Abstract
We consider parameter estimation and statistical inference of high-dimensional undirected graphical models for mixed data comprising both ordinal and continuous variables. We propose a flexible model called Latent Mixed Gaussian Copula Model that simultaneously deals with such mixed data by assuming that the observed ordinal variables are generated by latent variables. For parameter estimation, we introduce a convenient rank-based ensemble approach to estimate the latent correlation matrix, which can be subsequently applied to recover the latent graph structure. In addition, based on the ensemble estimator, we develop test statistics via a pseudo-likelihood approach to quantify the uncertainty associated with the low dimensional components of high-dimensional parameters. Our theoretical analysis shows the consistency of the estimator and asymptotic normality of the test statistic. Experiments on simulated and real gene expression data are conducted to validate our approach.

1 Introduction
Graphical models have become a fundamental tool to study the complex dependence structures over random variables. In a wide variety of applications such as social science, genetics, statistical physics and image processing, learning high dimensional undirected graphical models is of great interest. As a class of popular tools, Gaussian graphical models (GGMs) assume $\mathbf{X} \in \mathbb{R}^d$ follows a zero mean multivariate Gaussian distribution with covariance $\Sigma$, in which the conditional independence structure of variables is encoded by the precision matrix $\Omega = \Sigma^{-1}$: $X_i$ and $X_j$ are conditionally independent given remaining variables if and only if $\Omega_{ij} = 0$ (Lauritzen, 1996). There has been a large number of methods proposed for learning GGMs in high dimension, including the neighborhood selection (Meinshausen and Bühlmann, 2006), the penalized likelihood approach (Yuan and Lin, 2007; Friedman et al., 2008; Banerjee et al., 2008; Rothman et al., 2008), CLIME (Cai et al., 2011, 2016), among others. At the same time, the statistical inferential problems for high-dimensional GGMs attract increasing attention (Liu et al., 2013; Ren et al., 2015; Jankova et al., 2015; Janková and van de Geer, 2017). We refer to Cai et al. (2016); Fan et al. (2016); Drton and Maathuis (2017) for a more complete list of references on graphical models and Zhang and Zhang (2014); Javanmard and Montanari (2014); Van de Geer et al. (2014); Gu et al. (2015); Barber and Kolar (2015); Ning et al. (2017); Cai et al. (2017) for the recent advance in high-dimensional inference.

GGMs are not appropriate for modeling mixed data (i.e., a combination of categorical and continuous data), yet such type of data is ubiquitous in practice, particularly the mix of ordinal and continuous data (referred as ordinal-mixed data) are . For instance, in genetics the data may contain both continuous gene expression values as well as ordinal disease stages and phenotypic effects. In food sensory analysis, continuous rating and ordinal scales are commonly combined to describe different aspects of a food product. Directly treating the ordinal levels as numerical values may not truly reflect the dependence structure between ordinal and continuous variables. Nevertheless, high dimensional graphical models tailored to ordinal-mixed data have attracted less attention. Moreover, how to perform statistical inference on this type of model is largely unknown.

In this paper we propose a unified framework for estimation and statistical inference of the graphical model named Latent Mixed Gaussian Copula Model, which unifies and extends the models in (Ashford and Snowden, 1970; Amemiya, 1974; Guo et al., 2015; Fan et al.,...
2 Background

2.1 Notation.

Let $\mathbf{A} = (A_{jk}) \in \mathbb{R}^{d \times d}$ and $\mathbf{v} = (v_1, \ldots, v_d)^T \in \mathbb{R}^d$. We use $\mathbf{v}_S$ as the subvector of $\mathbf{v}$ with entries indexed by the set $S$. For a vector $\mathbf{v}$, we define $||\mathbf{v}||_1 = \sum_{i=1}^d v_i$, $||\mathbf{v}||_2 = (\sum_{i=1}^d v_i^2)^{1/2}$ and $||\mathbf{v}||_\infty = \max_{1 \leq i \leq d} |v_i|$. For a matrix $\mathbf{A}$, we define the matrix $\ell_1$ norm $||\mathbf{A}||_1 = \max_{1 \leq j \leq d} \sum_{i=1}^d |A_{ij}|$ and the matrix elementwise maximum norm $||\mathbf{A}||_{\max} = \max_{ij} |A_{ij}|$. We use $\mathbf{A} \succ 0$ to indicate that $\mathbf{A}$ is positive definite. We use $a_i \preceq b_i$ if $a_i \leq C b_i$ for some constant $C > 0$. We use $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ to denote the smallest and largest eigenvalues of $\mathbf{A}$.

