ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

*J. R. Statist. Soc.* B (2017)
**79**, *Part* 2, *pp.* 405–421

# High dimensional semiparametric latent graphical model for mixed data

Jianqing Fan, Han Liu and Yang Ning

*Princeton University, USA*

and Hui Zou

*University of Minnesota, Minneapolis, USA*

[Received April 2014. Final revision December 2015]

**Summary.** We propose a semiparametric latent Gaussian copula model for modelling mixed multivariate data, which contain a combination of both continuous and binary variables. The model assumes that the observed binary variables are obtained by dichotomizing latent variables that satisfy the Gaussian copula distribution. The goal is to infer the conditional independence relationship between the latent random variables, based on the observed mixed data. Our work has two main contributions: we propose a unified rank-based approach to estimate the correlation matrix of latent variables; we establish the concentration inequality of the proposed rank-based estimator. Consequently, our methods achieve the same rates of convergence for precision matrix estimation and graph recovery, as if the latent variables were observed. The methods proposed are numerically assessed through extensive simulation studies, and real data analysis.

*Keywords*: Discrete data; Gaussian copula; Latent variable; Mixed data; Non-paranormal; Rank-based statistic

## 1. Introduction

Graphical models (Lauritzen, 1996) have been widely used to explore the dependence structure of multivariate distributions, arising in many research areas including machine learning, image analysis, statistical physics and epidemiology. In these applications, the data that are collected often have high dimensionality and low sample size. Under this high dimensional setting, parameter estimation and edge structure learning in the graphical model attract increasing attention in statistics. Owing to mathematical simplicity and wide applicability, Gaussian graphical models have been extensively studied by Meinshausen and Bühlmann (2006), Yuan and Lin (2007), Rothman *et al.* (2008), Friedman *et al.* (2008, 2010), d'Aspremont *et al.* (2008), Rocha *et al.* (2008), Fan *et al.* (2009), Peng *et al.* (2009), Lam and Fan (2009), Yuan (2010), Cai *et al.* (2011) and Zhang and Zou (2014), among others. To relax the Gaussian model assumption, Xue and Zou (2012) and Liu *et al.* (2009, 2012) proposed a semiparametric Gaussian copula model for modelling continuous data by allowing for monotonic univariate transformations. Recently, there has been a large body of work in the machine learning literature focusing on the computational aspect of graphical model estimation; see Hsieh *et al.* (2011, 2013), Rolfs *et al.* (2012), Oztoprak *et al.* (2012) and Treister and Turek (2014), among others.

*Address for correspondence*: Han Liu, Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA.
E-mail: hanliu@princeton.edu

1369–7412/17/79405

Both Gaussian and Gaussian copula models are tailored only for modelling continuous data. However, many multivariate problems may contain discrete data or data of hybrid types with both discrete and continuous variables. For instance, genomic data such as DNA nucleotides data may take binary values. In social science, covariate information that is collected from sample surveys often contains both continuous and discrete variables. For binary data, Xue *et al.* (2012), Höfling and Tibshirani (2009) and Ravikumar *et al.* (2010) proposed a penalized pseudolikelihood approach under the Ising model. Recently, there has been a sequence of work studying the mixed graphical model. For instance, Lee and Hastie (2014) proposed a penalized composite likelihood method for pairwise graphical models with mixed Gaussian and multinomial data. Later, Cheng *et al.* (2013) extended the model by incorporating further interaction terms. A non-parametric approach based on random forests was proposed by Fellinghauer *et al.* (2013). Recently, Yang *et al.* (2014a) and Chen *et al.* (2015) proposed exponential family graphical models, which allow the conditional distribution of nodes to belong to the exponential family. Later, a semiparametric exponential family graphical model was studied by Yang *et al.* (2014b).

In many applications, it is often reasonable to assume that the discrete variable is obtained by discretizing a latent (unobserved) variable (Skrondal and Rabe-Hesketh, 2007). For instance, in psychology, the latent variables can represent abstract concepts such as human feeling or recognition that exist in hidden form but are not directly measurable and, instead, they can be measured indirectly by some surrogate variables. In the analysis of gene expression data, there is often unwanted variation between different experiments which is known as batch effects (McCall *et al.*, 2014; Lazar *et al.*, 2013). To remove them, a commonly used procedure is to dichotomize the numerical expression data into 0–1 binary data (McCall and Irizarry, 2011). In many social science studies, the responses are often collected from a survey, which may take the form of yes–no or categorical answers. Since in all these applications the existence of latent variables seems reasonable, the modelling of these types of discrete data can be improved by incorporating this latent variable structure.

In this paper, we consider a generative modelling approach and propose a latent Gaussian copula model for mixed data. The model assumes that the observed discrete data are generated by dichotomizing a latent continuous variable at some unknown cut-off. In addition, the latent variables for the binary components combined with the observed continuous variables jointly satisfy the Gaussian copula distribution. The model proposed extends the Gaussian copula model (Xue and Zou, 2012; Liu *et al.*, 2009, 2012) and the latent Gaussian model (Han and Pan, 2012) to account for mixed data. In this modelling framework, our goal is to infer the conditional independence structure between latent variables, which provides deeper understandings of the unknown mechanism than that between the observed binary surrogates. Under the latent Gaussian copula model, the conditional independence structure is characterized by the sparsity pattern of the latent precision matrix.

Our work has two major contributions. Our first contribution is to propose a unified rank-based estimation procedure. The framework proposed extends the existing rank-based method by Xue and Zou (2012) and Liu *et al.* (2012) to a more challenging setting with mixed data. To the best of our knowledge, this paper for the first time proposes such a generalized notion of a rank-based estimator for mixed data. Given the new rank-based estimator, the existing graph estimation procedures, such as the graphical lasso (Friedman *et al.*, 2008), the constrained $l_1$-minimization for inverse matrix estimation estimator CLIME (Cai *et al.*, 2011) and the adaptive graphical lasso (Lam and Fan, 2009), can be directly used to infer the latent precision matrix. Our second contribution is to establish concentration inequalities for the generalized rank-based estimator. Based on this result, the estimator of the precision matrix achieves the same rates of convergence and model selection consistency, as if the latent variables were observed.

Compared with existing methods for mixed data, our model and estimation procedures are different. The work by Lee and Hastie (2014), Cheng *et al.* (2013), Yang *et al.* (2014a) and Chen *et al.* (2015) essentially models the nodewise conditional distribution by generalized linear models. In contrast, the latent Gaussian copula model is a generative model which combines continuous and discrete data through a deeper layer of unobserved variables. In addition, the model is semiparametric and allows more complicated joint distributions of continuous and discrete data. The existing methods by Lee and Hastie (2014), Cheng *et al.* (2013), Yang *et al.* (2014a) and Chen *et al.* (2015) cannot offer such flexibility for modelling the interaction between the mixed variables. Compared with the non-parametric approach in Fellinghauer *et al.* (2013), our semiparametric approach can be much more efficient, which is demonstrated through extensive numerical studies. A composite likelihood method was proposed by Han and Pan (2012) for latent Gaussian models. However, such an approach cannot be applied to mixed data in high dimensional settings, owing to the high computational cost for maximizing the composite likelihood. Instead, our rank-based estimation method is computationally much more convenient.

The rest of the paper is organized as follows. In Section 2, we review the Gaussian copula model. In Section 3, we define the latent Gaussian copula model for mixed data. In Section 4, we propose a general rank-based estimation framework for mixed data. In Section 5, we consider latent graph estimation based on the rank-based approach proposed. We conduct extensive simulation studies and apply our methods to a real data example in Sections 6 and 7 respectively. Discussion and concluding remarks are presented in Section 8. The programs that were used to analyse the data can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

For the following development, we introduce some notation. Let $\mathbf{M} = (M_{jk}) \in \mathbb{R}^{d\times d}$ and $\mathbf{v} = (v_1, \ldots, v_d)^{\mathrm{T}} \in \mathbb{R}^d$ be a $d \times d$ matrix and a $d$-dimensional vector. We denote $\mathbf{v}_I$ to be the subvector of $\mathbf{v}$ whose entries are indexed by a set $I$ and $\mathbf{v}_{-I}$ to be the subvector of $\mathbf{v}$ with $\mathbf{v}_I$ removed. We define $\|\mathbf{M}\|_{\max} := \max\{|M_{ij}|\}$ as the matrix elementwise maximum norm, $\|\mathbf{M}\|_1 = \Sigma_{1\leqslant i\leqslant d}\,\Sigma_{1\leqslant j\leqslant d}\,|M_{ij}|$ as the elementwise $L_1$-norm, $\|\mathbf{M}\|_2$ as the spectral norm and $\|\mathbf{M}\|_{\mathrm{F}}$ as the Frobenius norm.

## 2. Gaussian copula model

In multivariate analysis, the Gaussian model is commonly used because of its mathematical simplicity (Lauritzen, 1996). Although the Gaussian model has been widely applied, the normality assumption is rather restrictive. To relax this assumption, Xue and Zou (2012) and Liu *et al.* (2009, 2012) proposed a semiparametric Gaussian copula model.

