effects in an exponential family model. In the latter model, each interaction effect can be interpreted as the conditional log-odds-ratio of the two associated variables given all other variables. In both models, a zero element in the inverse covariance matrix or a zero interaction effect determines a conditionally independent relationship between the corresponding nodes in the network.

Estimation of such models from high-dimensional data under a sparsity assumption has attracted a lot of interest in the statistics and machine learning literature, including regularized likelihood and regression methods, for example, see Yuan and Lin (2007); Banerjee, El Ghaoui, and d'Aspremont (2008); Friedman, Hastie, and Tibshirani (2008); Rothman et al. (2008); Fan, Feng, and Wu (2009); Meinshausen and Buhlmann (2006);

Jian Guo is Assistant Professor, Department of Biostatistics, Harvard University, Boston, MA 02138 (E-mail: *jguo@hsph.harvard.edu*). Elizaveta Levina is Professor (E-mail: *elevina@umich.edu*), George Michailidis is Professor (E-mail: *gmichail@umich.edu*), and Ji Zhu is Professor (E-mail: *jizhu@umich.edu*), Department of Statistics, University of Michigan, Ann Arbor, MI 48109.

Journal of Computational and Graphical Statistics, Volume 24, Number 1, Pages 183–204
DOI: 10.1080/10618600.2014.889023
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jcgs*.

Rocha, Zhao, and Yu (2008); Peng et al. (2009) and references therein. For a Markov network, direct estimation of a regularized likelihood is infeasible due to the intractable partition function in the likelihood. Instead, existing methods in the literature employ variants of approximation estimation methods. Examples include the surrogate likelihood methods (Banerjee, El Ghaoui, and d'Aspremont 2008; Kolar and Xing 2008) and the pseudo-likelihood methods (Höefling and Tibshirani 2009; Ravikumar, Wainwright, and Lafferty 2010; Guo et al. 2010).

In many applications involving categorical data, an ordering of the categories can be safely assumed. For example, in marketing studies consumers rate their preferences for a wide range of products. Similarly, computer recommender systems use customer ratings to make purchase recommendations to new customers; this constitutes a key aspect of the business model behind Netflix, Amazon, Tripadvisor, etc. (Koren, Bell, and Volinsky 2009).

Ordinal variables are also an integral part of survey data, where respondents rate items or express level of agreement/disagreement on issues/topics under consideration. Such responses correspond to Likert items, and a popular model to analyze such data is the polychotomous Rasch model (von Davier and Carstensen 2010) that obtains interval level estimates on a continuum—an idea that we explore in this work as well. Ordinal response variables in regression analysis give rise to variants of the classical linear model, including the proportional odds model (Walker and Duncan 1967; McCullagh 1980), the partial proportional odds model (Peterson 1990), the probit model (Bliss 1935; Albert and Chib 1993; Chib and Greenberg 1998), etc. A comprehensive review of ordinal regression was given by McCullagh and Nelder (1989) and O'Connell (2005).

In this article, we introduce a graphical model for ordinal variables. It is based on the assumption that the ordinal scales are generated by discretizing the marginal distributions of a latent multivariate Gaussian distribution and the dependence relationships of these ordinal variables are induced by the underlying Gaussian graphical model. In this context, an EM-like algorithm is appropriate for estimating the underlying latent network, which presents a number of technical challenges that have to be addressed for successfully pursuing this strategy.

Our work is related to Albert and Chib (1993), Chib and Greenberg (1998), and Stern, Herbrich, and Graepel (2009) in the sense that they are all built on the probit model and/or the EM algorithmic framework. Albert and Chib (1993) proposed an MCMC algorithm for the probit-model-based univariate ordinal regression problem, where an ordinal response is fitted on a number of covariates, while Chib and Greenberg (1998) can be considered an extension to the multivariate case. Stern, Herbrich, and Graepel (2009) aimed to build an online recommender system via collaborative filtering and applied the discretization/thresholding idea in the probit model to the ordinal matrix factorization problem. Our model, on the other hand, has a completely different motivation from these works. Our objective is to explore associations between a set of ordinal variables, rather than prediction and/or regression problems. Nevertheless, the EM framework employed is related to that in Chib and Greenberg (1998), but due to the different goal, the form of the likelihood function of the proposed model is different from that of the ordinal regression problem. Further, as seen in Section 2, we do not use any MCMC or Gibbs sampling scheme.

The remainder of the article is organized as follows. Section 2 presents the probit graphical model and discusses algorithmic and model selection issues. Section 3 evaluates the performance of the proposed method on several synthetic examples and Section 4 applies the model to two data examples, one on movie ratings and the other on a national educational longitudinal survey study.

2. METHODOLOGY

2.1 The Probit Graphical Model

Suppose we have p ordinal random variables $X_1, \dots, X_p$, where $X_j \in \{1, 2, \dots, K_j\}$ for some integer K_j, which is the number of the ordinal levels in variable j. In the proposed probit graphical model, we assume that there exist p latent random variables $Z_1, \dots, Z_p$ from a joint Gaussian distribution with mean zero and covariance matrix $\boldsymbol{\Sigma} = (\sigma_{j,j'})_{p\times p}$. Without loss of generality, we further assume that Z_j's have unit variances ($\sigma_{j,j} = 1$ for $j = 1, \dots, p$), that is, the Z_j's marginally follow standard Gaussian distributions. Each observed variable X_j is discretized from its latent counterpart Z_j. Specifically, for the jth variable ($j = 1, \dots, p$), we assume that $(-\infty, +\infty)$ is split into K_j disjointed intervals by a set of thresholds $-\infty = \theta_0^{(j)} < \theta_1^{(j)} < \cdots < \theta_{K_j-1}^{(j)} < \theta_{K_j}^{(j)} = +\infty$, such that $X_j = k$ if and only if Z_j falls in the interval $[\theta_{k-1}^{(j)}, \theta_k^{(j)})$. Thus,

$$\Pr(X_j = k) = \Pr\left(\theta_{k-1}^{(j)} \le Z_j < \theta_k^{(j)}\right) = \Phi\left(\theta_k^{(j)}\right) - \Phi\left(\theta_{k-1}^{(j)}\right), \tag{1}$$

where $\Phi(\cdot)$ denotes the cumulative density function of the standard normal distribution.

Let $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1} = (\omega_{j,j'})_{p\times p}$, $\boldsymbol{\Theta} = \{\theta_k^{(j)} : j = 1, \dots, p; k = 1, \dots, K_j\}$, $\boldsymbol{X} = (X_1, \dots, X_p)$, $\boldsymbol{Z} = (Z_1, \dots, Z_p)$. Let $C(\boldsymbol{X}, \boldsymbol{\Theta})$ be the hypercube defined by $[\theta_{X_1-1}^{(1)}, \theta_{X_1}^{(1)}) \times \cdots \times [\theta_{X_p-1}^{(p)}, \theta_{X_p}^{(p)})$. Then we can write the joint density function of $(\boldsymbol{X}, \boldsymbol{Z})$ as

$$\mathrm{f}_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{x}, z; \boldsymbol{\Omega}, \boldsymbol{\Theta}) = \mathrm{f}(z; \boldsymbol{\Omega}) \prod_{j=1}^{p} \mathrm{f}_{\boldsymbol{\Theta}}(x_j | z_j; \boldsymbol{\Theta}) = \frac{\det(\boldsymbol{\Omega})}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2} z \boldsymbol{\Omega} z^{\mathrm{T}}\right) \mathrm{I}(z \in C(\boldsymbol{x}, \boldsymbol{\Theta})), \tag{2}$$

where $\mathrm{I}(\cdot)$ is the indicator function. Thus, the marginal probability density function of the observed $\boldsymbol{X}$ is given by

$$\mathrm{f}_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\Omega}, \boldsymbol{\Theta}) = \int_{z\in\mathbb{R}^p} \mathrm{f}_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{x}, z; \boldsymbol{\Omega}, \boldsymbol{\Theta}) dz. \tag{3}$$

We refer to (1)–(3) as the *probit graphical model*, which is motivated by the probit regression model (Bliss 1935; Albert and Chib 1993; Chib and Greenberg 1998) and the polychotomous Rasch model (von Davier and Carstensen 2010).

To fit the probit graphical model, we propose maximizing an ℓ_1-regularized log-likelihood of the observed data. Let $x_{i,j}$ and $z_{i,j}$ be the ith realizations of the observed variable X_j and the latent variable Z_j, respectively, with $\boldsymbol{x}_i = (x_{i,1}, \dots, x_{i,p})$ and $\boldsymbol{z}_i = (z_{i,1}, \dots, z_{i,p})$. The criterion is given by

$$\sum_{i=1}^{n} \log \mathrm{f}_{\boldsymbol{X}}(\boldsymbol{x}_i; \boldsymbol{\Omega}, \boldsymbol{\Theta}) - \lambda \sum_{j\neq j'} |\omega_{j,j'}|. \tag{4}$$

The tuning parameter λ in (4) controls the degree of sparsity in the underlying network. When λ is large enough, some $\omega_{j,j'}$'s can be shrunken to zero, resulting in the removal of the corresponding links in the underlying network. Numerically, it is difficult to maximize criterion (4) directly, because of the integral in (3). Next, we introduce an EM-type algorithm to maximize (4) in an iterative manner.

2.2 AN ALGORITHM FOR FITTING THE PROBIT GRAPHICAL MODEL

Criterion (4) depends on the parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\Omega}$ and the latent variable $\boldsymbol{Z}$. The former has a closed-form estimator. Specifically, for each $j = 1, \ldots, p$, we set

$$\widehat{\theta}_k^{(j)} = \begin{cases} -\infty, & \text{if } k = 0; \\ \Phi^{-1}\left(n^{-1}\sum_{i=1}^{n} \mathrm{I}(x_{i,j} < k)\right), & \text{if } k = 1, \ldots, K_j - 1; \\ +\infty, & \text{if } k = K_j; \end{cases} \tag{5}$$

where Φ is the cumulative distribution function of the standard normal. One can show that $\widehat{\boldsymbol{\Theta}}$ consistently estimates $\boldsymbol{\Theta}$. The estimation of $\boldsymbol{\Omega}$, on the other hand, is nontrivial due to the multiple integrals in (3). To address this problem, we apply the EM algorithm to optimizing (4), where the latent variables $z_{i,j}$'s ($i = 1, \ldots, n; j = 1, \ldots, p$) are treated as "missing data" and are imputed in the E-step, and the parameter $\boldsymbol{\Omega}$ is estimated in the M-step.