2.2 Latent Gaussian Copula Model

Given a $d$-dimensional continuous variable $\mathbf{Z}$, we say $\mathbf{Z}$ satisfies the Gaussian copula model (Liu et al., 2009, 2012; Xue et al., 2012), i.e., $\mathbf{Z} \sim \text{NPN}(0, \Sigma, \mathbf{f})$, if $f(\mathbf{Z}) := (f_1(Z_1), \ldots, f_d(Z_d))^T \sim \text{N}_d(0, \Sigma)$ with $\Sigma_{jj} = 1$ and some monotonic transformations $(f_1, \ldots, f_d)$, where $\text{N}_d(\mu, \Sigma)$ denotes the multivariate Gaussian distribution with dimension $d$, mean $\mu$ and covariance $\Sigma$. To deal with ordinal-mixed data, we consider the following extension.

Definition 1. Latent Mixed Gaussian Copula Model Assume that $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, where $\mathbf{X}_1$ represents $d_1$-dimensional ordinal variables and $\mathbf{X}_2$ represents $d_2$-dimensional continuous variables. Suppose $X_j \in \{0, \ldots, N_j\}$ for $1 \leq j \leq d_1$. The random vector $\mathbf{X}$ satisfies the latent mixed Gaussian copula model, if there is a $d_1$-dimensional random vector $\mathbf{Z}_1 = (Z_1, \ldots, Z_{d_1})^T$ such that $\mathbf{Z} := (\mathbf{Z}_1, \mathbf{X}_2) \sim \text{NPN}(0, \Sigma, \mathbf{f})$ and

$$X_j = \sum_{k=1}^{N_j} \mathbb{I}(Z_j > C^k_j) \quad \forall j = 1, \ldots, d_1$$

where $C_{1j}^1 < C_{2j}^2 < \ldots < C_{N_j}^{N_j}$ is a sequence of constants. Let $\mathcal{C} = (C_1, \ldots, C_{d_1})$ where $C_j = (C_{1j}^1, \ldots, C_{N_j}^{N_j})$ and $\mathcal{N} = (N_1, \ldots, N_{d_1})$. We denote $\mathbf{X} \sim \text{LMNPN}(0, \Sigma, \mathbf{f}, \mathcal{C}, \mathcal{N})$.

The proposed model accounts for the ordinal-mixed data by assuming a latent continuous vector $\mathbf{Z}$ that generates the ordinal vector $\mathbf{X}_1$ via some unknown cutoffs $\mathcal{C}$. Let $\mathbf{\Omega} = \Sigma^{-1}$ be the precision matrix. The zero pattern of $\mathbf{\Omega}$ encodes the underlying conditional independence among latent variables $\mathbf{Z}$, even if the latent variables are not directly observable. Therefore it suffices to estimate the sparsity pattern of the latent precision matrix $\mathbf{\Omega}$. The proposed model is invariant to re-ordering the ordinal variables. Specifically, if $\mathbf{X}$ is a vector of ordinal variables and denote
$X_j^* = N_j - X_j$, the latent correlation matrix $\Sigma$ is invariant in the sense that if $\mathbf{X} \sim \text{LMNPN}(0, \Sigma, f, C, N)$ then $\mathbf{X}^* \sim \text{LMNPN}(0, \Sigma, f^*, C^*, N)$ for some $f^*$ and $C^*$. In addition, the model parameters are not fully identifiable; only the transformed unknown cutoffs $f_j(C^*_j)$ rather than the original cutoffs $C^*_j$ are identifiable for the ordinal entries. To ease notation, we define $\Delta_j(p) = f_j(C^*_j)$.

We note that when the input only contains ordinal data, and the monotonic transformations are trivial in the sense that $f(x) = x$, the proposed model is equivalent to the probit graphical model (Ashford and Sowden, 1970; Amemiya, 1974; Guo et al., 2015; Suggala et al., 2017). In addition, when all ordinal variables only contain two levels, the proposed model reduces to the model for binary-mixed data in Fan et al. (2017). We can see that the family of latent mixed Gaussian copula models is strictly larger than the aforementioned ones.

3 Methodology

3.1 Estimate the Latent Correlation Matrix

Given $n$ independent observations $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim \text{LMNPN}(0, \Sigma, f, C, N)$, in this section we introduce a procedure to estimate the latent covariance matrix $\Omega$. Our main idea is to firstly estimate the latent correlation matrix, which is then plugged into a quadratic optimization framework to reconstruct the sparsity pattern of $\Omega$.

Due to the existence of ordinal data, direct estimation of ordinal-ordinal and ordinal-continuous correlations become difficult. Under the framework of GGMs, Chen et al. (2014); Yang et al. (2014a,b); Lee and Hastie (2015); Cheng et al. (2017) resolve this issue by imposing additional distributional assumptions. However, under the proposed Latent Mixed Gaussian Copula Models, such methods are infeasible. Rank-based approach, on the other hand can be applied under the proposed framework (Liu et al., 2012; Xue et al., 2012; Fan et al., 2017), where in brief, the authors connect the latent correlation that is hard to estimate to a rank correlation coefficient such as Kendall’s $\tau$ or Spearman’s $\rho$, which are convenient to estimate.