*Definition 1* (Gaussian copula model). A random vector $\mathbf{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$ is sampled from the Gaussian copula model, if there is a set of monotonically increasing transformations $f = (f_j)_{j=1}^d$, satisfying $f(\mathbf{X}) = (f_1(X_1), \ldots, f_d(X_d))^{\mathrm{T}} \sim N_d(0, \mathbf{\Sigma})$ with $\Sigma_{jj} = 1$ for any $1 \leqslant j \leqslant d$. Then we denote $\mathbf{X} \sim \mathrm{NPN}(0, \mathbf{\Sigma}, f)$.

Under the Gaussian copula model, the sparsity pattern of $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ encodes the conditional independence between $\mathbf{X}$. Specifically, $X_i$ and $X_j$ are independent given the remaining variables $\mathbf{X}_{-(i,j)}$ if and only if $\Omega_{ij} = 0$. Hence, inferring the graph structure under the Gaussian copula model can be accomplished by estimating $\mathbf{\Omega}$.

## 3. Latent Gaussian copula model for mixed data

Despite the flexibility of the Gaussian copula model (Xue and Zou, 2012; Liu *et al.*, 2009, 2012), it can handle only continuous data. In this section, we extend the model to account for mixed data. We call it the latent Gaussian copula model.

*Definition 2* (latent Gaussian copula model for mixed data). Assume that $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, where $\mathbf{X}_1$ represents $d_1$-dimensional binary variables and $\mathbf{X}_2$ represents $d_2$-dimensional continuous variables. The random vector $\mathbf{X}$ satisfies the latent Gaussian copula model, if there is a $d_1$-dimensional random vector $\mathbf{Z}_1 = (Z_1, \ldots, Z_{d_1})^{\mathrm{T}}$ such that $\mathbf{Z} := (\mathbf{Z}_1, \mathbf{X}_2) \sim \mathrm{NPN}(0, \boldsymbol{\Sigma}, f)$ and

$$X_j = I(Z_j > C_j) \qquad \text{for all } j = 1, \ldots, d_1,$$

where $I(\cdot)$ is the indicator function and $\mathbf{C} = (C_1, \ldots, C_{d_1})$ is a vector of constants. Then we denote $\mathbf{X} \sim \mathrm{LNPN}(0, \boldsymbol{\Sigma}, f, \mathbf{C})$, where $\boldsymbol{\Sigma}$ is the latent correlation matrix. When $\mathbf{Z} \sim N(0, \boldsymbol{\Sigma})$, we say that $\mathbf{X}$ satisfies the latent Gaussian model $\mathrm{LN}(0, \boldsymbol{\Sigma}, \mathbf{C})$.

In the latent Gaussian copula model, the 0–1 binary components $\mathbf{X}_1$ are generated by a latent continuous random vector $\mathbf{Z}_1$ truncated at some unknown constants $\mathbf{C}$. Combining with the continuous components $\mathbf{X}_2$, $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{X}_2)$ satisfies the Gaussian copula model. Owing to the flexibility of the Gaussian copula model, the distribution of the latent variable $\mathbf{Z}$ can be skewed or multimodal. We also note that the latent correlation matrix $\boldsymbol{\Sigma}$ is invariant, if $\mathbf{X}$ is a vector of binary variables and $X_j$ is recoded as $X_j^* = 1 - X_j$ for $j = 1, \ldots, d$. In other words, if $\mathbf{X} \sim \mathrm{LNPN}(0, \boldsymbol{\Sigma}, f, \mathrm{C})$ then $\mathbf{X}^* \sim \mathrm{LNPN}(0, \boldsymbol{\Sigma}, f^*, \mathbf{C}^*)$ for some $f^*$ and $\mathbf{C}^*$, where $\mathbf{X}^* = (X_1^*, \ldots, X_d^*)^{\mathrm{T}}$. We defer the details to the on-line supplementary material. Let $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ denote the latent precision matrix. Similarly to Liu *et al.* (2009), the zero pattern of $\boldsymbol{\Omega}$ characterizes the conditional independence between the latent variables $\mathbf{Z}$. Thus, our goal reduces to inferring the sparsity pattern of the latent precision matrix $\boldsymbol{\Omega}$ even though latent variables are not directly observable.

The latent Gaussian copula model suffers from the identifiability issue. To see the reason, consider the following joint probability mass function of the binary component $\mathbf{X}_1$ at a point $\mathbf{x}_1 \in \{0, 1\}^{d_1}$,

$$\mathbb{P}(\mathbf{X}_1 = \mathbf{x}_1; \mathbf{C}, \boldsymbol{\Sigma}, f) = \frac{1}{(2\pi)^{d_1/2} |\boldsymbol{\Sigma}_{11}|^{1/2}} \int_{\mathbf{u} \in U} \exp\left\{ -\frac{1}{2} \mathbf{u}^{\mathrm{T}} (\boldsymbol{\Sigma}_{11})^{-1} \mathbf{u} \right\} \mathrm{d}\mathbf{u}, \tag{3.1}$$

where $\mathbf{u} = (u_1, \ldots, u_{d_1})$ and the integration region is $U = U_1 \times \ldots \times U_{d_1}$ with $U_j = [f_j(C_j), \infty]$ if $x_j = 1$ and $U_j = [-\infty, f_j(C_j)]$ otherwise for $j = 1, \ldots, d_1$. By equation (3.1), we find that only $f_j(C_j)$ is identifiable for the binary component. For notational simplicity, we denote $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_{d_1})$, where $\Delta_j = f_j(C_j)$.

Another consequence of the identifiability constraint is that the proposed latent Gaussian copula model is equivalent to the latent Gaussian model for binary outcomes. This is expected, because the binary outcomes contain little information to identify the marginal transformations, whose effect can be offset by properly shifting the cut-off constants in the latent Gaussian model. However, when the observed variable $\mathbf{X}$ has both continuous and discrete components, the family of latent Gaussian copula models is strictly larger than the latent Gaussian model. This is because, by incorporating the marginal transformations, the joint distribution of a continuous variable $X_j$ and a discrete variable $X_k$ is more flexible. Hence, the proposed latent Gaussian copula model can better explain the association between mixed variables than the latent Gaussian model, which is the main advantage of the model proposed.

## 4. A unified rank-based estimation framework

### *4.1. Methodology*

Assume that we observe $n$ independent vector-valued data $\mathbf{X}_1,\ldots,\mathbf{X}_n \sim \mathrm{LNPN}(0,\boldsymbol{\Sigma},f,\mathbf{C})$. In this section, we propose a generalized rank-based estimator of $\boldsymbol{\Sigma}$. Because the latent variable $\mathbf{Z}_1$ is not observable, the existing rank-based method in Xue and Zou (2012) and Liu *et al.* (2012) cannot be applied to estimate $\boldsymbol{\Sigma}$. The main contribution of this paper is to introduce a unified rank-based estimation framework, which can handle mixed data.

Consider the following Kendall's $\tau$ calculated from the observed data $(X_{1j},X_{1k}),\ldots,(X_{nj},X_{nk})$:

$$\hat{\tau}_{jk}=\frac{2}{n(n-1)}\sum_{1\leqslant i<i'\leqslant n}\mathrm{sgn}\{(X_{ij}-X_{i'j})(X_{ik}-X_{i'k})\}, \tag{4.1}$$

where $X_{ij}$ and $X_{ik}$ are possibly binary components of $\mathbf{X}_i$. Here, we define $\mathrm{sgn}(0)=0$. Although $\hat{\tau}_{jk}$ quantifies certain correlation between $X_{ij}$ and $X_{ik}$, it does not directly estimate the latent correlation parameter $\Sigma_{jk}$. Our main idea is to construct a bridge function, such that it can connect Kendall's $\tau$ to $\Sigma_{jk}$. For this, we first define the population Kendall's $\tau$ as $\tau_{jk}=\mathbb{E}(\hat{\tau}_{jk})$. By equation (4.1), we can show that

$$\tau_{jk}=2\,\mathbb{P}(X_{ij}-X_{i'j}>0,X_{ik}-X_{i'k}>0)-2\,\mathbb{P}(X_{ij}-X_{i'j}>0,X_{ik}-X_{i'k}<0). \tag{4.2}$$

Since $X_{ij}-X_{i'j}>0$ is equivalent to $f_j(X_{ij})-f_j(X_{i'j})>0$ for any monotonically increasing function $f_j(\cdot)$, the right-hand side of equation (4.2) is a function of $\Sigma_{jk}$ and independent of $f$. Thus, we can denote this function by $F(\Sigma_{jk})$, where the concrete form of $F(\cdot)$ will be described case by case in the later development. We call this function $F(\cdot)$ the bridge function, since it establishes the connection between the latent correlation $\Sigma_{jk}$ and the population Kendall's $\tau$ $\tau_{jk}$. Provided that $F(\cdot)$ is invertible, we have $\Sigma_{jk}=F^{-1}(\tau_{jk})$. Therefore, a plugged-in estimator of $\Sigma_{jk}$ is given by $\hat{\Sigma}_{jk}=F^{-1}(\hat{\tau}_{jk})$.