E-step. Suppose $\widehat{\boldsymbol{\Omega}}$ is the updated estimate of $\boldsymbol{\Omega}$ from the M-step. Then the E-step computes the conditional expectation of the joint log-likelihood given the estimates $\widehat{\boldsymbol{\Theta}}$ and $\widehat{\boldsymbol{\Omega}}$, which is usually called the Q-function in the literature:

$$\mathrm{Q}(\boldsymbol{\Theta}, \boldsymbol{\Omega}) = \sum_{i=1}^{n} \mathrm{E}_{Z}[\log \mathrm{f}_{X,Z}(\boldsymbol{x}_i, \boldsymbol{Z}; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})] = \frac{n}{2}[\log\det(\boldsymbol{\Omega}) - \mathrm{trace}(\boldsymbol{S}\boldsymbol{\Omega}) - p\log(2\pi)]. \tag{6}$$

Here $\boldsymbol{S}$ is a $p \times p$ matrix whose (j, j')th element is $s_{j,j'} = n^{-1}\sum_{i=1}^{n} \mathrm{E}(z_{i,j}z_{i,j'} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$ $(1 \le j, j' \le p)$. The distribution of $\boldsymbol{z}_i$ conditional on $\boldsymbol{x}_i$ is equal to that of $\boldsymbol{z}_i$ conditional on $\boldsymbol{z}_i \in C(\boldsymbol{x}_i, \boldsymbol{\Theta})$, which follows a truncated multivariate Gaussian distribution on the hypercube $C(\boldsymbol{x}_i, \boldsymbol{\Theta})$. Therefore, $\mathrm{E}(z_{i,j}z_{i,j'} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$ is the second moment of a truncated multivariate Gaussian distribution and it can be directly estimated using the algorithms proposed by Tallis (1961), Lee (1979), Leppard and Tallis (1989), and Manjunath and Wilhelm (2012). Nevertheless, the computational cost of these direct estimation algorithms is extremely high and thus not suitable for even moderate size problems. An alternative approach is based on the Markov chain Monte Carlo (MCMC) method. Specifically, we randomly generate a sequence of samples from the conditional distribution $f_{Z|X}(\boldsymbol{z}_i \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$ using a Gibbs sampler from a multivariate truncated normal distribution (Kotecha and Djuric 1999) and then $\mathrm{E}(z_{i,j}z_{i,j'} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$ is estimated by the empirical conditional second moment from these samples. Although the MCMC approach is faster than the direct estimation method, it is still not efficient for large-scale problems. To address this computational issue, we develop an efficient approximate estimation algorithm, discussed in Section 2.3.

M-step. The M-step updates $\mathbf{\Omega}$ by maximizing the ℓ_1-regularized Q-function (up to a constant and a factor):

$$\tilde{\Omega} = \arg\max_{\mathbf{\Omega}} \log\det(\mathbf{\Omega}) - \mathrm{trace}(\boldsymbol{S}\mathbf{\Omega}) - \lambda \sum_{j \neq j'} |\omega_{j,j'}|. \tag{7}$$

The optimization problem (7) can be solved efficiently by existing algorithms such as the graphical lasso (Friedman, Hastie, and Tibshirani 2008) and SPICE (Rothman et al. 2008). However, the estimated covariance matrix, $\widetilde{\mathbf{\Sigma}} = \widetilde{\mathbf{\Omega}}^{-1}$, does not necessarily have unit diagonal elements postulated by the probit graphical model. Therefore, we postprocess $\widetilde{\mathbf{\Sigma}}$ by scaling it to a unit-diagonal matrix $\widehat{\mathbf{\Sigma}}$ and update $\widehat{\mathbf{\Omega}} = \widehat{\mathbf{\Sigma}}^{-1}$, which will be used in the E-step of the next iteration.

2.3 APPROXIMATING THE CONDITIONAL EXPECTATION

Note that when $j = j'$, the corresponding conditional expectation is the conditional second moment $\mathrm{E}(z_{i,j}^2 \mid \boldsymbol{x}_i; \widehat{\mathbf{\Theta}}, \widehat{\mathbf{\Omega}})$; when $j \neq j'$, we use a mean field theory approach (Peterson and Anderson 1987) to approximate it as $E(z_{i,j}z_{i,j'} \mid \boldsymbol{x}_i; \widehat{\mathbf{\Theta}}, \widehat{\mathbf{\Omega}}) \approx E(z_{i,j} \mid \boldsymbol{x}_i; \widehat{\mathbf{\Theta}}, \widehat{\mathbf{\Omega}}) E(z_{i,j'} \mid \boldsymbol{x}_i; \widehat{\mathbf{\Theta}}, \widehat{\mathbf{\Omega}})$. Note that the approximation decouples the "interaction" between the two variables $z_{i,j}$ and $z_{i,j'}$. Therefore, one would expect that the approximation performs well when z_j and $z_{j'}$ are close to independence given all other random variables, which often holds when $\mathbf{\Omega}$ or the corresponding graph is sparse. With this approximation, it is sufficient to estimate the first moment $E(z_{i,j} \mid \boldsymbol{x}_i; \widehat{\mathbf{\Theta}}, \widehat{\mathbf{\Omega}})$ and the second moment $E(z_{i,j}^2 \mid \boldsymbol{x}_i; \widehat{\mathbf{\Theta}}, \widehat{\mathbf{\Omega}})$. In general, the latent variable $z_{i,j}$ not only depends on $x_{i,j}$, but also on all other observed variables $\boldsymbol{x}_{i,-j} = (x_{i,1}, \dots, x_{i,j-1}, x_{i,j+1}, \dots, x_{i,p})$. We can write the first and second conditional moments as

$$E(z_{i,j} \mid \boldsymbol{x}_i; \widehat{\mathbf{\Theta}}, \widehat{\mathbf{\Omega}}) = E[E(z_{i,j} \mid \boldsymbol{z}_{i,-j}, x_{i,j}; \widehat{\mathbf{\Theta}}, \widehat{\mathbf{\Omega}}) \mid \boldsymbol{x}_i; \widehat{\mathbf{\Theta}}, \widehat{\mathbf{\Omega}}], \tag{8}$$

$$E\left(z_{i,j}^2 \mid \boldsymbol{x}_i; \widehat{\mathbf{\Theta}}, \widehat{\mathbf{\Omega}}\right) = E\left[E\left(z_{i,j}^2 \mid \boldsymbol{z}_{i,-j}, x_{i,j}; \widehat{\mathbf{\Theta}}, \widehat{\mathbf{\Omega}}\right) \mid \boldsymbol{x}_i; \widehat{\mathbf{\Theta}}, \widehat{\mathbf{\Omega}}\right], \tag{9}$$

where $\boldsymbol{z}_{i,-j} = (z_{i,1}, \dots, z_{i,j-1}, z_{i,j+1}, \dots, z_{i,p})$. The inner expectations in (8) and (9) are relatively straightforward to compute: given the parameter estimate $\widehat{\mathbf{\Omega}}$, $z_{i,1}, \dots, z_{i,p}$ jointly follow a multivariate Gaussian distribution with mean zero and covariance matrix $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{\Omega}}^{-1}$. A property of the Gaussian distribution is that the conditional distribution of $z_{i,j}$ given $\boldsymbol{z}_{i,-j}$ is also Gaussian, with mean $\widetilde{\mu}_{i,j} = \widehat{\mathbf{\Sigma}}_{j,-j}\widehat{\mathbf{\Sigma}}_{-j,-j}^{-1}\boldsymbol{z}_{i,-j}{}^{\mathsf{T}}$ and variance $\widetilde{\sigma}_{i,j}^2 = 1 - \widehat{\mathbf{\Sigma}}_{j,-j}\widehat{\mathbf{\Sigma}}_{-j,-j}^{-1}\widehat{\mathbf{\Sigma}}_{-j,j}$. Moreover, given the observed data $x_{i,j}$, conditioning $z_{i,j}$ on $\boldsymbol{z}_{i,-j}, x_{i,j}$ in (8) is equivalent to conditioning on $\boldsymbol{z}_{i,-j}, \theta_{x_{i,j}-1}^{(j)} \leq z_{i,j} < \theta_{x_{i,j}}^{(j)}$, which follows a truncated Gaussian distribution on the interval $[\theta_{x_{i,j}-1}^{(j)}, \theta_{x_{i,j}}^{(j)})$. The following lemma gives the closed-form expressions for the first and second moments of the truncated Gaussian distribution.

Lemma 1. Suppose that a random variable Y follows the Gaussian distribution with mean μ_0 and variance σ_0^2. For any constants $t_1 < t_2$, let $\xi_1 = (t_1 - \mu_0)/\sigma_0$ and $\xi_2 = (t_2 - \mu_0)/\sigma_0$. Then the first and second moments of Y truncated to the interval (t_1, t_2) are given by

$$\mathrm{E}(Y \mid t_1 < Y < t_2) = \mu_0 + \tfrac{\phi(\xi_1)-\phi(\xi_2)}{\Phi(\xi_2)-\Phi(\xi_1)}\sigma_0, \tag{10}$$

$$\mathrm{E}(Y^2 \mid t_1 < Y < t_2) = \mu_0^2 + \sigma_0^2 + 2\tfrac{\phi(\xi_1)-\phi(\xi_2)}{\Phi(\xi_2)-\Phi(\xi_1)}\mu_0\sigma_0 + \tfrac{\xi_1\phi(\xi_1)-\xi_2\phi(\xi_2)}{\Phi(\xi_2)-\Phi(\xi_1)}\sigma_0^2, \tag{11}$$

where $\phi(\cdot)$ is the probability density function of the standard normal.

For more properties of the truncated Gaussian distribution, see Johnson, Kotz, and Balakrishnan (1994).