Nevertheless, constructing an explicit connection between the rank correlation coefficients and the parameter of interest is complicated for ordinal variables, especially when the number of levels within the ordinal variable becomes large. More importantly, the convergence rate of the estimator cannot be derived when the connection function becomes involved. We therefore propose an ensemble approach: we firstly binarize the ordinal variable at each level and construct a set of preliminary rank-based estimators. At the next stage, we combine the preliminary estimators into a single but strong estimator, of which the theoretical property and empirical advantage over a single preliminary estimator is confirmed in section 4 and 5.

Formally, given an ordinal variable $X_{ij}$ for the $i$th observation, we consider the binary form with respect to level $p$ as $X_{ij}^{(p)} = 1(X_{ij} \geq p)$, $p = 1, \ldots, N_j$. In the case where $X_{ij}$ and $X_{ik}$ are both ordinal, consider the Kendall’s $\tau$ calculated from the binarized observed data

$$\hat{\tau}_{jk}^{(p,q)} = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < i' \leq n} \text{sign}((X_{ij}^{(p)} - X_{i'j}^{(p)})(X_{ik}^{(q)} - X_{i'k}^{(q)})$$

with $\text{sign}(0) = 0$.

Define $\Phi_2(u, v, s) = \int_{x_1 < u} \int_{x_2 < v} \phi_2(x_1, x_2; s)dx_1dx_2$ as the cumulative distribution function of the standard bivariate Gaussian distribution with correlation $s$, and $\Phi(\cdot)$ as the cumulative distribution of the standard Gaussian distribution. To recover the (latent) correlation between variable $j$ and $k$ from $\hat{\tau}_{jk}^{(p,q)}$, it can be shown that

$$\hat{\tau}_{jk}^{(p,q)} := \mathbb{E} [\hat{\tau}_{jk}^{(p,q)}] = F(\Sigma_{jk}; \hat{\Delta}_j^{(p)}, \hat{\Delta}_k^{(q)})$$

where

$$F(s; \Delta_j, \Delta_k) = 2\{\Phi_2(\Delta_j, \Delta_k, s) - \Phi(\Delta_j)\Phi(\Delta_k)\}$$

is an invertible function with respect to $s$.

In practice, $\hat{\Delta}_j^{(p)}$ and $\hat{\Delta}_k^{(q)}$ are unknown. Using the first moment condition, $\hat{\Delta}_j^{(p)}$ can be estimated by $\hat{\Delta}_j^{(p)} = \Phi^{-1}(1 - \frac{1}{n} \sum_{i=1}^n \tau_{ij}^{(p)})$, and similarly for $\hat{\Delta}_k^{(q)}$. Thus, we define the base estimate of $\Sigma_{jk}$ with respect to levels $p$ and $q$ as

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

Similarly, in the case where variable $X_j$ is ordinal and variable $X_k$ is continuous, the base estimate with respect to level $p$ is given by

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

where

$$\hat{\tau}_{jk}^{(p)} = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < i' \leq n} \text{sign}((X_{ij}^{(p)} - X_{i'j}^{(p)})(X_{ik} - X_{i'k}))$$

and

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

Finally, when both $X_{ij}$ and $X_{ik}$ are continuous, the estimator of $\Sigma_{jk}$ is given by

$$\hat{R}_{jk} = \sin(\frac{\pi}{2} \hat{\tau}_{jk}).$$
where \( \hat{\tau}_{jk} \) is the original Kendall’s \( \tau \) estimator Kendall (1948).

Apparently, the base estimates in (1) and (2) depend on the levels at which the data are binarized. When the ordinal data only contain two levels, the binarization step is unnecessary and \( \hat{R}_{jk}^{(p,q)} \) is independent of \( (p,q) \). However, for the ordinal data, one may construct multiple base estimators \( \{ \hat{R}_{jk}^{(p,q)} : p = 1, \ldots, N_j; q = 1, \ldots, N_k \} \) for the same target \( \Sigma_{jk} \). The final estimator can be potentially more efficient when the base estimators are properly aggregated.

The ensemble estimator for the ordinal-by-ordinal entry can be generally defined as

\[
\hat{R}_{jk} = \sum_{1 \leq p \leq N_j, 1 \leq q \leq N_k} w_{jk}^{(p,q)} \hat{R}_{jk}^{(p,q)},
\]

where for each \( (j,k) \) the weight \( w_{jk}^{(p,q)} \) is defined as

\[
0 \leq w_{jk}^{(p,q)} \leq 1, \quad \text{and} \quad \sum_{1 \leq p \leq N_j, 1 \leq q \leq N_k} w_{jk}^{(p,q)} = 1. \tag{4}
\]

Similarly, the ensemble estimator for the ordinal-by-continuous entry is defined as

\[
\hat{R}_{jk} = \sum_{1 \leq p \leq N_j} w_{jk}^{(p)} \hat{R}_{jk}^{(p)}.
\]

In the following, we illustrate two concrete examples for aggregating the base estimators.