When both $X_{ij}$ and $X_{ik}$ are continuous variables, the bridge function $F(\cdot)$ has the explicit form $F(\Sigma_{jk})=2\sin^{-1}(\Sigma_{jk})/\pi$, as shown in Kendall (1948). Thus, the rank-based estimator of $\Sigma_{jk}$, when both $X_{ij}$ and $X_{ik}$ are continuous, is

$$\hat{R}_{jk}=\sin\left(\frac{\pi}{2}\hat{\tau}_{jk}\right). \tag{4.3}$$

In what follows, we focus on the calculation of the bridge function $F(\cdot)$ on the two cases: case I, both $X_{ij}$ and $X_{ik}$ are binary variables, and case II, $X_{ij}$ is binary and $X_{ik}$ is continuous. By symmetry, the case that $X_{ij}$ is continuous and $X_{ik}$ is binary is identical to case II.

In case I, since $\mathrm{sgn}(X_{ij}-X_{i'j})=X_{ij}-X_{i'j}$, a direct calculation of the population Kendall's $\tau$ $\tau_{jk}=\mathbb{E}(\hat{\tau}_{jk})$ yields

$$\begin{aligned}\tau_{jk}&=2\,\mathbb{E}(X_{ij}X_{ik})-2\,\mathbb{E}(X_{ij})\mathbb{E}(X_{ik})\\&=2\,\mathbb{P}\{f_j(Z_{ij})>\Delta_j,f_k(Z_{ik})>\Delta_k\}-2\,\mathbb{P}\{f_j(Z_{ij})>\Delta_j\}\,\mathbb{P}\{f_k(Z_{ik})>\Delta_k\}\\&=2\{\Phi_2(\Delta_j,\Delta_k,\Sigma_{jk})-\Phi(\Delta_j)\,\Phi(\Delta_k)\}.\end{aligned} \tag{4.4}$$

Here, we denote $\Phi_2(u,v,t)=\int_{x_1<u}\int_{x_2<v}\phi_2(x_1,x_2;t)\,\mathrm{d}x_1\mathrm{d}x_2$ by the cumulative distribution function of the standard bivariate normal distribution, where $\phi_2(x_1,x_2;t)$ is the probability density function of the standard bivariate normal distribution with correlation $t$. Let $\Phi(\cdot)$ be the cumulative distribution of the standard normal distribution.

To emphasize the dependence of the bridge function on $\Delta_j$ and $\Delta_k$, we denote equation (4.4) by

$$F(t;\Delta_j,\Delta_k)=2\{\Phi_2(\Delta_j,\Delta_k,t)-\Phi(\Delta_j)\,\Phi(\Delta_k)\}. \tag{4.5}$$

As a special case, when $\Delta_j = \Delta_k = 0$, by Sheppard's theorem (Sheppard, 1899), $F(t;0,0)$ can be further simplified to $F(t;0,0) = (1/\pi)\sin^{-1}(t)$, i.e. $\tau_{jk} = (1/\pi)\sin^{-1}(\Sigma_{jk})$. As will be shown in lemma 2, we verify that $F(t;\Delta_j,\Delta_k)$ is invertible with respect to $t$, and we denote the inverse function by $F^{-1}(\tau;\Delta_j,\Delta_k)$. Thus, given $\Delta_j$ and $\Delta_k$, we can estimate $\Sigma_{jk}$ by $F^{-1}(\hat{\tau}_{jk};\Delta_j,\Delta_k)$. In practice, the cut-off values $\Delta_j$ and $\Delta_k$ can be estimated from the moment equation $\mathbb{E}(X_{ij}) = 1 - \Phi(\Delta_j)$. Namely, $\Delta_j$ can be estimated by $\hat{\Delta}_j = \Phi^{-1}(1 - \bar{X}_j)$, where $\bar{X}_j = \Sigma_{i=1}^n X_{ij}/n$. Thus, the rank-based estimator of $\Sigma_{jk}$, when both $X_{ij}$ and $X_{ik}$ are binary, is

$$\hat{R}_{jk} = F^{-1}(\hat{\tau}_{jk};\hat{\Delta}_j,\hat{\Delta}_k). \tag{4.6}$$

Given $\hat{\Delta}_j$ and $\hat{\Delta}_k$, the estimator $\hat{R}_{jk}$ is the root of the equation $F(t;\hat{\Delta}_j,\hat{\Delta}_k) = \hat{\tau}_{jk}$. As seen in lemma 2 below, the function $F(t;\hat{\Delta}_j,\hat{\Delta}_k)$ is strictly increasing, and therefore its root can be easily solved by using Newton's method.

In case II, when $X_{ij}$ is binary and $X_{ik}$ is continuous, the following lemma establishes the bridge function that connects the population Kendall's $\tau$ to $\mathbf{\Sigma}$ for mixed data.

*Lemma 1.* When $X_{ij}$ is binary and $X_{ik}$ is continuous, $\tau_{jk} = \mathbb{E}(\hat{\tau}_{jk})$ is given by $\tau_{jk} = F(\Sigma_{jk};\Delta_j)$, where

$$F(t;\Delta_j) = 4\,\Phi_2(\Delta_j, 0, t/\sqrt{2}) - 2\,\Phi(\Delta_j). \tag{4.7}$$

Moreover, for fixed $\Delta_j$, $F(t;\Delta_j)$ is an invertible function of $t$. In particular, when $\Delta_j = 0$, we have $F(t,0) = (2/\pi)\sin^{-1}(t/\sqrt{2})$, and hence $\Sigma_{jk} = \sqrt{2}\sin(\pi\tau_{jk}/2)$.

Similarly, the unknown parameter $\Delta_j$ can be estimated by $\hat{\Delta}_j = \Phi^{-1}(1 - \bar{X}_j)$, where $\bar{X}_j = \Sigma_{i=1}^n X_{ij}/n$. When $X_{ij}$ is binary and $X_{ik}$ is continuous, the rank-based estimator is defined as

$$\hat{R}_{jk} = F^{-1}(\hat{\tau}_{jk};\hat{\Delta}_j), \tag{4.8}$$

where $F^{-1}(\tau,\Delta_j)$ is the inverse function of $F(t,\Delta_j)$ for fixed $\Delta_j$.

Thus, combining these three cases, the rank-based estimator of $\mathbf{\Sigma}$ is given by $\hat{\mathbf{R}}$, where $\hat{\mathbf{R}}$ is a symmetric matrix with $\hat{R}_{jj} = 1$, $\hat{R}_{jk} = \sin(\pi\hat{\tau}_{jk}/2)$ for $d_1 + 1 \leqslant j < k \leqslant d$, $\hat{R}_{jk} = F^{-1}(\hat{\tau}_{jk};\hat{\Delta}_j,\hat{\Delta}_k)$ for $1 \leqslant j < k \leqslant d_1$ and $\hat{R}_{jk} = F^{-1}(\hat{\tau}_{jk};\hat{\Delta}_j)$ for $1 \leqslant j \leqslant d_1$, $d_1 + 1 \leqslant k \leqslant d$.

### 4.2. Theoretical results

In this section, we establish concentration results for the rank-based estimator, which plays the key role in the theory of graph estimation and model selection. We first consider case I, where both $X_{ij}$ and $X_{ik}$ are binary. The following lemma justifies that the inverse function of $F(t;\Delta_j,\Delta_k)$ exists, such that the rank-based estimator $\hat{R}_{jk}$ in equation (4.6) is well defined.

*Lemma 2.* For any fixed $\Delta_j$ and $\Delta_k$, $F(t;\Delta_j,\Delta_k)$ in equation (4.5) is a strictly increasing function on $t \in (-1, 1)$. Thus, the inverse function $F^{-1}(\tau;\Delta_j,\Delta_k)$ exists.

To study the theoretical properties of $\hat{\mathbf{R}}$, we assume the following regularity conditions.

*Assumption 1.* There is a constant $\delta > 0$ such that $|\Sigma_{jk}| \leqslant 1 - \delta$, for any $1 \leqslant j < k \leqslant d_1$.

*Assumption 2.* There is a constant $M$ such that $|\Delta_j| \leqslant M$, for any $j = 1, \ldots, d_1$.

Conditions 1 and 2 are adopted for technical reasons and they impose little restriction in practice. Specifically, condition 1 rules out the singular case that $f_j(Z_{ij})$ and $f_k(Z_{ik})$ are perfectly collinear. Condition 2 is used to control the variation of $F^{-1}(\tau;\Delta_j,\Delta_k)$ with respect to $(\tau;\Delta_j,\Delta_k)$. The following theorem establishes the convergence rate of $\hat{R}_{jk} - \Sigma_{jk}$ uniformly over $1 \leqslant j, k \leqslant d_1$.

*Theorem 1.* Under assumptions 1 and 2, for any $t > 0$ we have

$$\mathbb{P}\Big( \sup_{1 \leqslant j,k \leqslant d_1} |\hat{R}_{jk} - \Sigma_{jk}| > t \Big) \leqslant 2d_1^2 \exp\Big( -\frac{nt^2}{8L_2^2} \Big) + 4d_1^2 \exp\Big( -\frac{nt^2\pi}{16^2 L_1^2 L_2^2} \Big) + 4d_1^2 \exp\Big( -\frac{M^2 n}{2L_1^2} \Big), \qquad (4.9)$$

where $L_1$ and $L_2$ are positive constants given in lemmas A.2 and A.1 in the on-line supplementary materials, respectively, i.e., for some constant $C$ independent of $(n, d)$, $\sup_{1 \leqslant j,k \leqslant d_1} |\hat{R}_{jk} - \Sigma_{jk}| \leqslant C\surd\{\log(d)/n\}$ with probability greater than $1 - d^{-1}$.