Letting $\delta_{i,j,k} = (\theta_k^{(j)} - \widetilde{\mu}_{i,j})/\widetilde{\sigma}_{i,j}$ and applying Lemma 1 to the conditional expectations in (8) and (9), we obtain

$$\mathrm{E}(z_{i,j}|\boldsymbol{z}_{i,-j}, x_{i,j}; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}) = \widetilde{\mu}_{i,j} + a_{i,j}\widetilde{\sigma}_{i,j}, \tag{12}$$

$$\mathrm{E}\big(z_{i,j}^2|\boldsymbol{z}_{i,-j}, x_{i,j}; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}\big) = \widetilde{\mu}_{i,j}^2 + \widetilde{\sigma}_{i,j}^2 + 2a_{i,j}\widetilde{\mu}_{i,j}\widetilde{\sigma}_{i,j} + b_{i,j}\widetilde{\sigma}_{i,j}^2\ , \tag{13}$$

where

$$a_{i,j} = \frac{\phi(\delta_{i,j,x_{i,j}-1}) - \phi(\delta_{i,j,x_{i,j}})}{\Phi(\delta_{i,j,x_{i,j}}) - \Phi(\delta_{i,j,x_{i,j}-1})}\ , \quad b_{i,j} = \frac{\delta_{i,j,x_{i,j}-1}\phi(\delta_{i,j,x_{i,j}-1}) - \delta_{i,j,x_{i,j}}\phi(\delta_{i,j,x_{i,j}})}{\Phi(\delta_{i,j,x_{i,j}}) - \Phi(\delta_{i,j,x_{i,j}-1})}\ .$$

Next, we plug Equations (12) and (13) into (8) and (9), respectively. Since $\widetilde{\mu}_{i,j}$, $a_{i,j}$, and $b_{i,j}$ depend on the latent variables $z_{i,-j}$'s, the outer expectations in (8) and (9) depend on $\mathrm{E}(\widetilde{\mu}_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$, $\mathrm{E}(a_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$, $\mathrm{E}(b_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$, and $\mathrm{E}(a_{i,j}\widetilde{\mu}_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$. Note that $\widetilde{\mu}_{i,j}$ is a linear function of $z_{i,-j}$ and $\widetilde{\sigma}_{i,j}$ is a constant irrelevant to the latent data. For each $i = 1, \ldots, n$ and $j = 1, \ldots, p$, the conditional expectation of $\widetilde{\mu}_{i,j}$ is

$$\mathrm{E}(\widetilde{\mu}_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Omega}}) = \widehat{\boldsymbol{\Sigma}}_{j,-j}\widehat{\boldsymbol{\Sigma}}_{-j,-j}^{-1}\mathrm{E}\big(\boldsymbol{z}_{i,-j}^{\mathsf{T}}|\boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Omega}}\big). \tag{14}$$

However, $a_{i,j}$ and $b_{i,j}$ are nonlinear functions of $\widetilde{\mu}_{i,j}$, and thus of $\boldsymbol{z}_{i,-j}$. Using the first-order delta method, we approximate their conditional expectations by

$$\mathrm{E}(a_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Omega}}) \approx \frac{\phi(\widetilde{\delta}_{i,j,x_{i,j}-1}) - \phi(\widetilde{\delta}_{i,j,x_{i,j}})}{\Phi(\widetilde{\delta}_{i,j,x_{i,j}}) - \Phi(\widetilde{\delta}_{i,j,x_{i,j}-1})}, \tag{15}$$

$$\mathrm{E}(b_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Omega}}) \approx \frac{\widetilde{\delta}_{i,j,x_{i,j}-1}\phi(\widetilde{\delta}_{i,j,x_{i,j}-1}) - \widetilde{\delta}_{i,j,x_{i,j}}\phi(\widetilde{\delta}_{i,j,x_{i,j}})}{\Phi(\widetilde{\delta}_{i,j,x_{i,j}}) - \Phi(\widetilde{\delta}_{i,j,x_{i,j}-1})}, \tag{16}$$

where $\widetilde{\delta}_{i,j,x_{i,j}} = [\theta_k^{(j)} - \mathrm{E}(\widetilde{\mu}_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Omega}})]/\widetilde{\sigma}_{i,j}$. Finally, we approximate $\mathrm{E}(a_{i,j}\widetilde{\mu}_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}) \approx \mathrm{E}(a_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})\mathrm{E}(\widetilde{\mu}_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$. Therefore, (8) and (9) can be approximated by

$$\mathrm{E}(z_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}) \approx \widehat{\boldsymbol{\Sigma}}_{j,-j}\widehat{\boldsymbol{\Sigma}}_{-j,-j}^{-1}\mathrm{E}\big(\boldsymbol{z}_{i,-j}^{\mathsf{T}} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}\big) + \frac{\phi(\widetilde{\delta}_{i,j,x_{i,j}-1}) - \phi(\widetilde{\delta}_{i,j,x_{i,j}})}{\Phi(\widetilde{\delta}_{i,j,x_{i,j}}) - \Phi(\widetilde{\delta}_{i,j,x_{i,j}-1})}\widetilde{\sigma}_{i,j} \tag{17}$$

$$\begin{aligned}\mathrm{E}\big(z_{i,j}^2 \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}\big) \approx{} & \widehat{\boldsymbol{\Sigma}}_{j,-j}\widehat{\boldsymbol{\Sigma}}_{-j,-j}^{-1}\mathrm{E}\big(\boldsymbol{z}_{i,-j}^{\mathsf{T}}\boldsymbol{z}_{i,-j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}\big)\widehat{\boldsymbol{\Sigma}}_{-j,-j}^{-1}\widehat{\boldsymbol{\Sigma}}_{j,-j}^{\mathsf{T}} + \widetilde{\sigma}_{i,j}^2 \\ & + 2\frac{\phi(\widetilde{\delta}_{i,j,x_{i,j}-1}) - \phi(\widetilde{\delta}_{i,j,x_{i,j}})}{\Phi(\widetilde{\delta}_{i,j,x_{i,j}}) - \Phi(\widetilde{\delta}_{i,j,x_{i,j}-1})}\big[\widehat{\boldsymbol{\Sigma}}_{j,-j}\widehat{\boldsymbol{\Sigma}}_{-j,-j}^{-1}\mathrm{E}\big(\boldsymbol{z}_{i,-j}^{\mathsf{T}} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}\big)\big]\widetilde{\sigma}_{i,j} \\ & + \frac{\delta_{i,j,x_{i,j}-1}^{(j)}\phi(\widetilde{\delta}_{i,j,x_{i,j}-1}) - \widetilde{\delta}_{i,j,x_{i,j}}\phi(\widetilde{\delta}_{i,j,x_{i,j}})}{\Phi(\widetilde{\delta}_{i,j,x_{i,j}}) - \Phi(\widetilde{\delta}_{i,j,x_{i,j}-1})}\widetilde{\sigma}_{i,j}^2.\end{aligned} \tag{18}$$

Equations (17) and (18) establish the recursive relationships among the elements in $\mathrm{E}(z_i \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$ and $\mathrm{E}({z_i}^{\mathsf{T}} z_i \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$, respectively, giving a natural iterative procedure for estimating these quantities. Algorithm 1 summarizes the main steps of the proposed combined estimation procedure outlined in Sections 2.2 and 2.3.

In Algorithm 1, Lines 1–2 initialize the conditional expectation $E(z_{i,j} \mid \boldsymbol{x}_i)$ and the parameter estimate $\widehat{\boldsymbol{\Omega}}$. Lines 3–16 establish the outer loop which iteratively computes the E-step and the M-step. In the E-step, Lines 5–14 consist of the inner loop which recursively estimates the first and second moments of $z_{i,j}$ conditional on $\boldsymbol{x}_i$. The complexity of the inner loop is $O(np^2)$, which is the same as that of the graphical lasso algorithm in the M-step. Therefore, the overall complexity of Algorithm 1 is $O(Mnp^2)$, where M is the number of EM steps required for convergence. In our numerical studies, we found M is often smaller than 50. For a more concrete idea about the computational cost, we note that on a linux server with four 1G Dual-Core AMD Opteron Processors and 4GB RAM, it takes about 2 min for the proposed algorithm to complete the fitting on a simulated dataset in Section 3 with $n = 200$ observations and $p = 50$ variables.

Algorithm 1 The EM Algorithm for estimating $\boldsymbol{\Omega}$

```
1:  Initialize $\mathrm{E}(z_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}) \approx \mathrm{E}(z_{i,j} \mid x_{i,j}; \widehat{\boldsymbol{\Theta}})$, $\mathrm{E}(z_{i,j}^2 \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}) \approx \mathrm{E}(z_{i,j}^2 \mid x_{i,j}; \widehat{\boldsymbol{\Theta}})$
    and $\mathrm{E}(z_{i,j} z_{i,j'} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}) \approx \mathrm{E}(z_{i,j} \mid x_{i,j}; \widehat{\boldsymbol{\Theta}})\mathrm{E}(z_{i,j'} \mid x_{i,j'}; \widehat{\boldsymbol{\Theta}})$ for $i = 1, \dots, n$ and
    $j, j' = 1, \dots, p$;
2:  Initialize $s_{j,j'}$ for $1 \le j, j' \le p$ using the Line 1 above, and then estimate $\widehat{\boldsymbol{\Omega}}$ by maximizing criterion (7);
    {Start outer loop}
3:  repeat
4:    E-step: estimate $\boldsymbol{S}$ in (6);
      {Start inner loop}
5:    repeat
6:      for $i = 1$ to n do
7:        if $j = j'$ then
8:          Update $\mathrm{E}(z_{i,j}^2 \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$ using RHS of Equation (18) for $j = 1, \dots, p$;
9:        else
10:         Update $\mathrm{E}(z_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$ using RHS of Equation (17) for $j = 1, \dots, p$ and
            then set $\mathrm{E}(z_{i,j} z_{i,j'} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}) = \mathrm{E}(z_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})\mathrm{E}(z_{i,j'} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$ for $1 \le$
            $j \ne j' \le p$;
11:       end if
12:     end for
13:     Update $s_{j,j'} = 1/n \sum_{i=1}^{n} \mathrm{E}(z_{i,j} z_{i,j'} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$ for $1 \le j, j' \le p$;
14:   until The inner loop converges;
15:   M-step: update $\widehat{\boldsymbol{\Omega}}$ by maximizing criterion (7);
16: until The outer loop converges.
```