**Example 1. Simple Average Ensemble** A natural idea is to take simple average of base estimators for each entry. This is equivalent to the ensemble estimator with the uniform weight \( w_{jk}^{(p,q)} = 1/(N_j N_k) \) and \( w_{jk}^{(p)} = 1/N_j \). Such simple approach is known to be efficient when the base estimators are calculated from independent data, see (Battey et al., 2015; Lee et al., 2015). However, our base estimators are possibly correlated and therefore the simple average estimator may not be optimal. As confirmed in section 5 ensemble estimators with a non-uniform weight may outperform the simple average method.

**Example 2. Data Adaptive Ensemble** Consider the following weighted ensemble rank-based estimator, where the weights are determined by the bivariate log likelihood with plug-in estimates of cutoffs. Intuitively, the base estimators associated with higher likelihood should enjoy higher weights, because the estimate is likely to be closer to the truth. Specifically, in the case where \( X_{ij} \) and \( X_{ik} \) are both ordinal, we define the weight as

\[
w_{jk}^{(p,q)} = \frac{\exp \left\{ l(\hat{R}_{jk}^{(p,q)}) \right\}}{\sum_{1 \leq m \leq N_j} \sum_{1 \leq n \leq N_k} \exp \left\{ l(\hat{R}_{jk}^{(m,n)}) \right\}},
\]

where

\[
l(\hat{R}_{jk}^{(p,q)}) = \sum_{i=1}^{n} \log \left( \int_{U_{ij}} \int_{U_{jk}} \phi_2(x_1, x_2; \hat{R}_{jk}^{(p,q)}) dx_1 dx_2 \right)
\]

is the bivariate log likelihood function and

\[
\hat{U}_{ij} = [\hat{\Delta}_j (X_{ij}), \hat{\Delta}_j (X_{ij} + 1)]
\]

with \( \hat{\Delta}_j (0) = -\infty \) and \( \hat{\Delta}_j (N_j + 1) = \infty \). The weight \( w_{jk}^{(p,q)} \) is proportional to the bivariate likelihood function of \( (X_{ij}, X_{ik}) \) evaluated at the estimate \( \hat{R}_{jk}^{(p,q)} \). Thus, unlike the simple average method, we can adaptively reweight the base estimator \( \hat{R}_{jk}^{(p,q)} \) according to whether it provides a good fit to the data. Similarly, when \( X_{ij} \) is ordinal and \( X_{ik} \) is continuous we have

\[
w_{jk}^{(p)} = \frac{\exp \left\{ l(\hat{R}_{jk}^{(p)}) \right\}}{\sum_{1 \leq m \leq N_j} \exp \left\{ l(\hat{R}_{jk}^{(m)}) \right\}},
\]

where

\[
l(\hat{R}_{jk}^{(p)}) = \sum_{i=1}^{n} \log \left( \int_{U_{ij}} \phi_2(x_1, X_{ik}; \hat{R}_{jk}^{(p)}) dx_1 \right).
\]

### 3.2 Learning and testing the latent graph structure

To learn the latent graph structure, it suffices to estimate the sparsity pattern of the precision matrix \( \Omega \). Write \( \Omega = (\beta_1, \ldots, \beta_d) \) and denote the true value of \( \Omega \) by \( \Omega^* = (\beta_{11}, \ldots, \beta_{dd}) \). We estimate \( \hat{\beta}_k \) by the following quadratic loss with \( \ell_1 \) penalty

\[
\hat{\beta}_k = \arg \min_{\beta} \left\{ \frac{1}{2} \beta^T \hat{R} \hat{\beta} - \beta^T \lambda \right\}, \tag{5}
\]

where \( \lambda \) is a tuning parameter and \( \hat{R} \) is the projection of \( \hat{R} \) onto the cone of positive definite matrices, that is, \( \hat{R} = \arg \min_{R \succ 0} \| \hat{R} - R \|_{\max} \). Similar to Liu and Luo (2012); Cai et al. (2011), the final estimator \( \hat{\Omega} \) is obtained after a symmetrization step

\[
\hat{\Omega}_{jk} = \hat{\Omega}_{kj} = \hat{\beta}_{jk} \mathbb{1} (|\hat{\beta}_{jk}| < |\hat{\beta}_{kj}|) + \hat{\beta}_{kj} \mathbb{1} (|\hat{\beta}_{kj}| < |\hat{\beta}_{jk}|).
\]

We note that similar to the Gaussian copula model, we may modify existing procedures by plugging \( \hat{R} \) into these methods, such as graphical lasso Friedman et al. (2008) and CLIME Cai et al. (2011), to learn the latent precision matrix. However these methods do not provide a good framework for statistical inference for our model. In contrast, the proposed method not only gives a consistent estimator but for testing and constructing confidence region, our approaches also build upon the estimator defined in (5).
A similar quadratic optimization approach to (5) with the sample covariance matrix is called SCIO (Liu and Luo, 2012). In our approach, we plug in the projected estimator \( \hat{R} \) rather than the ensemble estimator \( \tilde{R} \) itself, since \( \hat{R} \) may not be positive semidefinite. In practice, we find that the optimization (5) tends to be more robust than that with \( \tilde{R} \). More importantly, with the projected estimator \( \hat{R} \), one can establish a variety of error bounds for \( \beta \), which are essential for the statistical inference. The theoretical properties of \( \hat{R} \) are established in section 4.