Now we consider case II, where $X_{ij}$ is binary and $X_{ik}$ is continuous. The following concentration result similar to theorem 1 holds.

*Theorem 2.* Under assumptions 1 and 2, for any $t > 0$ we have

$$\mathbb{P}\Big( \sup_{1 \leqslant j \leqslant d_1, d_1+1 \leqslant k \leqslant d} |\hat{R}_{jk} - \Sigma_{jk}| > t \Big) \leqslant 2d_1 d_2 \exp\Big( -\frac{nt^2}{8L_3^2} \Big) + 2d_1 d_2 \exp\Big( -\frac{nt^2\pi}{12^2 L_1^2 L_3^2} \Big) + 2d_1 d_2 \exp\Big( -\frac{M^2 n}{2L_1^2} \Big),$$

where $L_1$ and $L_3$ are positive constants given in lemmas A.2 and A.3 in the on-line supplementary material respectively. i.e., for some constant $C$ independent of $(n, d)$,

$$\sup_{1 \leqslant j \leqslant d_1,\, d_1+1 \leqslant k \leqslant d} |\hat{R}_{jk} - \Sigma_{jk}| \leqslant C\surd\{\log(d)/n\}$$

with probability greater than $1 - d^{-1}$.

Analogously to theorems 1 and 2, for continuous components, the following lemma in Liu *et al.* (2012) provides the upper bound for $\sup_{d_1+1 \leqslant j,\, k \leqslant d} |\hat{R}_{jk} - \Sigma_{jk}|$.

*Lemma 3.* For $n > 1$, with probability greater than $1 - d_2^{-1}$, we have

$$\sup_{d_1+1 \leqslant j,k \leqslant d} |\hat{R}_{jk} - \Sigma_{jk}| \leqslant 2.45\pi \sqrt{\Big\{ \frac{\log(d_2)}{n} \Big\}}.$$

Combining theorems 1 and 2 and lemma 3, we finally obtain the concentration inequality for $|\hat{R}_{jk} - \Sigma_{jk}|$ uniformly over $1 \leqslant j, k \leqslant d$.

*Corollary 1.* Under assumptions 1 and 2, with probability greater than $1 - d^{-1}$, we have

$$\sup_{1 \leqslant j,k \leqslant d} |\hat{R}_{jk} - \Sigma_{jk}| \leqslant C \sqrt{\Big\{ \frac{\log(d)}{n} \Big\}},$$

where $C$ is a constant independent of $(n, d)$.

## 5. Latent graph structure learning for mixed data

The structure of the latent graph is characterized by the sparsity pattern of the inverse correlation matrix $\mathbf{\Omega}$. In this section, we show that a simple modification of the existing estimators for the Gaussian graphical model can be used to estimate $\mathbf{\Omega}$. For concreteness, we demonstrate the modification for the graphical lasso estimator (Friedman *et al.*, 2008), CLIME (Cai *et al.*, 2011) and adaptive graphical lasso estimator (Fan *et al.*, 2009, 2014), which are given as follows: for the graphical lasso,

$$\hat{\boldsymbol{\Omega}} = \underset{\boldsymbol{\Omega}\succeq 0}{\arg\min}\{\mathrm{tr}(\hat{\mathbf{R}}\boldsymbol{\Omega}) - \log|\boldsymbol{\Omega}| + \lambda \sum_{j\neq k} |\Omega_{jk}|\}; \tag{5.1}$$

for the adaptive graphical lasso,

$$\hat{\boldsymbol{\Omega}} = \underset{\boldsymbol{\Omega}\succeq 0}{\arg\min}\{\mathrm{tr}(\hat{\mathbf{R}}\boldsymbol{\Omega}) - \log|\boldsymbol{\Omega}| + \sum_{j\neq k} p_\lambda(|\Omega_{jk}|)\}; \tag{5.2}$$

for CLIME,

$$\hat{\boldsymbol{\Omega}} = \arg\min \|\boldsymbol{\Omega}\|_1, \qquad \text{subject to } \|\hat{\mathbf{R}}\boldsymbol{\Omega} - \mathbf{I}_d\|_{\max} \leqslant \lambda, \tag{5.3}$$

where $\mathbf{I}_d$ is a $d \times d$ identity matrix, $\lambda$ is a tuning parameter and $p_\lambda(\theta)$ is a folded concave penalty function such as the smoothly clipped absolute deviation penalty (Fan and Li, 2001) and minimax concave penalty (Zhang, 2010). Compared with the original formulation of the graphical lasso, CLIME and adaptive graphical lasso estimators, the modification that we conduct is that the sample covariance matrix is now replaced by the rank-based estimator $\hat{\mathbf{R}}$. The same modification can be also applied to other existing Gaussian graphical model estimators with the sample covariance matrix as the input.

However, one potential issue with the rank-based estimator is that $\hat{\mathbf{R}}$ may not be positive semidefinite. Since we do not penalize the diagonal elements of $\boldsymbol{\Omega}$ in equations (5.1) and (5.2), the resulting estimator may diverge to $\infty$. Even though optimization problem (5.1) remains convex, the computational algorithms in Friedman *et al.* (2008) and Hsieh *et al.* (2011), among others, may not converge. To regularize the estimator further, we can project $\hat{\mathbf{R}}$ into the cone of positive semidefinite matrices, i.e.

$$\hat{\mathbf{R}}_p = \underset{\mathbf{R}\succeq 0}{\arg\min}\, \|\hat{\mathbf{R}} - \mathbf{R}\|_{\max}. \tag{5.4}$$

The smoothed approximation method in Nesterov (2005) can be used to calculate $\hat{\mathbf{R}}_p$; see also Liu *et al.* (2012) and Zhao *et al.* (2014) for some computationally efficient algorithms. Hence, we can replace $\hat{\mathbf{R}}$ in problems (5.1) and (5.2) by $\hat{\mathbf{R}}_p$. The following corollary shows that a similar error bound holds for the projected estimator $\hat{\mathbf{R}}_p$ in equation (5.4).

*Corollary 2.* Under assumptions 1 and 2, with probability greater than $1 - d^{-1}$, we have

$$\|\hat{\mathbf{R}}_p - \boldsymbol{\Sigma}\|_{\max} \leqslant C\sqrt{\left\{\frac{\log(d)}{n}\right\}},$$

where $C$ is a constant that is independent of $(n, d)$.

By corollaries 1 and 2, under assumptions 1 and 2, the graphical lasso (5.1), adaptive graphical lasso (5.2) and CLIME (5.3) with $\hat{\mathbf{R}}$ or $\hat{\mathbf{R}}_p$ enjoy the same theoretical properties as those established by Raskutti *et al.* (2008), Fan *et al.* (2014) and Cai *et al.* (2011) respectively. Thus, our estimator achieves the same rate of convergence for estimating $\boldsymbol{\Omega}$ and model selection consistency, as if the latent variables $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ were observed. We refer the reader to the corresponding references for the detailed theoretical results.

The optimization problem (5.2) is non-convex because of the penalty function. In practice, we suggest use of the local linear approximation algorithm that was proposed by Zou and Li (2008) and Fan *et al.* (2014) to solve problem (5.2). In our context, we can solve the weighted $l_1$-penalization problem

$$\hat{\mathbf{\Omega}}_A = \mathop{\arg\min}_{\mathbf{\Omega}\succeq 0}\{\mathrm{tr}(\hat{\mathbf{R}}_p\mathbf{\Omega}) - \log|\mathbf{\Omega}| + \sum_{j\neq k} p'_\lambda(|\hat{\Omega}^0_{jk}|)|\Omega_{jk}|\}, \tag{5.5}$$

where $p'_\lambda(\theta)$ is the derivative of $p_\lambda(\theta)$ with respect to $\theta$ and $\hat{\mathbf{\Omega}}^0 = (\hat{\Omega}^0_{jk})$ is an initial estimator of $\mathbf{\Omega}$ which can be taken as the graphical lasso or CLIME estimator.

To select the tuning parameter $\lambda$ in the graphical lasso (5.1), adaptive graphical lasso (5.2) and CLIME (5.3), we suggest use of the high dimensional Bayesian information criterion that was proposed by Wang *et al.* (2013) and Fan and Tang (2013). In particular, we use $\hat{\mathbf{\Omega}}_\lambda$ to denote the estimators (5.1), (5.2) and (5.3) corresponding to the tuning parameter $\lambda$. The high dimensional Bayesian information criterion is defined as

$$\mathrm{HBIC}(\lambda) = \mathrm{tr}(\hat{\mathbf{R}}\hat{\mathbf{\Omega}}_\lambda) - \log|\hat{\mathbf{\Omega}}_\lambda| + C_n\frac{\log(d)}{n}s_\lambda,$$

where, as suggested by Wang *et al.* (2013) and Fan and Tang (2013), we take $C_n = \log\{\log(n)\}$ and $s_\lambda$ is the number of edges corresponding to $\hat{\mathbf{\Omega}}_\lambda$. Then, the tuning parameter is chosen by $\lambda_{\mathrm{HBIC}} = \arg\min_{\lambda\in\Lambda}\mathrm{HBIC}(\lambda)$, where $\Lambda$ is a sequence of values for $\lambda$. This procedure is further empirically assessed by simulation studies.