2.4 Model Selection

In the probit graphical model, the tuning parameter λ controls the sparsity of the resulting estimator and it can be selected using cross-validation. Specifically, we randomly split the observed data $\boldsymbol{X}$ into D subsets of similar sizes and denote the index set of the observations in the dth subset by $\mathcal{T}_d$ $(d = 1, \ldots, D)$. For any prespecified λ, we denote by $\widehat{\boldsymbol{\Omega}}_\lambda^{[-d]}$ the maximizer of the criterion (4) estimated by Algorithm 1 using all observations except those in $\mathcal{T}_d$. We also denote by $\widehat{\boldsymbol{\Theta}}^{[-d]}$ and $\boldsymbol{S}^{[d]} = (s_{j,j'}^{[d]})_{p\times p}$ the analogs of $\widehat{\boldsymbol{\Theta}}$ and $\boldsymbol{S}$ in Section 2.2, but computed from the data in $\mathcal{T}_d^c$ and $\mathcal{T}_d$, respectively. In particular, an element of $\boldsymbol{S}^{[d]}$ is defined as $s_{j,j'}^{[d]} = |\mathcal{T}_d|^{-1} \sum_{i\in\mathcal{T}_d} \mathrm{E}(z_{i,j} z_{i,j'} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}^{[-d]}, \widehat{\boldsymbol{\Omega}}_\lambda^{[-d]})$, for $1 \le j, j' \le p$, where $|\mathcal{T}_d|$ is the cardinality of $\mathcal{T}_d$. Given $\widehat{\boldsymbol{\Theta}}^{[-d]}$ and $\widehat{\boldsymbol{\Omega}}_\lambda^{[-d]}$, $\boldsymbol{S}^{[d]}$ can be estimated by the algorithm introduced in Section 2.3, that is, the inner loop of Algorithm 1. Thus, the optimal tuning parameter can be selected by maximizing the following criterion:

$$\max_{\lambda} \sum_{d=1}^{D} \log\det\left(\widehat{\boldsymbol{\Omega}}_\lambda^{[-d]}\right) - \mathrm{trace}\left(\boldsymbol{S}^{[d]}\widehat{\boldsymbol{\Omega}}_\lambda^{[-d]}\right) - p\log(2\pi). \tag{19}$$

We note that we have also considered the AIC and BIC type criteria for choosing the tuning parameter λ. We found that AIC performs the worst among the three due to estimating many zero parameters as nonzeros (Lian 2011); BIC and cross-validation tend to have similar performances in estimating zero parameters as zeros, but BIC also tends to estimate the nonzero parameters as zeros. Therefore, we choose to use cross-validation. Due to space limitation, the results are not included.

3. NUMERICAL EXAMPLES

In this section, we use two sets of simulated experiments to illustrate the performance of the probit graphical model. The first set aims at comparing the computational cost of the three methods that estimate the Q-function in the E-step, namely the direct computation, the MCMC sampling and the approximation algorithm. The second set compares the performance of the probit graphical model using the approximation algorithm to that of the Gaussian graphical model.

3.1 Computational Cost and Performance

Note that the computational costs of the direct estimation and the MCMC sampling are extremely high when p is even of moderate size. Therefore, in this experiment, we simulate a low-dimensional dataset with $p = 5$ variables and $n = 10$ observations. Specifically, we define the underlying inverse covariance matrix $\boldsymbol{\Omega}$ as a tri-diagonal matrix with 1's on the main diagonal and 0.5 on the first subdiagonal. The corresponding covariance matrix is then scaled so that all the variances are equal to 1. Then, for $i = 1, \ldots, n$, we generate the latent data $\boldsymbol{z}_i = (z_{i,1}, \ldots, z_{i,p})$ from $N(\boldsymbol{0}, \boldsymbol{\Sigma})$ and discretize them as follows: for each

Table 1. The numbers are the mean CPU times for different tuning parameter values and 20 replications, with median absolute deviations in parentheses (in sec)

Method	CPU time in sec
Direct	3310.21 (199.95)
Gibbs sampler	46.17 (1.51)
TMG	303.94 (11.05)
Proposed approximation	0.04 (0.03)

NOTE: Direct: direct computation. Gibbs sampler: the regular Gibbs sampler; TMG: the Gibbs sampler proposed by Pakman and Paninski (2012) via the R package "tmg"; Proposed approximation: the approximation approach proposed in our article.

$j = 1, \ldots, p$, set

$$\theta_k^{(j)} = \begin{cases} -\infty, & \text{if } k = 0; \\ \Phi^{-1}(0.2) & \text{if } k = 1; \\ \Phi^{-1}(0.4) & \text{if } k = 2; \\ +\infty, & \text{if } k = 3; \end{cases} \tag{20}$$

and $x_{i,j} = \sum_{k=0}^{2} \mathrm{I}(z_{i,j} \geq \theta_k^{(j)})$ $(i = 1, \ldots, n;\ j = 1, \ldots, p)$, that is, the value of $x_{i,j}$ is k if it locates in interval $[\theta_{k-1}^{(j)}, \theta_k^{(j)})$.

The probit graphical model is applied using four estimation methods in the E-step, namely the direct computation, a standard Gibbs sampling, the Gibbs sampler proposed by Pakman and Paninski (2012) and the approximation algorithm proposed in this article. The procedure is repeated for 20 times, and the computational costs are shown in Table 1. We can see that the median CPU time of the approximation algorithm is only about 1/1000 of that of the Gibbs sampling and about 1/80,000 of that of the direct computation. To further compare the estimation accuracy of these methods, we use the Frobenius and entropy loss metrics that are defined next:

$$\mathrm{FL} = \frac{\sum_{1 \leq j < j' \leq p} (\omega_{j,j'} - \widehat{\omega}_{j,j'})^2}{\sum_{1 \leq j < j' \leq p} \omega_{j,j'}^2}, \tag{21}$$

$$\mathrm{EL} = \mathrm{trace}(\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\Omega}}) - \log[\det(\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\Omega}})] - p, \tag{22}$$

where $\widehat{\boldsymbol{\Omega}}$ denotes the estimated network.

The performance of the three estimation methods is depicted in Figure 1. It can be seen that the direct computation and Gibbs sampling methods are fairly similar in performance (the result using the R package "tmg" is almost identical to that of the standard Gibbs sampling and not shown); this is expected since they can all be considered "exact" approaches. In terms of the Frobenius and entropy losses, the approximation algorithm lags slightly behind its competitors when the tuning parameter λ is relatively small, whereas for larger λ it outperforms them. This is because in this simulation study, the true $\boldsymbol{\Omega}$ is very sparse and the mean field approximation also happens to implicitly enforce a conditional independence structure on the $\boldsymbol{S}$ matrix. These findings suggest that the proposed approximation algorithm achieves its orders of magnitude computational savings over the competitors with minimal degradation in performance.

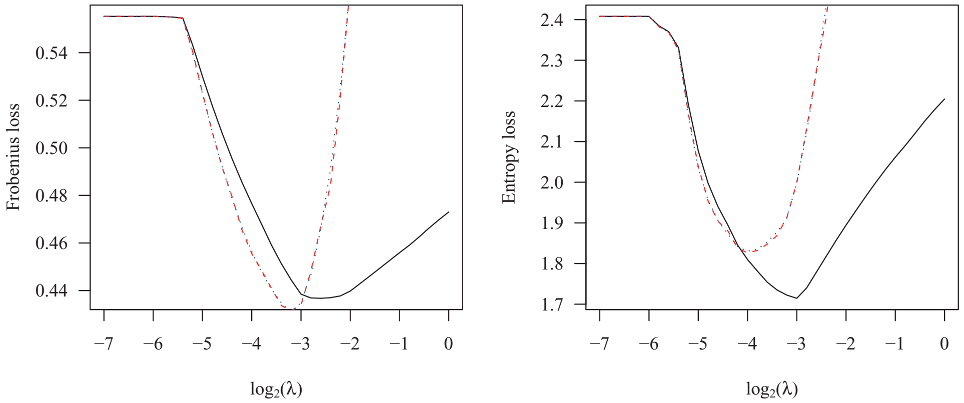

Figure 1. Comparison of Frobenius loss and entropy loss over different values of the tuning parameter. The direct computation, the MCMC sampling and the approximation algorithm are, respectively, represented by blue dotted, red dashed, and black solid lines.

3.2 Experiments With Different Types of Graphs

In this section, we evaluate the performance of the proposed method by simulation studies. These examples simulate four types of network structures: a scale-free graph, a hub graph, a nearest-neighbor graph and a block graph. Each network consists of $p = 50$ nodes. The details of these networks are described as follows:

Example 1: Scale-free graph. A scale-free graph has a power-law degree distribution and can be simulated by the Barabasi-Albert algorithm (Barabasi and Albert 1999). A realization of a scale-free network is depicted in Figure 2(a).

Example 2: Hub graph. A hub graph consists of a few high-degree nodes (hubs) and a large amount of low-degree nodes. In this example, we follow the simulation setting in Peng et al. (2009) and generate a hub graph by inserting a few hub nodes into a very sparse graph. Specifically, the graph consists of three hubs with degrees around eight, and the other 47 nodes with degrees at most three. An example of the hub graph is shown in Figure 2(b).