In the following, we consider how to quantify the estimation uncertainty of the conditional independence between two pre-specified variables of interest. In particular, we aim to test the null hypothesis \( H_0 : \Omega_{jk} = 0 \). The existing high-dimensional inference methods for Gaussian (or transelliptical) graphical models rely on the Gaussian (or elliptical) structures and are not directly applied to our problem. In general, the maximum likelihood estimation is the default approach for statistical inference. However, our likelihood function is highly intractable due to the presence of high-dimensional integral for the ordinal data, which makes the likelihood based inference infeasible. To tackle this challenge, we borrow the loss function (5) in the proposed estimation procedure and view it as a pseudo-likelihood, and extend the likelihood based inference using the score function along the Stein’s least favorable direction (Ning et al., 2017) to our pseudo-likelihood.

To simplify notation, denote the \( k \)th column of \( \Omega \) as \( \beta \), which is further partitioned as \( \beta = (\theta, \gamma^T)^T \). Consider a partitioning of \( \Sigma \) as

\[
\Sigma = \begin{bmatrix}
\Sigma_{11} & \Sigma_{12} \\
\Sigma_{21} & \Sigma_{22}
\end{bmatrix}
\]

where \( \Sigma_{11} \) is scalar and \( \Sigma_{22} \) is \((d-1) \times (d-1)\). It can be shown that \( \nu^* = (1, -w^*^T)^T \) is the least favorable direction for the pseudo-likelihood (5), where \( w^* = \Sigma_{22}^{-1} \Sigma_{21} \). The gradient of the pseudo-likelihood (i.e., pseudo-score) is simply \( \hat{R} \beta - e_k \).

To construct the test statistic, one has to analyze the stochastic structure of the estimator \( \hat{R} \). However, this analysis seems mathematically intractable due to the projection of the estimator. To get around this issue, when defining the test statistic based on the pseudo-score in (5), we replace \( \hat{R} \) in the loss function by the ensemble estimator \( \tilde{R} \). With a nice U-statistics structure of \( \tilde{R} \), the pseudo-score function with \( \tilde{R} \) is more convenient.

We define the pseudo-score function in the direction of \( \nu^* \) as

\[
S(\beta) = \nu^{*T} (\tilde{R} \beta - e_k).
\]

By the definition of \( \nu^* \), one can estimate \( \nu^* \) by \( \hat{\nu} = (1, -\hat{w}^T)^T \), where

\[
\hat{w} = \arg\min \|w\|_1 \ s.t. \ |\tilde{R}_{12} - w^T \tilde{R}_{22}| \leq \lambda', \quad (6)
\]

where \( \lambda' \) is a tuning parameter. Since \( w \) is a sparse vector by \( \|w^*\|_0 = \|\Omega_{21}^*\|_0 \), the estimator \( \hat{w} \) in (6) is a sparse approximation of the true parameter \( w^* \). Denote \( \Phi^* = \lim_{n \to \infty} \text{Var}_\beta ((1/\sqrt{n})(\tilde{R} \beta^* - e_k)) \) and \( \sigma^* = \nu^{*T} \Phi^* \nu^* \). Let \( \hat{\beta}_0 = (0, \gamma^T)^T \) be the estimate under the null hypothesis and \( \hat{\sigma} \) be a consistent estimator of \( \sigma^* \). Our score test statistic is defined as

\[
T_n = \sqrt{n} \hat{S}(\hat{\beta}_0)/\hat{\sigma}^{1/2},
\]

where \( \hat{S}(\beta) = \hat{w}^T (\tilde{R} \beta - e_k) \). In the next section, we will show that \( T_n \to_d N(0,1) \) under \( H_0 \), and thus it can be used to evaluate the validity of the null hypothesis. The test statistic \( T_n \) can be equivalently used to construct confidence interval for \( \theta \); due to space limit please see the supplementary materials for details.

4 Theoretical properties\(^1\)

In this section we analyze the theoretical properties of the ensemble rank-based estimator of \( \Sigma \) and pseudo-score test statistic. Here we focus on the simple average estimator, and the theory can be easily generalized to the weighted case. We consider the following assumptions:

Assumption 1. \( |\tilde{\Omega}_{jk}| \leq 1 - \delta, \forall 1 \leq j < k \leq d_1 \) for some constant \( \delta > 0 \).

Assumption 2. \( |\tilde{\Omega}_{jk}| \leq M, \forall j = 1, \ldots, d_1, \, p = 1, \ldots, N_j \) for some constant \( M \).

Assumption 3. \( \|\tilde{\Omega}_*\|_1 \leq M_n \).

Assumption 4. \( 1/\tau \leq \lambda_{\min}(\Sigma^*) \) for some constant \( \tau > 0 \).