## 6. Simulation studies

### 6.1. Data generation

To evaluate the accuracy of graphical estimation, we adopt similar data-generating procedures to that in Liu *et al.* (2012). To generate the inverse correlation matrix $\mathbf{\Omega}$, we set $\Omega_{jj} = 1$, and $\Omega_{jk} = ta_{jk}$, if $j \neq k$. Here, $t$ is a constant which is chosen to guarantee the positive definiteness of $\mathbf{\Omega}$, and $a_{jk}$ is a Bernoulli random variable with a success probability $p_{jk} = (2\pi)^{-1/2}\exp\{\|z_j - z_k\|_2/(2c_1)\}$, where $z_j = (z_j^{(1)}, z_j^{(2)})$ is independently generated from a bivariate uniform $[0,1]$ distribution, and $c_1$ is chosen such that there are about 200 edges in the graph. We choose $t = 0.15$. In the simulation studies, we consider three possible values for the dimensionality of the graph: $d = 50, 250, 3000$, which represent small, moderate and large-scale graphs. Since $\mathbf{\Sigma}$ needs to be a correlation matrix, we rescale the covariance matrix such that the diagonal elements of $\mathbf{\Sigma}$ are 1.

Assume the cut-off $C \sim \mathrm{Unif}[-1, 1]$. Consider the following four data-generating scenarios.

(a) Simulate data $\mathbf{X} = (X_1, \ldots, X_d)$, where $X_j = I(Z_j > C_j)$, for all $j = 1, \ldots, d$, and $\mathbf{Z} \sim N(0, \mathbf{\Sigma})$.

(b) Simulate data $\mathbf{X} = (X_1, \ldots, X_d)$, where $X_j = I(Z_j > C_j)$, for all $j = 1, \ldots, d$, and $\mathbf{Z} \sim N(0, \mathbf{\Sigma})$, where 10% entries in each $\mathbf{Z}$ are randomly sampled and replaced by $-5$ or 5.

(c) Simulate data $\mathbf{X} = (X_1, \ldots, X_d)$, where $X_j = I(Z_j > C_j)$, for $j = 1, \ldots, d/2$, $\mathbf{Z} \sim N(0, \mathbf{\Sigma})$ and $X_j = Z_j$, for $j = d/2+1, \ldots, d$.

(d) Simulate data $\mathbf{X} = (X_1, \ldots, X_d)$, where $X_j = I(Z_j > C_j)$, for $j = 1, \ldots, d/2$, $\mathbf{Z} \sim \mathrm{NPN}(0, \mathbf{\Sigma}, f)$ and $X_j = Z_j$, for $j = 1, \ldots, d/2$, where $f_j(x) = x^3$ for $j = d/2+1, \ldots, d$.

In scenarios (a) and (b), the binary data are generated. In particular, scenario (a) corresponds to the latent Gaussian model and scenario (b) represents the setting where the binary data can be misclassified because of the outliers of the latent variable. Scenarios (c) and (d) correspond to the mixed data generated from the latent Gaussian model and the latent Gaussian copula model respectively. The sample size is $n = 200$ when $d = 50$ and $d = 250$. For the large-scale graph with $d = 3000$, we use $n = 600$. We repeat the simulation 100 times.

### 6.2. Estimation error

In this section, we examine the empirical estimation error for the precision matrix. Here, we compare five estimation methods:

(a) the latent graphical lasso estimator L-GLASSO in problem (5.1),
(b) the latent adaptive graphical lasso estimator L-GSCAD in problem (5.2),
(c) the approximate sparse maximum likelihood estimator AMLE in Banerjee *et al.* (2008),
(d) ZR-GLASSO (where 'ZR' denotes rank-based correlation of random variable **Z**) and
(e) ZP-GLASSO (where 'ZP' denotes the Pearson correlation of **Z**).

The weight in L-GSCAD is based on the smoothly clipped absolute deviation penalty with $a=3.7$ and the estimator is calculated by the local linear approximation algorithm (Zou and Li, 2008; Fan *et al.*, 2014). AMLE refers to the graphical lasso estimator with the modified sample covariance matrix $\tilde{\mathbf{\Sigma}}$, where

$$\tilde{\mathbf{\Sigma}}=\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i-\bar{\mathbf{X}})(\mathbf{X}_i-\bar{\mathbf{X}})^{\mathrm{T}}+\frac{1}{3}, \qquad \bar{\mathbf{X}}=\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i,$$

as the input. In ZR-GLASSO and ZP-GLASSO, we assume the latent variable **Z** is observed. In particular, the rank-based covariance matrix of **Z** (Liu *et al.*, 2012) and the sample covariance matrix of **Z** are plugged into the graphical lasso procedure. Since **Z** represents the latent variable, ZR-GLASSO and ZP-GLASSO are often unavailable in real applications. Here, we use these two estimators as benchmarks to quantify the loss of information of our proposed estimators constructed on the basis of the observed data **X**. We find that the CLIME estimator (5.3) has similar performance to the L-GLASSO estimator. Hence, we present only the results for L-GLASSO. We also examine the performance of a naive estimator corresponding to the graphical lasso estimator with the sample covariance matrix of **X** as the input. This estimator has similar performance to AMLE. For simplicity, we report only the latter.

We note that the competing methods for mixed data (Lee and Hastie, 2014; Fellinghauer *et al.*, 2013; Cheng *et al.*, 2013; Yang *et al.*, 2014a; Chen *et al.*, 2015) do not consider the problem of precision matrix estimation and therefore they are not suitable for comparison from the precision matrix estimation perspective. Later, we shall compare their performance in terms of graph structure recovery in Section 6.4.

Table 1 reports the mean estimation error of $\hat{\mathbf{\Omega}}-\mathbf{\Omega}$ in terms of the Frobenius and the matrix $L_1$-norms. The entries for L-GLASSO and L-GSCAD are calculated under the tuning parameter chosen by the HBIC-method. For the remaining procedures, similar HBIC-methods are used to determine $\lambda$. It is seen that L-GLASSO has smaller estimation errors than AMLE under all scenarios. This becomes more transparent, as the dimension grows. In addition, the folded concave estimator L-GSCAD further reduces the estimation error of L-GLASSO, which is consistent with the literature. Compared with the estimation errors of the benchmarks ZR-GLASSO and ZP-GLASSO, Table 1 suggests that the proposed estimators L-GLASSO and L-GSCAD suffer little loss of information for $d=50, 250$ and only moderate loss of information for the very high dimensional setting with $d=3000$. Additional simulation results in the on-line supplementary material show that the conclusions are stable with respect to the signal strength of the true precision matrix.

### 6.3. Graph recovery

Define the number of false positive FP($\lambda$) and true positive results TP($\lambda$) with regularization parameter $\lambda$ as the number of lower off-diagonal elements $(i, j)$ such that $\Omega_{ij}=0$ but $\hat{\Omega}_{ij}\neq 0$, and the number of lower off-diagonal elements $(i, j)$ such that $\Omega_{ij}\neq 0$ and $\hat{\Omega}_{ij}\neq 0$. Define the false positive rate FPR($\lambda$) and true positive rate TPR($\lambda$) as

**Table 1.** Average estimation error of L-GLASSO, L-GSCAD, AMLE, ZR-GLASSO and ZP-GLASSO for $\hat{\Omega}-\Omega$ as measured by the matrix $L_1$-norm $L$ and the Frobenius norm $F$†