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

Example 4: Block graph. In this setting, we generate a graph using a random adjacency matrix generated from the stochastic block model. Specifically, for nodes 1–20 the probability of being linked is 0.2, for nodes 21–30 the probability of being linked is 0.5,

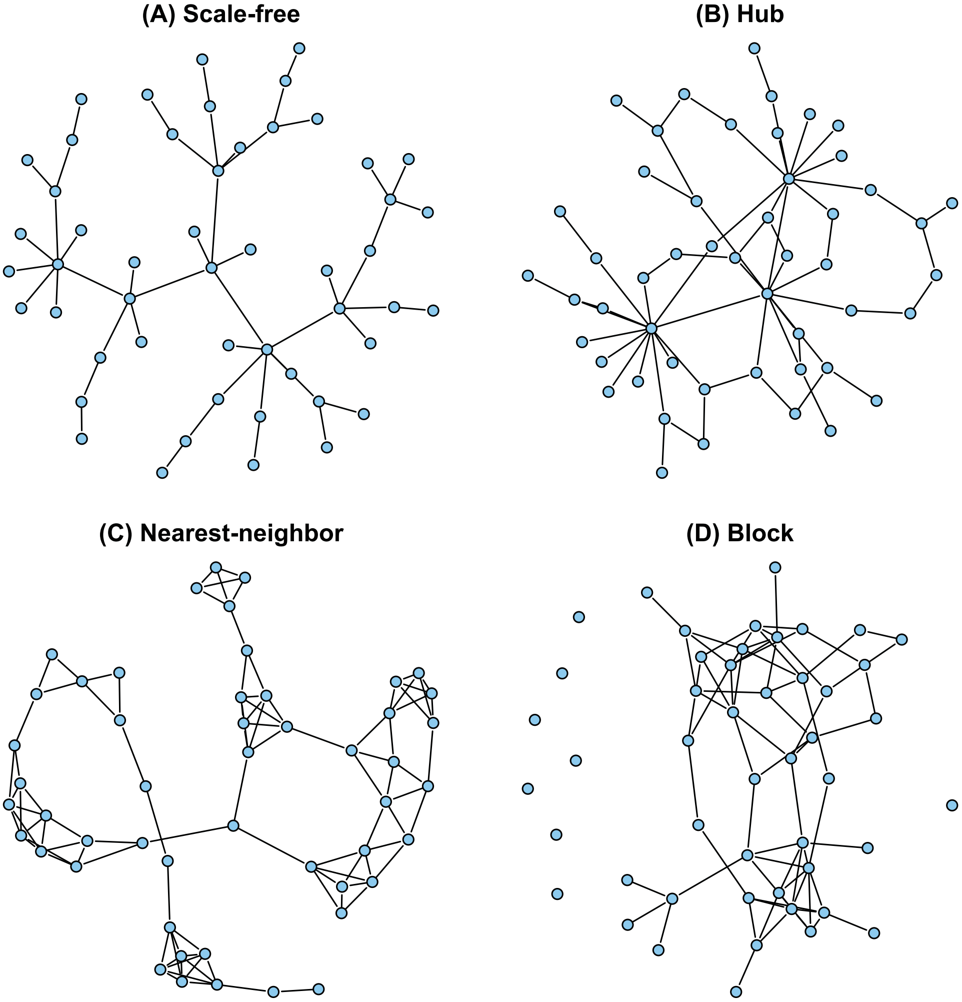

Figure 2. Illustration of the networks used in four simulated examples: Scale-free graph, hub graph, nearest-neighbor graph, and block graph.

whereas for all other pairs of nodes the probability of having a link is 0.02. Figure 2(d) illustrates such a random graph.

The ordinal data are generated as follows. First, we generate the inverse covariance matrix $\mathbf{\Omega}$ of the latent multivariate Gaussian distribution. Specifically, each off-diagonal element $\omega_{j,j'}$ is drawn uniformly from $[-1, -0.5] \cup [0.5, 1]$ if nodes j and j' are linked by an edge, otherwise $\omega_{j,j'} = 0$. Further, the diagonal elements were all set to be 2 to ensure positive definiteness, and the corresponding covariance matrix is scaled so that all the variances are equal to 1. Second, we generate the latent data $\boldsymbol{z}_i = (z_{i,1}, \dots, z_{i,p})$ as an iid sample from $N(\mathbf{0}, \mathbf{\Sigma})$. Finally, the continuous latent data $\boldsymbol{z}_i$'s are discretized into ordinal scale with three levels by thresholding. Specifically, for each $j = 1, \dots, p$, we set

$$\theta_k^{(j)} = \begin{cases} -\infty, & \text{if } k = 0; \\ \Phi^{-1}(0.1) & \text{if } k = 1; \\ \Phi^{-1}(0.2) & \text{if } k = 2; \\ +\infty, & \text{if } k = 3; \end{cases} \tag{23}$$

and set $x_{i,j} = \sum_{k=0}^{2} \mathrm{I}(z_{i,j} \geq \theta_k^{(j)})$ $(i = 1, \dots, n; j = 1, \dots, p)$. For each example, we considered different sample sizes, with $n = 50$, 100, 200, and 500.

We compare the proposed probit graphical model with two other methods. One consists of direct application of the graphical lasso to the ordinal data $\boldsymbol{X}$, ignoring their discrete nature. The second uses the graphical lasso on the latent continuous data $\boldsymbol{Z}$. We refer to the first one as the naive method and the second one as an oracle method because it represents an ideal situation where $\boldsymbol{Z}$ is exactly recovered. Of course, the latter never occurs with real data, but serves as a benchmark for comparison purposes. The receiver operating characteristic curve (ROC) was used to evaluate the accuracy of network structure estimation. The ROC curve plots the sensitivity (the proportion of correctly detected links) against the false positive rate (the proportion of misidentified zeros) over a range of values of the tuning parameter λ. The sensitivity and the false positive rate are defined as follows:

$$\text{Sensitivity} = \frac{\sum_{1 \leq j < j' \leq p} \mathcal{I}(\omega_{j,j'} \neq 0, \widehat{\omega}_{j,j'} \neq 0)}{\sum_{1 \leq j < j' \leq p} \mathcal{I}(\omega_{j,j'} \neq 0)}, \tag{24}$$

$$\text{False positive rate} = \frac{\sum_{1 \leq j < j' \leq p} \mathcal{I}(\omega_{j,j'} = 0, \widehat{\omega}_{j,j'} \neq 0)}{\sum_{1 \leq j < j' \leq p} \mathcal{I}(\omega_{j,j'} = 0)}, \tag{25}$$

where $\mathcal{I}(\cdot)$ is an indicator function whose value is 1 if the statement in the parenthesis is true, and is 0 if it is false. In addition, the Frobenius loss and the entropy loss defined in (21) were used to evaluate the performance of parameter estimation.

Figure 3 shows the ROC curves for all simulated examples. The curves are averaged over 50 replications. The oracle method provides a benchmark curve for each setting (blue dotted line in each panel). We can see that when the sample size is relatively small ($n =$ 50, 100, or 200), the probit model (dark solid line) dominates the naive method (red dashed line). When the sample size gets larger, the two methods exhibit similar performance.

Table 2 summarizes the parameter estimation measured by the Frobenius loss and the entropy loss. The results were again averaged over 50 repetitions and the tuning parameter λ was selected using the cross-validation introduced in Section 2.4. The oracle method evidently performs the best, as it should. Comparing the two methods based on the observed data $\boldsymbol{X}$, we can see that the Frobenius losses from the probit model are consistently lower than those from the naive method. The advantage is more significant when the sample size is moderate ($n = 100$ or 200). In terms of the entropy loss, we can see that the probit model outperforms the naive method for relatively large sample sizes, such as $n = 200$ and 500.

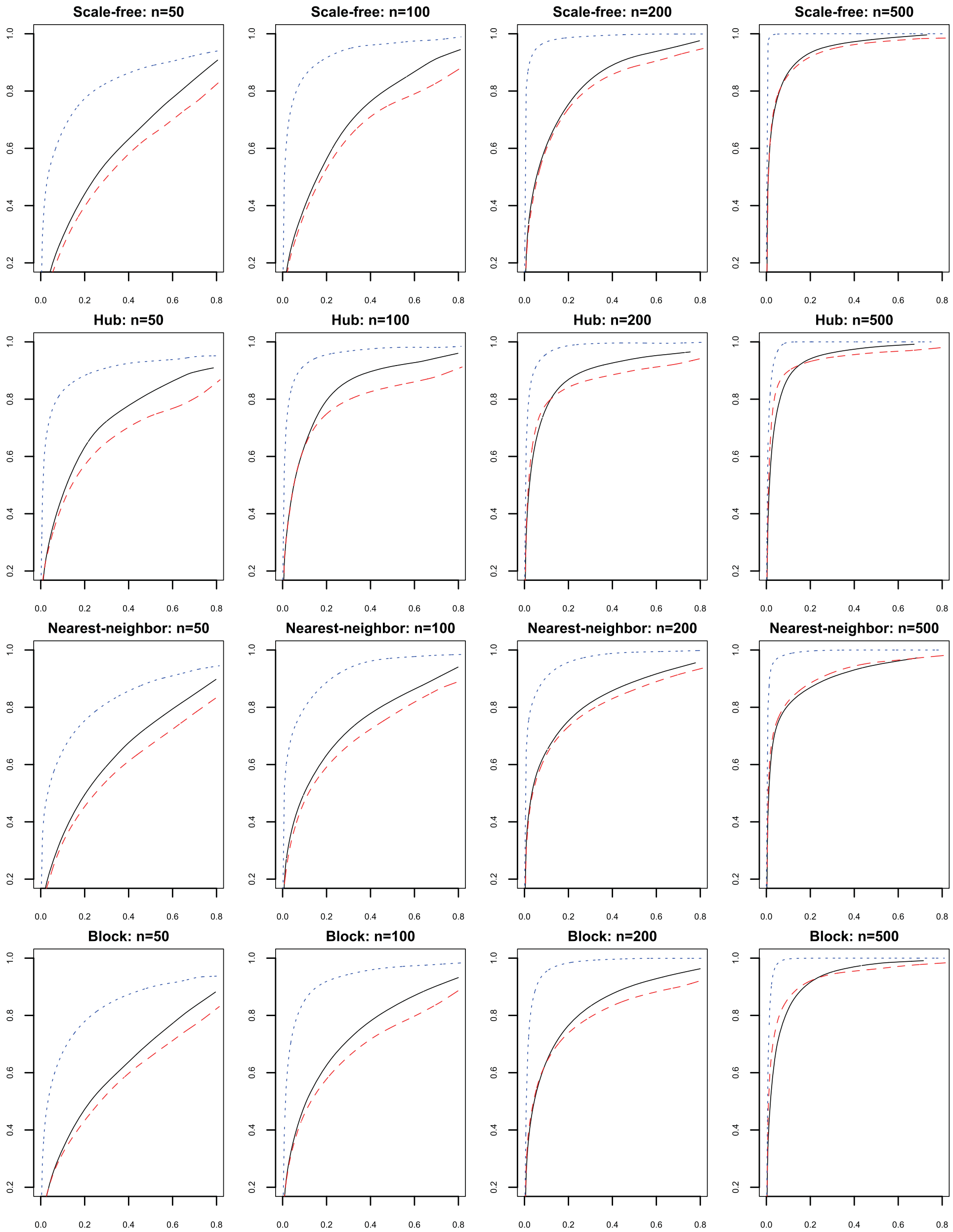

Figure 3. The ROC curves estimated by the probit graphical model (solid dark line), the oracle method (dotted blue line), and the naive method (dashed red line). The oracle method and the naive method simply apply the graphical lasso algorithm to the latent continuous data $\boldsymbol{Z}$ and the observed discrete data $\boldsymbol{X}$, respectively.