Assumptions 1 and 2 are mild technical assumptions. Assumption 1 makes sure that there is no perfect collinearity for any pair of \( f_j(Z_j) \) and \( f_k(Z_k) \). Assumption 2 is mainly used to control the variation of \( F^{-1}(\tau, \tilde{\Delta}_j, \tilde{\Delta}_k) \) with respect to \( (\tau, \tilde{\Delta}_j, \tilde{\Delta}_k) \). It can be shown that under these assumptions \( F^{-1}(\tau, \tilde{\Delta}_j, \tilde{\Delta}_k) \) is Lipschitz in \( \tau \) uniformly (Fan et al., 2017). Assumption 3 controls the magnitude of \( \Phi^* \), which is commonly used for the precision matrix estimation (Cai et al., 2011). Note that we allow \( M_n \) to increase with \( n \). Assumption 4 implies the standard restricted eigenvalue (RE) condition.

Proposition 1. Under Assumption 1 and 2, it holds that

\[
P(\|\tilde{R} - \Sigma^*\|_{\max} \leq C \sqrt{\frac{\log d}{n}}) \geq 1 - d^{-1},
\]

\(^1\)The proofs are given in the supplementary material.
where \( C \) is a constant independent of \( n \) and \( d \).

**Theorem 1.** Let \( s^* \) be the maximum degree of the graph and \( \lambda = C_n \sqrt{\log d / n} \). Under Assumption 1 - 4, with probability at least \( 1 - d^{-1} \), we have

\[
\|\hat{\beta} - \beta^*\|_1 \lesssim C_n s^* \sqrt{\frac{\log d}{n}},
\]

\[
(\hat{\beta} - \beta^*)^T \hat{R}(\hat{\beta} - \beta^*) \lesssim C_n^2 s^* \frac{\log d}{n}.
\]

Theorem 1 shows that our estimator \( \hat{\beta} \) achieves the same convergence rate as the SCIO estimator Liu and Luo (2012) and CLIME Cai et al. (2011) for GGMs.

To establish the asymptotic distribution of the test statistic \( T_n \), we additionally make the following assumption:

**Assumption 5.** \( \lambda_{\max}(\Sigma^*) \leq \kappa \) for some constant \( \kappa > 0 \).

This assumption has been made in Ren et al. (2015); Jankova et al. (2015); Janková and van de Geer (2017) to study the testing problems for GGMs. It is natural for many important classes of covariance matrices, e.g., bandable, Toeplitz, and sparse covariance matrices. The following theorem validates the asymptotic normality of the test statistic.

**Theorem 2.** Suppose Assumptions 1 - 5 hold. If \( \lambda = \lambda' = C_n \sqrt{\log d / n} \) and \( \sigma^* \geq K \) for some constant \( K \), then under the null hypothesis we have

\[
T_n \rightarrow_{d} N(0, 1).
\]

When \( C_n \) is a constant, our condition \( M_n^2 s^* \log d / \sqrt{n} = o(1) \) is identical to the best possible scaling under GGMs established by Ren et al. (2015). In general, our condition appears to be stronger, as \( M_n \) may increase with \( n \). This can be viewed as the price paid for estimating the mixed graphical model, due to the sophisticated U-statistic structure of \( \hat{R} \). In particular, we show in the proof that the remaining term in the expansion of \( T_n \) scales with \( M_n^2 s^* \log d / \sqrt{n} \), which needs to be small enough to prove the normal approximation result.

5 Simulation study

In this section we evaluate the empirical performance of our methods on a set of synthetic datasets. Our data generating procedure is similar to that in (Liu et al., 2012). Specifically, for the precision matrix \( \Omega \), we set \( \Omega_{ij} = 1 \), and \( \Omega_{jk} = t \alpha_{jk} \), where \( t \) is chosen to ensure the positive definiteness of \( \Omega \) and \( \alpha_{jk} \) is a Bernoulli random variable with success probability \( p_{jk} = (2\pi)^{-1/2} \exp(\{||z_j - z_k||^2 / 2s^2\}) \). Here \( z_j = (z_j^{(1)}, z_j^{(2)}) \) is sampled from a bivariate uniform distribution on \([0, 1]\), and \( s \) is chosen to control the sparsity level of \( \Omega \). In our case we choose \( s \) to make sure there are about 200 edges in each latent graph.

After the latent precision matrix \( \Omega \) is created, we take the inverse and rescale it to obtain the correlation matrix \( \Sigma \), where the diagonal elements are 1. To generate the observed ordinal data, for \( X_j \) with \( N_j + 1 \) levels, we sample a sequence of cutoffs \( C^{(k)}_j \sim \text{Uniform}[0, \Omega^{-1}(k - 0.5) \frac{1}{\sqrt{N_j}}] \) for \( k = 1, \ldots, N_j \). This procedure ensures the randomness of the simulated cutoffs and guarantees that the difference of the number of samples at each level is not too large. In the following we fix the sample size \( n = 200 \), choose \( t = 0.15 \) and repeat the simulation 50 times for each scenario in low dimensional case \( d = 50 \) and high dimensional case \( d = 250 \):

(a) Simulate \( X = (X_1, \ldots, X_d) \), where \( X_j = \sum_{i=1}^{2} \mathbb{1}(Z_j > C^{(i)}_j) \) and \( Z \sim N(0, \Sigma) \).