| *d* | *Scenario* | *Norm* | *Results for the following estimators:* | | | | |
|---|---|---|---|---|---|---|---|
| | | | *L-GLASSO* | *L-GSCAD* | *AMLE* | *ZR-GLASSO* | *ZP-GLASSO* |
| 50 | (a) | *F* | 3.74 (0.29) | 3.67 (0.25) | 4.52 (0.18) | 3.01 (0.18) | 2.88 (0.20) |
| | | *L* | 3.00 (0.49) | 2.94 (0.39) | 3.14 (0.42) | 2.87 (0.33) | 2.77 (0.35) |
| | (b) | *F* | 3.92 (0.24) | 3.77 (0.30) | 4.52 (0.20) | 3.59 (0.29) | 4.27 (0.41) |
| | | *L* | 3.11 (0.49) | 3.00 (0.45) | 3.32 (0.46) | 3.40 (0.46) | 3.59 (0.58) |
| | (c) | *F* | 3.66 (0.28) | 3.40 (0.20) | 4.50 (0.26) | 3.01 (0.18) | 2.88 (0.20) |
| | | *L* | 3.10 (0.52) | 3.02 (0.45) | 3.39 (0.55) | 2.87 (0.33) | 2.77 (0.35) |
| | (d) | *F* | 3.80 (0.35) | 3.56 (0.39) | 6.27 (0.77) | 3.04 (0.26) | 3.70 (0.24) |
| | | *L* | 3.23 (0.52) | 3.08 (0.46) | 4.54 (0.56) | 2.93 (0.36) | 3.10 (0.35) |
| 250 | (a) | *F* | 6.50 (0.31) | 6.12 (0.25) | 9.42 (0.10) | 5.50 (0.20) | 5.41 (0.23) |
| | | *L* | 3.55 (0.35) | 3.50 (0.29) | 3.70 (0.44) | 3.38 (0.27) | 3.37 (0.25) |
| | (b) | *F* | 6.56 (0.24) | 6.50 (0.30) | 9.40 (0.12) | 5.72 (0.22) | 6.70 (0.65) |
| | | *L* | 3.68 (0.30) | 3.66 (0.26) | 3.73 (0.26) | 3.49 (0.32) | 3.79 (0.33) |
| | (c) | *F* | 6.70 (0.38) | 6.43 (0.30) | 7.10 (0.26) | 5.50 (0.20) | 5.41 (0.23) |
| | | *L* | 3.66 (0.32) | 3.52 (0.35) | 4.64 (0.35) | 3.38 (0.27) | 3.37 (0.25) |
| | (d) | *F* | 6.99 (0.37) | 6.63 (0.34) | 9.34 (0.29) | 5.30 (0.17) | 5.55 (0.30) |
| | | *L* | 3.72 (0.40) | 3.57 (0.37) | 4.19 (0.33) | 3.40 (0.29) | 3.59 (0.30) |
| 3000 | (a) | *F* | 12.5 (1.43) | 10.8 (1.39) | 18.8 (2.45) | 7.97 (0.76) | 7.53 (0.77) |
| | | *L* | 2.52 (0.36) | 2.50 (0.34) | 3.42 (0.54) | 1.16 (0.20) | 1.22 (0.25) |
| | (b) | *F* | 12.7 (1.50) | 10.8 (1.39) | 18.9 (2.45) | 7.64 (0.70) | 9.62 (0.90) |
| | | *L* | 2.83 (0.47) | 2.77 (0.40) | 3.38 (0.60) | 1.11 (0.24) | 1.56 (0.43) |
| | (c) | *F* | 13.5 (1.78) | 11.0 (1.67) | 18.4 (1.88) | 7.97 (0.76) | 7.53 (0.77) |
| | | *L* | 3.35 (0.58) | 3.30 (0.50) | 3.96 (0.59) | 1.16 (0.20) | 1.22 (0.25) |
| | (d) | *F* | 13.0 (1.73) | 11.4 (1.58) | 19.9 (2.10) | 8.13 (0.85) | 8.33 (0.87) |
| | | *L* | 3.39 (0.55) | 3.30 (0.52) | 4.21 (0.66) | 1.09 (0.26) | 1.20 (0.31) |

†Numbers in parentheses are the simulation standard errors.

$$\mathrm{FPR}(\lambda)=\frac{\mathrm{FP}(\lambda)}{d(d-1)/2-|E|},$$
$$\mathrm{TPR}(\lambda)=\frac{\mathrm{TP}(\lambda)}{|E|},$$

where $|E|$ is the number of edges in the graph. Fig. 1 shows the plot of $\mathrm{TPR}(\lambda)$ against $\mathrm{FPR}(\lambda)$ for L-GLASSO, L-GSCAD, AMLE, ZR-GLASSO and ZP-GLASSO, when $d=50$. We find that L-GLASSO always yields higher TPR than AMLE for any fixed FPR under all four scenarios, and L-GSCAD improves L-GLASSO in terms of graph recovery. By comparing the receiver operating characteristic curves in scenarios (a) and (b), L-GLASSO and L-GSCAD are more robust to data misclassification than the benchmark estimators ZR-GLASSO and ZP-GLASSO. This robustness property demonstrates the advantage of the dichotomization method. In the absence of misclassification, it is seen that the receiver operating characteristic curves of L-GLASSO and ZR-GLASSO are similar, suggesting little loss of information for the graph recovery due to the dichotomization procedure.

### 6.4. Further comparison with competing approaches

In this section, we further compare the proposed estimator L-GLASSO with competing approaches for mixed graphical models. The following four estimators are considered in this study:

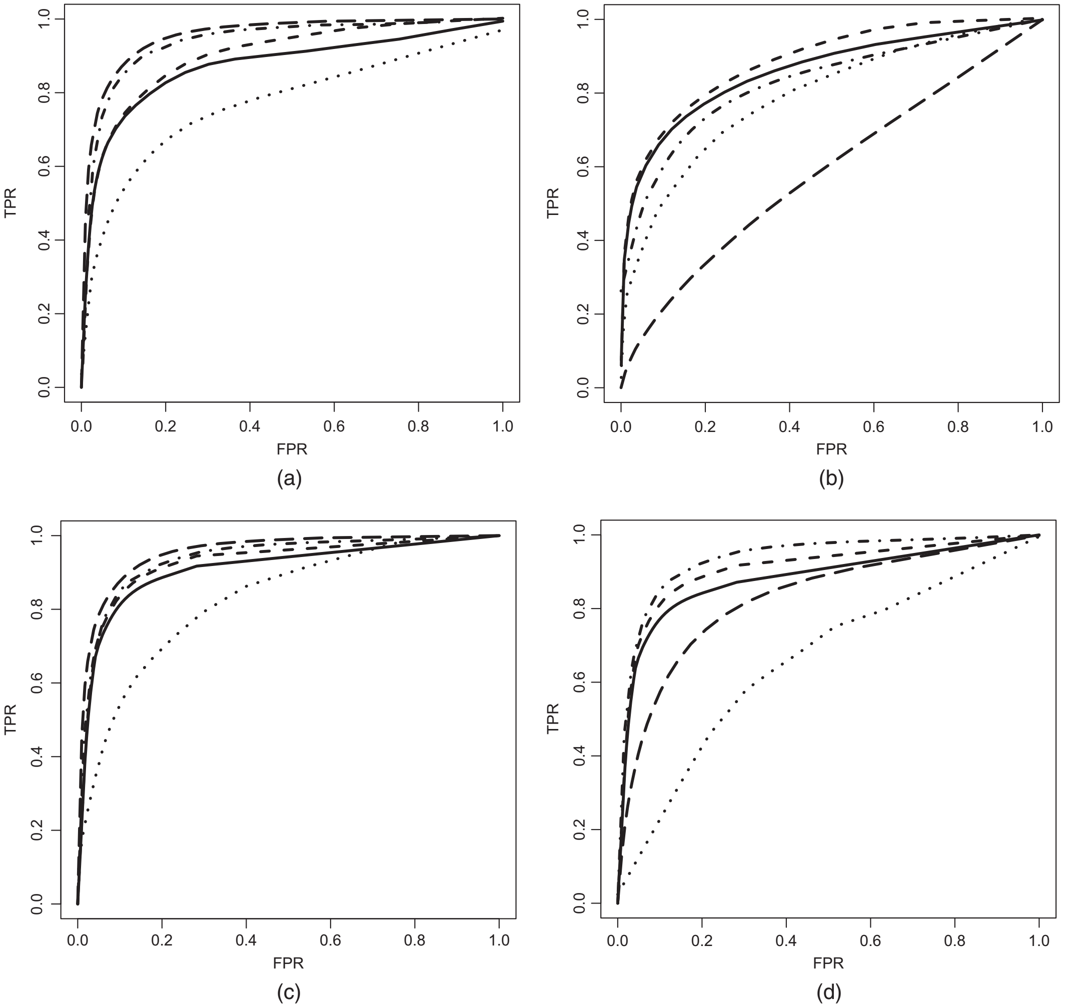

**Fig. 1.** TPR *versus* FPR for graph recovery of L-GLASSO (———), L-GSCAD (– – –), AMLE (· · · · · ·), ZR-GLASSO (·–·–·) and ZP-GLASSO (— —), when $d = 50$: (a) scenario (a); (b) scenario (b); (c) scenario (c); (d) scenario (d)

Nodewise-1, PMLE, Nodewise-2 and Forest. Specifically, the Nodewise-1 estimator refers to penalized nodewise regression based on the pairwise exponential family (Chen *et al.*, 2015; Yang *et al.*, 2014a); the PMLE estimator refers to the penalized pseudolikelihood estimator in the mixed graphical model (Lee and Hastie, 2014); the Nodewise-2 estimator refers to weighted $L_1$-penalized nodewise regression (Cheng *et al.*, 2013); and finally the Forest estimator refers to the random-forests estimator for mixed graphical models (Fellinghauer *et al.*, 2013).