Table 2. The Frobenius loss and the entropy loss estimated by the probit graphical model, the oracle method and the naive method

		Frobenius loss			Entropy loss		
Example	n	Gaussian	Oracle	Probit	Gaussian	Oracle	Probit
Scale-free	50	2.3 (0.12)	0.7 (0.05)	2.2 (0.13)	12.0 (0.73)	3.1 (0.29)	23.1 (1.83)
	100	2.2 (0.13)	0.4 (0.08)	1.7 (0.09)	9.4 (0.68)	1.9 (0.29)	10.1 (0.45)
	200	1.7 (0.12)	0.3 (0.02)	1.2 (0.04)	6.4 (0.33)	1.1 (0.10)	5.4 (0.26)
	500	0.9 (0.05)	0.1 (0.01)	0.7 (0.04)	3.3 (0.19)	0.5 (0.05)	2.7 (0.19)
Hub	50	1.2 (0.06)	0.3 (0.02)	1.1 (0.04)	21.2 (1.32)	5.8 (0.70)	29.4 (1.76)
	100	1.1 (0.10)	0.1 (0.01)	0.8 (0.03)	15.9 (1.03)	3.2 (0.27)	15.1 (0.64)
	200	0.8 (0.05)	0.1 (0.01)	0.6 (0.01)	11.9 (0.39)	1.8 (0.23)	10.4 (0.33)
	500	0.6 (0.02)	0.0 (0.00)	0.5 (0.01)	9.1 (0.16)	0.7 (0.06)	7.5 (0.16)
Nearest-neighbor	50	1.4 (0.04)	0.6 (0.02)	1.3 (0.06)	16.5 (0.80)	5.6 (0.30)	25.6 (2.04)
	100	1.3 (0.08)	0.4 (0.02)	1.0 (0.02)	12.1 (0.52)	3.5 (0.36)	12.4 (0.76)
	200	1.0 (0.04)	0.2 (0.01)	0.7 (0.03)	8.6 (0.32)	2.0 (0.11)	7.5 (0.17)
	500	0.6 (0.03)	0.1 (0.01)	0.5 (0.02)	5.5 (0.12)	0.8 (0.02)	4.5 (0.19)
Random-block	50	1.8 (0.05)	0.7 (0.05)	1.7 (0.04)	14.8 (1.04)	4.7 (0.46)	23.5 (1.76)
	100	1.6 (0.16)	0.4 (0.02)	1.3 (0.03)	10.7 (1.10)	2.9 (0.27)	11.3 (0.46)
	200	1.3 (0.05)	0.2 (0.03)	0.9 (0.05)	7.2 (0.19)	1.6 (0.11)	6.3 (0.32)
	500	0.7 (0.03)	0.1 (0.01)	0.6 (0.03)	4.1 (0.15)	0.7 (0.06)	3.5 (0.13)

NOTE: The oracle method and the naive method simply apply the graphical lasso algorithm to the latent continuous data $\boldsymbol{Z}$ and the observed discrete data $\boldsymbol{X}$, respectively. The results are averaged over 50 repetitions and the corresponding standard deviations are recorded in the parentheses.

4. DATA EXAMPLES

4.1 Application to Movie Rating Records

In this section, we apply the probit graphical model to Movielens, a dataset containing rating scores for 1682 movies by 943 users. The rating scores have five levels, where 1 corresponds to strong dissatisfaction and 5 to strong satisfaction. More than 90% of the entries are missing in the full data matrix; for this reason, we consider a subset of the data containing 193 users and 32 movies, with 15% missing values. The missing values were imputed by the median of the observed movie ratings.

The estimated network for these 32 movies is shown in Figure 4. We can see that the estimated network consists of a large connected community as well as a few isolated nodes. The large community mainly consists of mass marketed commercial movies, dominated by science fiction, and action films. These movies are characterized by high production budgets, state of the art visual effects, and famous directors and actors. Examples in this data subset include the Star Wars franchise (“Star Wars” (1977), “The Empire Strikes Back” (1980), and “Return of the Jedi” (1983), directed/produced by Lucas), the Terminator series (1984, 1991) directed by Cameron, the Indiana Jones franchise (“Raiders of Lost Ark” (1981), “The Last Crusade” (1989), directed by Spielberg), the Alien series, etc. As expected, movies within the same series are most strongly associated. Further, “Raiders of the Lost Ark” (1981) and “Back to the Future” (1985) form two hub nodes each having 16 connections to other movies and their common feature is that they were directed/produced by Spielberg.

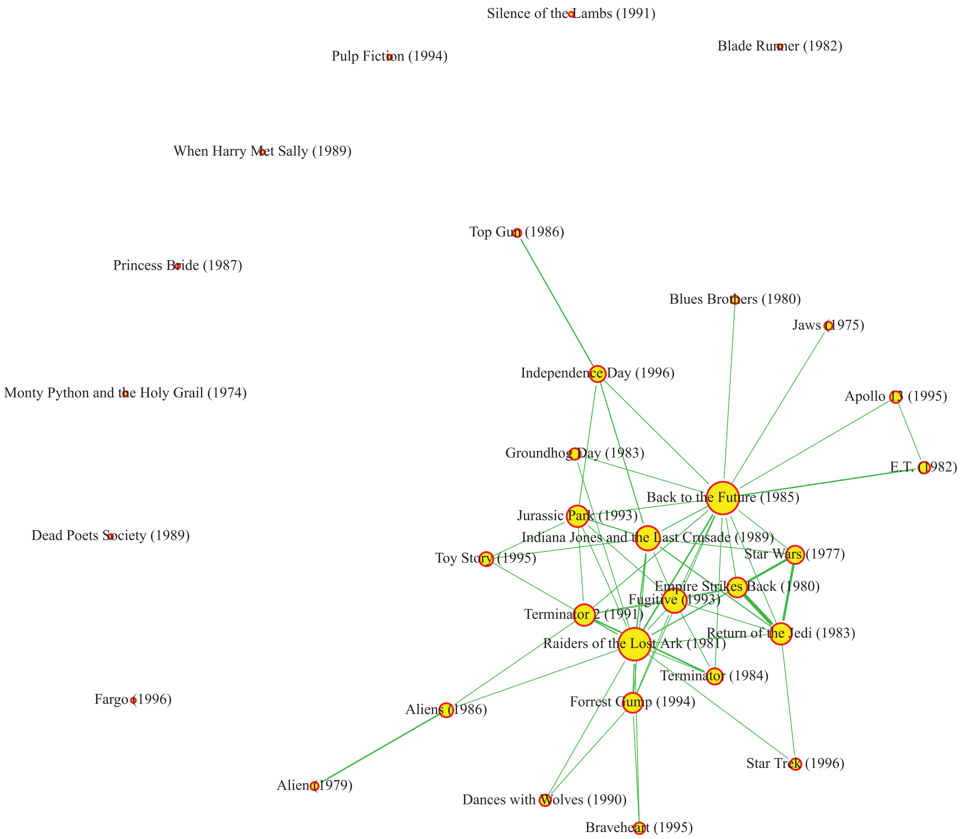

Figure 4. The network estimated by the probit graphical model. The nodes represent the movies labeled by their titles. The area of a node is proportional to its degree and the width of a link is proportional to the magnitude of the corresponding partial correlations.

On the other hand, isolated nodes tend to represent “artsier” movies, such as crime films and comedies whose popularity relies more on the plot and the cast than on big budgets and special effects, many with cult status among their followers. Examples include “Pulp Fiction” (1994) (one of the most popular Tarantino movies), “Fargo” (1996) (a quintessential Coen brothers movie), “When Harry Met Sally” (1989), and “Princess Bride” (1987). These films have no significant connections in the network, either with each other or with the commercial movies in the large community. This is likely due to two reasons: (1) we restricted the dataset to movies rated by a substantial fraction of the users, so while there probably are connections from “Fargo” to other Coen brothers movies, the other ones did not appear in this set; and (2) there is a greater heterogeneity of genres in this set than among the large group of science-fiction and action films. In other words, liking “When Harry Met Sally” (a romantic comedy) does not make one more likely to enjoy “Silence of the Lambs” (a thriller/horror movie), whereas liking “Terminator” suggests you are more likely to enjoy “Alien.” A more complete analysis of this dataset is an interesting topic for future work and requires a more sophisticated way of dealing with missing data, which is not the focus of the current article.

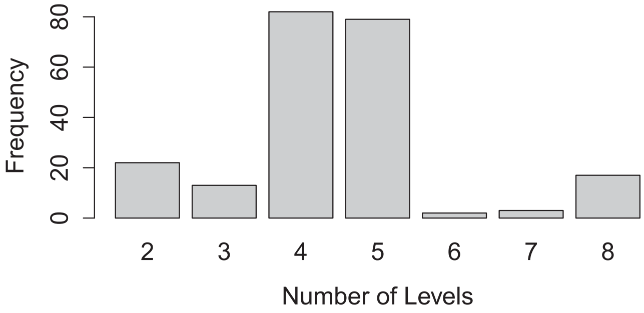

Figure 5. Histogram of the number of options in 218 survey questions.

4.2 National Education Longitudinal Survey Study

The data for the second example come from the National Educational Longitudinal Study of 1988 (NELS:88), whose objective was to assess student attitudes toward a number of questions about their school, education, and activities. The data used were obtained from the study's website *http://nces.ed.gov/surveys/nels88/* and correspond to a sample of 12,144 students of eighth-graders. We selected 218 questions with ordinal and/or binary responses that focused on diverse issues, including school work and home experiences, educational and occupational aspirations, access to educational resources and other support, as well as student background and school characteristics. Ordinal responses were chosen from the following options: "OFTEN", "SOMETIMES," "RARELY," and "NEVER," while binary ones corresponded to a "YES/NO" answer. Figure 5 depicts the histogram of the frequency of options in 218 survey questions.