(b) Simulate \( X = (X_1, \ldots, X_d) \), where \( X_j = \sum_{i=1}^{4} \mathbb{1}(Z_j > C^{(i)}_j) \) and \( Z \sim N(0, \Sigma) \).

(c) Simulate \( X = (X_1, \ldots, X_d) \), where \( X_j = \sum_{i=1}^{2} \mathbb{1}(Z_j > C^{(i)}_j) \) for \( j = 1, \ldots, d/2 \), \( Z \sim N(0, \Sigma) \), and \( X_{d-j} = Z_{d-j} \) for \( j = d/2 + 1, \ldots, d \).

(d) Simulate \( X = (X_1, \ldots, X_d) \), where \( X_j = \sum_{i=1}^{2} \mathbb{1}(Z_j > C^{(i)}_j) \) for \( j = 1, \ldots, d/2 \), \( Z \sim NPN(0, \Sigma, f) \), where \( f_j(x) = x^3 \) for \( j = d/2 + 1, \ldots, d \).

To investigate the empirical estimation error of the latent correlation and precision matrix, we compare our method LMNPN\(^2\) with (1) the rank-based estimator RBE (Fan et al., 2017) and (2) the elementwise maximum likelihood approach EMLE with plug-in estimates of cutoffs (Suggala et al., 2017). Since both method can not be directly applied to ordinal-mixed data, we adapt both methods to this case with simple extension. Due to space limit please refer to the supplementary material for more details.

Notice that the precision matrix estimation problem for mixed data is not considered in the class of node-wise regression methods (Lee and Hastie, 2015; Chen et al., 2014; Cheng et al., 2017; Yang et al., 2014a). We will only compare this class of methods for the graph recovery properties.

In Table 1 and 2 LMNPN\(_1\) stands for the LMNPN method with simple average ensemble and LMNPN\(_2\) stands for data adaptive ensemble. The numbers in the parenthesis are the simulation standard errors.
Table 1: Average estimation error of latent correlation matrix

<table>
<thead>
<tr>
<th>d</th>
<th>Scenario</th>
<th>LMNPN_1</th>
<th>LMNPN_2</th>
<th>RBE</th>
<th>EMLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>(a)</td>
<td>4.58</td>
<td>4.45</td>
<td>5.81</td>
<td>10.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.08)</td>
<td>(0.10)</td>
<td>(0.11)</td>
<td>(0.48)</td>
</tr>
<tr>
<td></td>
<td>(b)</td>
<td>4.02</td>
<td>3.87</td>
<td>6.37</td>
<td>7.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.09)</td>
<td>(0.11)</td>
<td>(0.17)</td>
<td>(0.26)</td>
</tr>
<tr>
<td></td>
<td>(c)</td>
<td>4.15</td>
<td>4.16</td>
<td>4.80</td>
<td>5.54</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.10)</td>
<td>(0.09)</td>
<td>(0.12)</td>
<td>(0.23)</td>
</tr>
<tr>
<td></td>
<td>(d)</td>
<td>4.10</td>
<td>4.18</td>
<td>4.78</td>
<td>5.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.11)</td>
<td>(0.10)</td>
<td>(0.18)</td>
<td>(0.14)</td>
</tr>
<tr>
<td>250</td>
<td>(a)</td>
<td>23.16</td>
<td>21.82</td>
<td>29.46</td>
<td>50.60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.06)</td>
<td>(0.09)</td>
<td>(0.15)</td>
<td>(0.87)</td>
</tr>
<tr>
<td></td>
<td>(b)</td>
<td>20.00</td>
<td>19.38</td>
<td>23.92</td>
<td>21.06</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.10)</td>
<td>(0.05)</td>
<td>(0.15)</td>
<td>(0.62)</td>
</tr>
<tr>
<td></td>
<td>(c)</td>
<td>20.49</td>
<td>19.10</td>
<td>23.78</td>
<td>21.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.08)</td>
<td>(0.11)</td>
<td>(0.03)</td>
<td>(0.34)</td>
</tr>
<tr>
<td></td>
<td>(d)</td>
<td>20.45</td>
<td>19.39</td>
<td>23.89</td>
<td>21.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.04)</td>
<td>(0.07)</td>
<td>(0.03)</td>
<td>(0.36)</td>
</tr>
</tbody>
</table>

Tables 1 and 2 report the mean estimation error in terms of the Frobenius norm. It suggests that the ensemble rank-based estimator suffers less loss of information than the RBE estimator and the EMLE estimator in terms of both correlation and precision matrix estimation. The improvement upon the RBE estimator to the proposed one is more transparent when the number of levels for the ordinal data grows from three to five, comparing case (a) and case (b). For the proposed method, the data adaptive ensemble performs generally better than simple average ensemble for latent correlation matrix estimation and achieves similar performance for latent precision matrix.