We adopt the same data-generating procedures. Fig. 2 displays the plot of TPR against FPR for graph recovery of L-GLASSO, Nodewise-1, PMLE, Nodewise-2 and Forest. In all four scenarios, L-GLASSO outperforms the existing estimators in terms of graph recovery. The estimators Nodewise-1 (Chen *et al.*, 2015; Yang *et al.*, 2014a) and PMLE (Lee and Hastie, 2014) have similar performance and both have lower TPR than the method proposed. This is because both of them are derived on the basis of the exponential family graphical model which

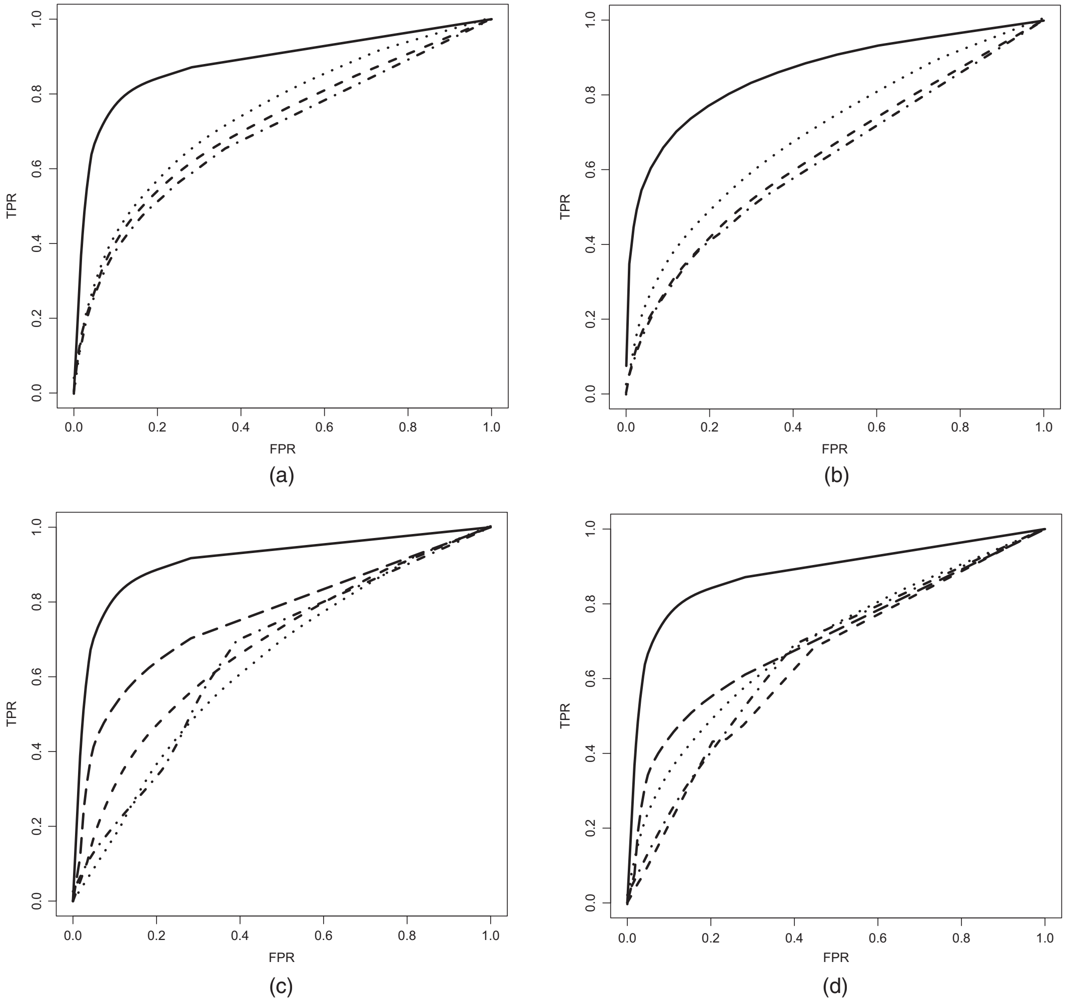

**Fig. 2.** TPR *versus* FPR for graph recovery of L-GLASSO (———), Nodewise-1 (— — —), PMLE (· · · · · ·), Nodewise-2 (— —) and Forest (· — · — ·) when $d=50$: (a) scenario (a); (b) scenario (b); (c) scenario (c); (d) scenario (d)

is different from the data-generating model. The Nodewise-2 estimator in Cheng *et al.* (2013) is identical to Nodewise-1 for the binary data in scenarios (a) and (b) and attempts to incorporate more sophisticated interaction than PMLE for mixed data. It shows improved performance in scenarios (c) and (d). Finally, as a non-parametric estimator, the Forest estimator (Fellinghauer *et al.*, 2013), tends to be less efficient than the parametric and semiparametric approaches. This explains the fact that our estimator L-GLASSO has higher TPR than does the Forest estimator. Further comparison of these estimators for $d=250$ demonstrates the same patterns; see the on-line supplementary material for details.

## 7. Analysis of Arabidopsis data

In this section, we consider the graph estimation for the Arabidopsis data set that was analysed by Lange and Ghassemian (2003), Wille *et al.* (2004) and Ma *et al.* (2007). As an illustration,

**Table 2.** Number of different edges among L-GLASSO *versus* AMLE, L-GLASSO *versus* Nodewise-1, L-GLASSO *versus* PMLE and L-GLASSO *versus* Forest in the Arabidopsis data

| *Total edges* | *Number of edges for the following methods:* | | | |
|---|---|---|---|---|
| | *AMLE* | *Nodewise-1* | *PMLE* | *Forest* |
| 80 | 27 | 29 | 30 | 34 |
| 60 | 19 | 20 | 24 | 20 |
| 45 | 15 | 17 | 14 | 12 |
| 25 | 7 | 9 | 10 | 6 |
| 10 | 6 | 5 | 6 | 3 |

we focus on 39 genes which are possibly related to the mevalonate or non-mevalonate pathway. In addition, 118 GeneChip (Affymetrix) microarrays are used to measure the gene expression values under various experimental conditions.

To remove the batch effects due to different experiments, we apply the adaptive dichotomization method that is implemented by the `ArrayBin` package in R (`https://cran.r-project.org/web/packages/ArrayBin/index.html`). This method transforms the numerical expression data into 0–1 binary data, where genes with higher expression values are encoded as 1 and genes with lower expression values are encoded as 0. Although the loss of information is inevitable in the discretization procedure, McCall and Irizarry (2011) argued that this procedure can potentially improve the accuracy of the statistical analysis. In contrast to Wille *et al.* (2004) and Ma *et al.* (2007) who imposed the Gaussian model assumption on the numerical expression values, we work on the derived binary data with the purpose of removing batch effects.

We compare the performance of our proposed L-GLASSO with several estimators, i.e. AMLE (Banerjee *et al.*, 2008), Nodewise-1 (Chen *et al.*, 2015), PMLE (Lee and Hastie, 2014) and Forest (Fellinghauer *et al.*, 2013). Note that the Nodewise-2 estimator in Cheng *et al.* (2013) is identical to Nodewise-1 in Chen *et al.* (2015) for binary data. The tuning parameters are selected separately, such that the estimated graphs have the same number of edges. The number of different edges for L-GLASSO *versus* AMLE, L-GLASSO *versus* Nodewise-1, L-GLASSO *versus* PMLE and L-GLASSO *versus* Forest is presented in Table 2. We find that our estimator produces 30–60% different edges compared with the existing methods, depending on the level of sparsity of the estimated graphs. When the number of estimated edges is small (i.e. 10 edges), the graph that is estimated by L-GLASSO is more concordant with that estimated by the non-parametric Forest estimator.

From a biological perspective, some well-known association patterns are identified by all the methods. For instance, when the number of total edges is 10, all four methods identify the gene–gene interaction between AACT2 and MK, and the interaction between AACT2 and FPPS2. These results are consistent with the findings in Wille *et al.* (2004). More importantly, many interesting association patterns are identified by L-GLASSO rather than by the existing methods. For instance, L-GLASSO is the only method that concludes that genes CMK and MCT, and CMK and MECPS are dependent. These genes are on the non-mevalonate pathway and are known to be associated in the literature (Hsieh and Goodman, 2005; Phillips *et al.*, 2008; Ruiz-Sola and Rodríguez-Concepción, 2012). Similarly, the association between genes MECPS and HDS supported by Phillips *et al.* (2008) is recovered by our estimator L-GLASSO and the non-parametric Forest estimator. Hence, we conclude that our method identifies some interesting dependence structure that is missed by the existing methods.

## 8. Discussion

In this paper, we propose a latent Gaussian copula model for mixed data. We assume that there is a deeper layer of unobserved driving factors that govern the observed mixed data. Thus, our primary interest is to learn the dependence structure of the latent variables. It is important to note that the conditional independence between latent variables (i.e. $Z_j$ and $Z_k$ are independent given $\mathbf{Z}_{-(j,k)}$) does not imply the conditional independence between observed binary variables (i.e. $X_j$ and $X_k$ are independent given $\mathbf{X}_{-(j,k)}$, where $X_j = I(Z_j > C_j)$).

Recently, Chandrasekaran *et al.* (2012) studied the latent variable graphical model. This model assumes that a subset of random variables is not observed. These variables are called latent variables or missing variables. This model is useful to account for unobserved confounding variables. In the current paper, we introduce latent variables to model the observed binary data. Hence, these two models are fundamentally different.

Although we focus on binary data in this paper, in principle, our methods can be extended to ordinal data with more than two categories. Specifically, once Kendall's $\tau$ has been defined, we can apply the proposed framework to derive the bridge function that connects the latent correlation matrix to the population of Kendall's $\tau$. However, unlike the binary case, the bridge function for ordinal data may not have a simple form and needs to be calculated case by case. One potentially unified approach to study ordinal data is to collapse the data into two categories. It is of interest to study the statistical properties of this procedure and to quantify the loss of information due to data collapse. We leave this problem for future investigations.

## 9. Supplementary materials

The supplementary material contains the proofs of the theoretical results, additional simulation studies and an analysis of a music data set.