The estimated network of the selected 218 survey questions is shown in Figure 6. It is apparent that the estimated network exhibits a strong clustering structure. For example, the set of the following nodes "F1S33A," "F1S33B," "F1S33C," "F1S33D," and "F1S33E" forms a cluster, separated from the remaining nodes. These five questions are a part of a sequence of similar questions, focusing on vocational coursework. Specifically, the question inquires whether "In your most recent or current VOCATIONAL course, how much emphasis did/does your teacher place on the following objectives?" and the specific objectives are listed in Table 3. It can be seen that questions "F1S33A"–"F1S33E" reflect different aspects of knowledge and analytical ability that a student should acquire from a vocational course, and therefore it is reasonable that they form a tight cluster. Similar

Table 3. Objectives in survey questions "In your most recent or current vocational course, how much emphasis did/does your teacher place on the following objectives?"

F1S33A	Teaching you skills you can use immediately
F1S33B	Teaching you facts, rules, and steps
F1S33C	Helping you understand how scientific ideas and mathematics are used in work
F1S33D	Thinking about what a problem means and the ways it might be solved
F1S33E	Helping you to understand mathematical and scientific ideas by helping you to manipulate physical objects (tools, machines, lab equipment)

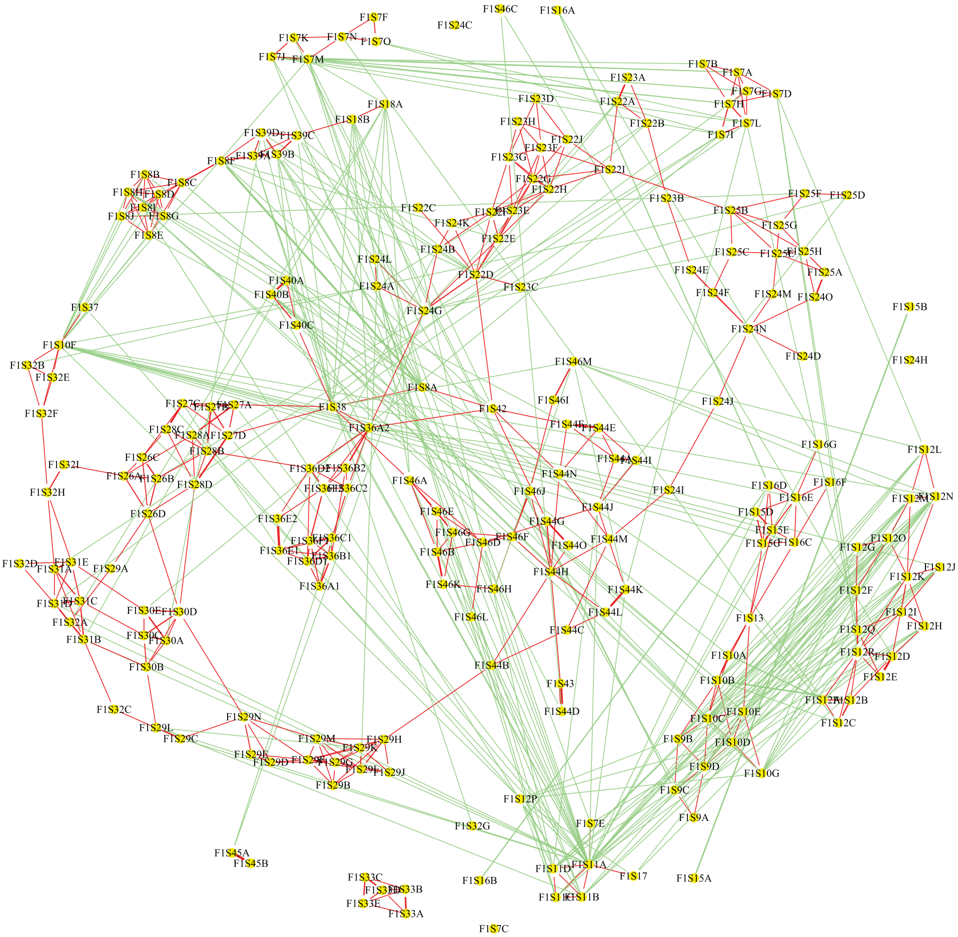

Figure 6. Layout of the network estimated by the proposed probit graphical model. The nodes represent the survey questions labeled by their code. The area of a node is proportional to its degree and the width of a link is proportional to the magnitude of the corresponding partial correlations. The red lines represent positive associations, while the light green lines negative ones.

clustering patterns can be observed in other parts of the graph, for example, serial "F1S7," serial "F1S8," serial "F1S12," serial "F1S25," etc.

Next, we focus on broad patterns revealed by the model, as depicted in Figure 6. The upper right corner captures relationships between serial questions broadly related to coursework (F1S22-F1S25) in various disciplines (mathematics, science, English, computer education), whereas in the lower left corner there are questions related to overall attitude and study patterns regarding mathematics and science classes (F1S26-F1S32). It is interesting to observe that the model does not discover any relationships between these two question clusters. In the center of the plot we find questions related to various life aspects and being successful/accomplishing them (F1S46) which is negatively associated with a cluster of questions related to working hard in school for good grades (F1S11). In the center, we also find a cluster of serial questions related to different ways of interacting with friends (F1S44)

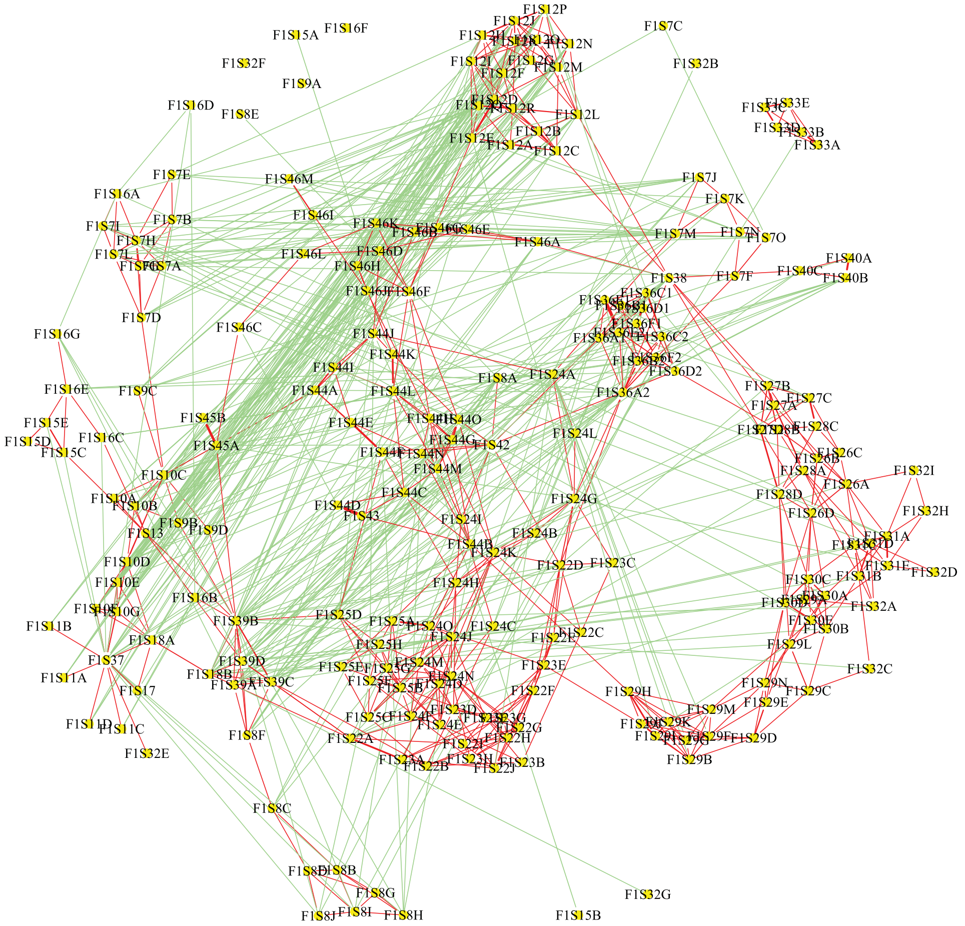

Figure 7. Layout of the estimated network by the graphical lasso algorithm. The nodes represent the survey questions labeled by their code. The area of a node is proportional to its degree and the width of a link is proportional to the magnitude of the corresponding partial correlations. The red lines represent positive associations, while the light green lines negative ones.

which is negatively correlated to questions related to students awards (F1S8). In the upper left corner we see the serial cluster on grades performance (F1S39) which is also negatively correlated with some of the questions related to amount of coursework in various subjects (F1S22 and F1S24). Finally, in the bottom right corner we encounter questions related to school attendance and attitude toward it (F1S10, F1S12).

Overall, the model reveals interesting and informative patterns, much more so than its Gaussian counterpart shown in Figure 7.

Next, we examined pairs of questions exhibiting the largest positive partial correlations (based on the theory of Gaussian graphical models, the partial correlation of variables j and j' is defined as $\rho_{j,j'} = -\omega_{j,j'}/\sqrt{\omega_{j,j}\omega_{j',j'}}$). The results are shown in Table 4. Among the top five ones, four pairs correspond to serial questions. The only exception is pair “F1S44D–F1S43,” although it inquires about extra reading, outside school. Analogously, Table 5 lists the pair of questions exhibiting the strongest negative partial

Table 4. List of pairs of questions with strongest positive partial correlations

Connection	Partial correlation	Description
F1S44D–F1S43	0.617981	F1S44D: How often do you spend time on reading for pleasure? F1S43: How much additional reading do you do each week on your own outside of school—not in connection with schoolwork?
F1S45A–F1S45B	0.443995	F1S45A: During the school year, how many hours a day do you on weekdays? F1S45B: During the school year, how many hours a day do you on weekends?
F1S36E1–F1S36E2	0.416786	F1S36E1: How much time do you spend on History homework in school each week? F1S36E2: How much time do you spend on History homework out of school each week?
F1S44E–F1S44F	0.398257	F1S44E: How often do you spend time on going to the park, gym, beach, or pool outside of school? F1S44F: How often do you spend time on playing ball or other sports with friends outside of school?
F1S12D–F1S12E	0.388861	F1S12D: How often do you feel it is "OK" for you to cheat on tests? F1S12E: How often do you feel it is "OK" for you to copy someone else's homework?

correlations. Note that question pairs "F1S8F–F1S8A," "F1S15B–F1S15A," "F1S16B–F1S16D" are composed of two opposite questions. It is interesting to observe that the model identifies the pair 'F1S10B–F1S12B," which can be interpreted that although students may skip class often they do not feel good about their action. A similar negative partial correlation is present in pair "F1S10A–F1S12A" that addresses a "coming to school late" issue.