5.1 Graph Structure Recovery

To investigate the graph recovery performance, we define the number of false positive \( \text{FP}(\lambda) \) and true positive \( \text{TP}(\lambda) \) given an estimator \( \hat{\Omega}^\lambda \) as

\[
\text{FP}(\lambda) := \left| \{(j, k) : \Omega^\lambda_{jk} = 0, \hat{\Omega}^\lambda_{jk} \neq 0\} \right|
\]

\[
\text{TP}(\lambda) := \left| \{(j, k) : \Omega^\lambda_{jk} \neq 0, \hat{\Omega}^\lambda_{jk} \neq 0\} \right|
\]

We further define the false positive rate \( \text{FPR}(\lambda) \) and true positive rate \( \text{TPR}(\lambda) \) as

\[
\text{FPR}(\lambda) = \frac{\text{FP}(\lambda)}{d(d-1)/2 - |E|}
\]

\[
\text{TPR}(\lambda) = \frac{\text{TP}(\lambda)}{|E|}
\]

In addition to the RBE and EMLE method, we compare the proposed LMNPN estimator via simple average ensemble with other approaches including: the weighted \( \ell_1 \) penalized nodewise regression \( \text{WNR} \) (Cheng et al., 2017) and pairwise exponential family nodewise regression \( \text{PNR} \) (Chen et al., 2014). According to the ROC curves in Figure 1, the proposed method generally outperforms competing methods.

Figure 1: Left (a) - (d): ROC curves for graph recovery of the proposed method (---), RBE (---), WNR (---), PNR (---) and EMLE (---) when \( d = 50 \) for scenario (a),(b),(c) and (d).

5.2 Asymptotic Normality

Finally, we empirically validate the asymptotic normality of the score test statistic. For simplicity, we
adopt scenario (c) to generate the mixed data. We consider the score test statistic $T_n$ for the null hypothesis $H_0 : \Omega_{jk} = 0$ for some pre-specified $(j, k)$. Two cases are studied: case (1) $X_{ij}$ and $X_{ik}$ are both ordinal and case (2) $X_{ij}$ is ordinal and $X_{ik}$ are continuous. The tuning parameters are chosen by the cross-validation. The results in Figure 2 implies that the proposed test statistic agrees well with the standard normal distribution, which validates our Theorem 2.

![Figure 2: Histograms for the score test statistic for (a): an ordinal-by-ordinal entry and (b): ordinal-by-continuous entry with the standard normal density superimposed over 500 iterations.](image)

6 Real data application

We analyze the NKI breast cancer dataset (http://compbio.dfci.harvard.edu/) that contains 24481 gene expression and multiple clinical information of 337 patients with primary breast carcinomas (Van’t Veer et al., 2002). The proposed method is advantageous when researchers want to simultaneously study the association among genes, clinical measures, and phenotype-genotype associations. For the purpose of our analysis, we filtered the data by removing all patients, gene expressions and clinical information with missing entries, and select the top 500 expressed loci. We then apply the proposed method on the filtered dataset with a tuning parameter that controls the sparsity level of the latent graph at around two percent.

According to Figure 3, the latent graph indicates associations between two ordinal clinical features (oestrogen receptor (ER) and tumor grade (Grade)) and several genes. In particular, ER has been shown to be a key driven factor of breast cancers (Carroll, 2016), and thus the strong associations between ER and multiple highly expressed gene loci in breast cancer patients can be justified. In addition, our latent graph also shows association between the breast tumor grade and locus NM_020974 (Homo sapiens signal peptide, CUB domain and EGF like domain containing 2 (SCUBE2), transcript variant 1, mRNA). According to Poorhosseini et al. (2016), over-expression of SCUBE2 has been shown to be associated with breast cancer recurrence. In another study, SCUBE2 has been shown to suppresses proliferation in breast tumor cell and confers a favorable prognosis in invasive breast cancer (Cheng et al., 2009). These results confirm the validity of the proposed method, and we expect other associations unveiled by the latent graph can be helpful for future research.

![Figure 3: The estimated latent graph for the mixed breast cancer dataset using the proposed method. Nodes (genes) without any edge are omitted.](image)

7 Conclusion

In this work we present and analyze a general framework for estimating and inferring a graphical model, called Latent Mixed Gaussian Copula Model. The proposed framework is shown to be effective both theoretically and empirically. Since most methods did not consider graphical models tailored to ordinal-mixed data and the fact that such mixed data are ubiquitous in practice, we believe our work fill this gap by providing a convenient and efficient modeling framework with theoretical guarantees for the community.

References


Huijie Feng, Yang Ning


Maurice George Kendall. Rank correlation methods. 1948.