## Acknowledgements

We thank the Joint Editor, the Associate Editor and the three referees for their helpful comments, which significantly improved the paper. Fan's research was supported by National Institutes of Health grant 2R01-GM072611-9 and National Science Foundation grants DMS-1206464 and DMS-1406266. Liu's research was supported by National Science Foundation grants III-1116730 and III-1332109, National Institutes of Health grants R01MH102339, R01GM083084 and R01HG06841, and Food and Drug Administration grant HHSF223201000072C. Zou's research was supported in part by National Science Foundation grant DMS-0846068.

## Reference

d'Aspremont, A., Banerjee, O. and El Ghaoui, L. (2008) First-order methods for sparse covariance selection. *SIAM J. Matr. Anal. Appl.*, **30**, 56–66.

Banerjee, O., El Ghaoui, L. and d'Aspremont, A. (2008) Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.*, **9**, 485–516.

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

Chandrasekaran, V., Parrilo, P. A. and Willsky, A. S. (2012) Latent variable graphical model selection via convex optimization. *Ann. Statist.*, **40**, 1935–1967.

Chen, S., Witten, D. M. and Shojaie, A. (2015) Selection and estimation for mixed graphical models. *Biometrika*, **102**, 47–64.

Cheng, J., Levina, E. and Zhu, J. (2013) High-dimensional mixed graphical models. *Preprint arXiv:1304.2810*. University of Michigan, Ann Arbor.

Fan, J., Feng, Y. and Wu, Y. (2009) Network exploration via the adaptive lasso and scad penalties. *Ann. Appl. Statist.*, **3**, 521–541.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.*, **96**, 1348–1360.

Fan, Y. and Tang, C. Y. (2013) Tuning parameter selection in high dimensional penalized likelihood. *J. R. Statist. Soc.* B, **75**, 531–552.

Fan, J., Xue, L. and Zou, H. (2014) Strong oracle optimality of folded concave penalized estimation. *Ann. Statist.*, **42**, 819–849.

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. *Computnl Statist. Data Anal.*, **64**, 132–152.

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

Friedman, J., Hastie, T. and Tibshirani, R. (2010) Applications of the lasso and grouped lasso to the estimation of sparse graphical models. *Technical Report*. Stanford University, Stanford.

Han, F. and Pan, W. (2012) A composite likelihood approach to latent multivariate Gaussian modeling of snp data with application to genetic association testing. *Biometrics*, **68**, 307–315.

Höfling, H. and Tibshirani, R. (2009) Estimation of sparse binary pairwise markov networks using pseudo-likelihoods. *J. Mach. Learn. Res.*, **10**, 883–906.

Hsieh, C.-J., Dhillon, I. S., Ravikumar, P. K. and Sustik, M. A. (2011) Sparse inverse covariance matrix estimation using quadratic approximation. In *Advances in Neural Information Processing Systems*. Red Hook: Curran Associates.

Hsieh, M.-H. and Goodman, H. M. (2005) The arabidopsis isph homolog is involved in the plastid nonmevalonate pathway of isoprenoid biosynthesis. *Plnt Physiol.*, **138**, 641–653.

Hsieh, C.-J., Sustik, M. A., Dhillon, I. S., Ravikumar, P. K. and Poldrack, R. (2013) Big & quic: sparse inverse covariance estimation for a million variables. In *Advances in Neural Information Processing Systems*. Red Hook: Curran Associates.

Kendall, M. G. (1948) *Rank Correlation Methods*. London: Griffin.

Lam, C. and Fan, J. (2009) Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.*, **37**, 42–54.

Lange, B. M. and Ghassemian, M. (2003) Genome organization in arabidopsis thaliana: a survey for genes involved in isoprenoid and chlorophyll metabolism. *Plnt Molec. Biol.*, **51**, 925–948.

Lauritzen, S. L. (1996) *Graphical Models*. Oxford: Clarendon.

Lazar, C., Meganck, S., Taminau, J., Steenhoff, D., Coletta, A., Molter, C., Weiss-Solís, D. Y., Duque, R., Bersini, H. and Nowé, A. (2013) Batch effect removal methods for microarray gene expression data integration: a survey. *Brief. Bioinform.*, **14**, 469–490.

Lee, J. D. and Hastie, T. J. (2014) Learning the structure of mixed graphical models. *J. Computnl Graph. Statist.*, **24**, 230–253.

Liu, H., Han, F., Yuan, M., Lafferty, J. D. and Wasserman, L. A. (2012) High dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.*, **40**, 2293–2326.

Liu, H., Lafferty, J. D. and Wasserman, L. A. (2009) The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.*, **10**, 2295–2328.

Ma, S., Gong, Q. and Bohnert, H. J. (2007) An arabidopsis gene network based on the graphical Gaussian model. *Genome Res.*, **17**, 1614–1625.

McCall, M. N. and Irizarry, R. A. (2011) Thawing frozen robust multi-array analysis (fRMA). *BMC Bioinform.*, **12**, article 369.

McCall, M. N., Jaffee, H. A., Zelisko, S. J., Sinha, N., Hooiveld, G., Irizarry, R. A. and Zilliox, M. J. (2014) The gene expression barcode 3.0: improved data processing and mining tools. *Nucleic Acids Res.*, **42**, D938–D943.

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

Nesterov, Y. (2005) Smooth minimization of non-smooth functions. *Math. Programng*, **103**, 127–152.

Oztoprak, F., Nocedal, J., Rennie, S. and Olsen, P. A. (2012) Newton-like methods for sparse inverse covariance estimation. In *Advances in Neural Information Processing Systems*. Red Hook: Curran Associates.

Peng, J., Wang, P., Zhou, N. and Zhu, J. (2009) Partial correlation estimation by joint sparse regression models. *J. Am. Statist. Ass.*, **104**, 735–746.

Phillips, M. A., León, P., Boronat, A. and Rodríguez-Concepción, M. (2008) The plastidial mep pathway: unified nomenclature and resources. *Trends Plnt Sci.*, **13**, 619–623.

Raskutti, G., Yu, B., Wainwright, M. J. and Ravikumar, P. K. (2008) Model selection in Gaussian graphical models: high-dimensional consistency of l1-regularized mle. In *Advances in Neural Information Processing Systems*. Red Hook: Curran Associates.

Ravikumar, P., Wainwright, M. J. and Lafferty, J. (2010) High-dimensional Ising model selection using l1-regularized logistic regression. *Ann. Statist.*, **38**, 1287–1319.

Rocha, G. V., Zhao, P. and Yu, B. (2008) A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (splice). *Preprint arXiv:0807.3734*. University of California at Berkeley, Berkeley.

Rolfs, B., Rajaratnam, B., Guillot, D., Wong, I. and Maleki, A. (2012) Iterative thresholding algorithm for sparse inverse covariance estimation. In *Advances in Neural Information Processing Systems*. Red Hook: Curran Associates.

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

Ruiz-Sola, M. Á. and Rodríguez-Concepción, M. (2012) Carotenoid biosynthesis in arabidopsis: a colorful pathway. *The Arabidopsis Book*, vol. 10. Rockville: American Society of Plant Biologists.

Sheppard, W. (1899) On the application of the theory of error to cases of normal distribution and normal correlation. *Philos. Trans. R. Soc. Lond.* A, **192**, 101–167.

Skrondal, A. and Rabe-Hesketh, S. (2007) Latent variable modelling: a survey. *Scand. J. Statist.*, **34**, 712–745.

Treister, E. and Turek, J. S. (2014) A block-coordinate descent approach for large-scale sparse inverse covariance estimation. In *Advances in Neural Information Processing Systems*.

Wang, L., Kim, Y. and Li, R. (2013) Calibrating non-convex penalized regression in ultra-high dimension. *Ann. Statist.*, **41**, 2505–2536.

Wille, A., Zimmermann, P., Vranová, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelic, A., von Rohr, P., Thiele, L., Zitzler, E., Gruissem, W. and Bühlmann, P. (2004) Sparse graphical Gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. *Genome Biol.*, **5**, article R92.

Xue, L. and Zou, H. (2012) Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann. Statist.*, **40**, 2541–2571.

Xue, L., Zou, H. and Cai, T. (2012) Nonconcave penalized composite conditional likelihood estimation of sparse Ising models. *Ann. Statist.*, **40**, 1403–1429.

Yang, E., Baker, Y., Ravikumar, P., Allen, G. and Liu, Z. (2014a) Mixed graphical models via exponential families. In *Proc. 17th Int. Conf. Artificial Intelligence and Statistics*. Red Hook: Curran Associates.

Yang, Z., Ning, Y. and Liu, H. (2014b) On semiparametric exponential family graphical models. *Preprint arXiv:1412.8697*. Princeton University, Princeton.

Yuan, M. (2010) High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.*, **11**, 2261–2286.

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

Zhang, C.-H. (2010) Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, **38**, 894–942.

Zhang, T. and Zou, H. (2014) Sparse precision matrix estimation via lasso penalized d-trace loss. *Biometrika*, **101**, 103–120.

Zhao, T., Roeder, K. and Liu, H. (2014) Positive semidefinite rank-based correlation matrix estimation with application to semiparametric graph estimation. *J. Computnl Graph. Statist.*, **23**, 895–922.

Zou, H. and Li, R. (2008) One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.*, **36**, 1509–1533.

*Supporting information*

Additional ‘supporting information’ may be found in the on-line version of this article:

Supplementary materials to “High dimensional semiparametric latent graphical model for mixed data”’.