Table 5. The list of pairs of questions with strongest negative partial correlations

Connection	Partial correlation	Description
F1S8F–F1S8A	−0.376025	F1S8F: Did you win any special recognition for good grades or honor roll? F1S8A: Haven't you won any awards or received recognition?
F1S10B–F1S12B	−0.281428	F1S10B: How many times did you cut or skip classes? F1S12B: How often do you feel it is "OK" for you to cut a couple of classes?
F1S15B–F1S15A	−0.259550	F1S15B: During your last absence from school, did anyone from the school call your home? F1S15A: The school did not do anything on your last absence from school.
F1S10A–F1S12A	−0.216677	F1S10A: How many times were you late for school in the first half of the current school year? F1S12A: How often do you feel it is "OK" for you to be late for school?
F1S16B–F1S16D	−0.214770	F1S16B: When you came back to school after your last absence, other students helped you catch up on the work you missed. F1S16D: When you came back to school after your last absence, you didn't need to catch up on work.

Overall, the proposed model identifies strong clustering patterns in the questions being asked in this survey, which primarily correspond to series of related in intent and purpose questions, thus indirectly validating its usefulness.

5. SUMMARY AND DISCUSSION

Ordinal data occur often in practice and are usually treated as continuous for most analyses, including estimating dependencies between the variables under consideration by fitting a graphical model. Our proposed model, explicitly takes into account the ordinal nature of the data in the graphical modeling step. While direct computation for the proposed model is expensive, the approximations employed allow us to efficiently fit high-dimensional models. On those datasets that the model can be fitted directly, our numerical results show that the approximations we make result in a minimal loss of accuracy. We leave the theoretical properties of both the exact estimator and its approximate version as a topic for future work.

The method proposed in this article can also be extended to fit the multivariate ordinal regression model, where multiple ordinal responses are fitted on a number of covariates. Specifically, suppose $W_{j1}, \ldots, W_{jm_j}$ are the covariates associated with the jth response. Following the notation in Section 2.1, let X_j denote the jth response, which is an ordinal variable, and Z_j the corresponding latent continuous variable. We may assume $Z_j = \alpha_{j0} + \alpha_{j1} W_{j1} + \cdots + \alpha_{jm_j} W_{jm_j} + \epsilon_j$, where α_{j0} is the intercept, and $\alpha_{j1}, \ldots, \alpha_{jm_j}$ are regression coefficients. In addition, we assume that $\epsilon_1, \ldots, \epsilon_p$ jointly follow a Gaussian distribution with mean zero and covariance matrix $\mathbf{\Sigma} = \mathbf{\Omega}^{-1}$. To estimate the regression coefficients, we may modify the M-step in Section 2.2 to estimate $\mathbf{\Omega}$ and $\alpha_{j\ell}$'s simultaneously. Rothman, Levina, and Zhu (2010) discussed a similar problem as the modified M-step, and the algorithm there can be directly applied.

ACKNOWLEDGMENTS

E. Levina's research is partially supported by NSF grants DMS-0805798, DMS-1106772, and DMS-1159005. G. Michailidis's research is partially supported by NIH grant 1RC1CA145444-0110. J. Zhu's research is partially supported by NSF grant DMS-0748389 and NIH grant R01GM096194.

[Received July 2012. Revised July 2013.]

REFERENCES

Albert, J., and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679. [184,185]

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008), "Model Selection Through Sparse Maximum Likelihood Estimation," *Journal of Machine Learning Research*, 9, 485–516. [183]

Barabasi, A.-L., and Albert, R. (1999), "Emergence of Scaling in Random Networks," *Science*, 286, 509–512. [192]

Bliss, C. (1935), "The Calculation of the Dosage-Mortality Curve," *Annals of Applied Biology*, 22, 134–167. [184,185]

Chib, S., and Greenberg, E. (1998), "Analysis of Multivariate Probit Models," *Biometrika*, 85, 347–361. [184,185]

Fan, J., Feng, Y., and Wu, Y. (2009), “Network Exploration via the Adaptive LASSO and SCAD Penalties,” *Annals of Applied Statistics*, 3, 521–541. [183]

Friedman, J., Hastie, T., and Tibshirani, R. (2008), “Sparse Inverse Covariance Estimation With the Graphical Lasso,” *Biostatistics*, 9, 432–441. [183,187]

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2010), “Joint Structure Estimation for Categorical Markov Networks,” Technical Report No. 507, Department of Statistics, University of Michigan, Ann Arbor. [184]

Höefling, H., and Tibshirani, R. (2009), “Estimation of Sparse Binary Pairwise Markov Networks Using Pseudo-Likelihoods,” *Journal of Machine Learning Research*, 10, 883–906. [184]

Johnson, N., Kotz, S., and Balakrishnan, N. (1994), *Continuous Univariate Distributions* (2nd ed., Vol. 1), New Jersey: Wiley. [188]

Kolar, M., and Xing, E. (2008), “Improved Estimation of High-Dimensional Ising Models,” available at *Eprint arXiv:0811.1239*. [184]

Koren, Y., Bell, R., and Volinsky, C. (2009), “Matrix Factorization Techniques for Recommender Systems,” *IEEE Computer*, 42, 30–37. [184]

Kotecha, J., and Djuric, P. (1999), “Gibbs Sampling Approach for Generation of Truncated Multivariate Gaussian Random Variables,” *IEEE Computer Society*, 3, 1757–1760. [186]

Lauritzen, S. (1996), *Graphical Models*, Oxford: Oxford University Press. [183]

Lee, L.-F. (1979), “On the First and Second Moments of the Truncated Multi-Normal Distribution and a Simple Estimator,” *Economics Letters*, 3, 165–169. [186]

Leppard, P., and Tallis, G. (1989), “Evaluation of the Mean and Covariance of the Truncated Multinormal,” *Applied Statistics*, 38, 543–553. [186]

Li, H., and Gui, J. (2006), “Gradient Directed Regularization for Sparse Gaussian Concentration Graphs, With Applications to Inference of Genetic Networks,” *Biostatistics*, 7, 302–317. [192]

Lian, H. (2011), “Shrinkage Tuning Parameter Selection in Precision Matrices Estimation,” *Journal of Statistical Planning and Inference*, 14, 2839–2848. [190]

Manjunath, B., and Wilhelm, S. (2012), “Moments Calculation for the Doubly Truncated Multivariate Normal Density,” available at *ArXiv e-prints*. [186]

McCullagh, P. (1980), “Regression Models for Ordinal Data,” *Journal of the Royal Statistical Society,* Series B, 42, 109–142. [184]

McCullagh, P., and Nelder, J. (1989), *Generalized Linear Models* (2nd ed.), London, UK: Chapman and Hall/CRC. [184]

Meinshausen, N., and Buhlmann, P. (2006), “High-Dimensional Graphs With the Lasso,” *The Annals of Statistics*, 34, 1436–1462. [183]

O’Connell, A. (2005), *Logistic Regression Models for Ordinal Response Variables* (1st ed.), Thousand Oaks, CA: Sage Publications, Inc. [184]

Pakman, A., and Paninski, L. (2012), “Exact Hamiltonian Monte Carlo for Truncated Multivariate Gaussians,” available at *ArXiv e-prints*. [191]

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009), “Partial Correlation Estimation by Joint Sparse Regression Model,” *Journal of the American Statistical Asscociation*, 104, 735–746. [184,192]

Peterson, B. (1990), “Partial Proportional Odds Models for Ordinal Response Variables,” *Applied Statistics*, 39, 205–217. [184]

Peterson, C., and Anderson, J. (1987), “A Mean Field Theory Learning Algorithm for Neural Networks,” *Complex Systems*, 1, 995–1019. [187]

Ravikumar, P., Wainwright, M., and Lafferty, J. (2010), “High-Dimensional Ising Model Selection Using ℓ_1-Regularized Logistic Regression,” *The Annals of Statistics*, 38, 1287–1319. [184]

Rocha, G., Zhao, P., and Yu, B. (2008), “A Path Following Algorithm for Sparse Pseudo-Likelihood Inverse Covariance Estimation (SPLICE),” Technical Report, arXiv:0807.3734, Department of Statistics, University of California, Berkeley. [184]

Rothman, A., Bickel, P., Levina, E., and Zhu, J. (2008), "Sparse Permutation Invariant Covariance Estimation," *Electronic Journal of Statistics*, 2, 494–515. [183,187]

Rothman, A., Levina, L., and Zhu, J. (2010), "Sparse Multivariate Regression With Covariance Estimation," *Journal of Computational and Graphical Statistics*, 19, 947–962. [202]

Stern, D., Herbrich, R., and Graepel, T. (2009), "Matchbox: Large Scale Online Bayesian Recommendations," in *Proceedings of World Wide Web 2009, Madrid, Spain*, pp. 111–120. [184]

Tallis, G. (1961), "The Moment Generating Function of the Truncated Multinormal Distribution," *Journal of the Royal Statistical Society,* Series B, 23, 223–229. [186]

von Davier, M., and Carstensen, C. (2010), *Multivariate and Mixture Distribution Rasch Models: Extensions and Applications* (1st ed.), New York: Springer. [184,185]

Walker, S., and Duncan, D. (1967), "Estimation of the Probability of an Event as a Function of Several Independent Variables," *Biometrika*, 54, 167–179. [184]

Yuan, M., and Lin, Y. (2007), "Model Selection and Estimation in the Gaussian Graphical Model," *Biometrika*, 94, 19–35. [183]