Revision: Point-to-point cover letter

In the response below, we address the reviewers' comments point-by-point. We want to thank the reviewers for their careful reading of the manuscript and their constructive comments. We have thoroughly revised the manuscript according to the reviewers' suggestions.

Reviewer 1

Reviewer Point P 1.1 — In Definition 2.1, when the model moves to the Gaussian copula, I think the mean parameter becomes redundant and cannot be separated from the transformations. Some comments on this would be helpful.

***Reply**: You are completely right. In fact, this is not the only identifiability issue that the latent Gaussian copula model (LGCM) faces. We added the following lines to the manuscript to address this and related issues:*

Changed: Several identifiability issues arise in the latent Gaussian copula class. First, the mean and the variances are not identifiable unless the monotone transformations f were restricted to preserve them. Note that this only affects the diagonal entries in $\mathbf{\Sigma}$, not the full covariance matrix. Therefore, without loss of generality, we assume the mean to be the zero vector and $\Sigma_{jj} = 1$ for all $j \in [d]$. Another identifiability issue relates to the unknown threshold parameters. To ease notation, let $\Gamma_j^r \equiv f_j(\gamma_j^r)$ and $\Gamma_j \equiv \{f_j(\gamma_j^r)\}_{r=0}^{l_j+1}$. In the LGCM, only the transformed thresholds Γ_j rather than the original thresholds are identifiable from the discrete variables. We assume, without loss of generality, that the transformed thresholds retain the limiting behavior of the original thresholds, i.e., $\Gamma_j^0 = -\infty$ and $\Gamma_j^{l_j+1} = \infty$.

Reviewer Point P 1.2 — The Section 3.2 on nonparanormal case 2 seems very interesting. However, their connection with ML under latent Gaussian model seems a little bit unclear. For example, the likelihood under latent Gaussian is given in eq (3). One natural idea is that, let's try to write down the analog of (3) under the nonparanormal. Of course, this will depend on the unknown function f, but we can plug in the estimator $\hat{f}$ in Section 3.2. This pseudo-likelihood approach is conceptually reasonable. Is there any connection with the proposed method in current Sec 3.2?

***Reply**: Thank you for pointing us in this direction. We engaged in a short comparison of the two approaches and added the paragraphs below to the Supplementary*

Materials. Formulating concentration inequalities for this pseudolikelihood approach becomes even more complex due to the polyserial likelihood function's intricate form and the multiple occurrences of the estimated transformations. Therefore, we lead with this empirical study to demonstrate the ad hoc estimator's computational efficiency. However, the pseudolikelihood approach is an interesting avenue for future research.

Changed: In this section, we conduct an empirical comparison of *Case II* estimators within the framework of the LGCM. Specifically, we examine the *Case II* MLE derived under the latent Gaussian model, as discussed in Section 2.1 of the Manuscript, which involves incorporating the estimated transformations $\hat{f}$ at appropriate locations. Furthermore, we investigate the ad hoc estimator presented in Section 3.2 of the Manuscript in more detail.

We start by rewriting Eq. (6), where we replace occurrences of the X_k with the corresponding transformation $\hat{f}_k(X_k)$. The resulting transformation-based first-order condition (FOC) for the *Case II* MLE under the LGCM becomes:

$$
\begin{aligned}
&\frac{\partial \ell_{jk}^{(n)}(\Sigma_{jk}, x_j^r, \hat{f}_k(x_k))}{\partial \Sigma_{jk}} \\
&= \sum_{i=1}^{n} \Bigg[\frac{1}{\Phi(\tilde{\Gamma}_j^r(\hat{f}_k)) - \Phi(\tilde{\Gamma}_j^{r-1}(\hat{f}_k))} (1 - (\Sigma_{jk})^2)^{-\frac{3}{2}} \\
&\qquad \Big[\phi(\tilde{\Gamma}_j^r(\hat{f}_k))(\Gamma_j^r \Sigma_{jk} - \hat{f}_k(\tilde{x}_{ik})) - \phi(\tilde{\Gamma}_j^{r-1}(\hat{f}_k))(\Gamma_j^{r-1} \Sigma_{jk} - \hat{f}_k(\tilde{x}_{ik})) \Big] \Bigg],
\end{aligned}
$$

where

$$
\hat{f}_k(\tilde{x}_{ik}) = \frac{\hat{f}_k(x_{ik}) - \hat{f}_k(\bar{x}_k)}{\sqrt{\hat{f}_k(s_k)^2}} \quad \text{and} \quad \tilde{\Gamma}_j^r(\hat{f}_k) = \frac{\Gamma_j^r - \Sigma_{jk} \hat{f}_k(\tilde{x}_{ik})}{\sqrt{1 - (\Sigma_{jk})^2}}.
$$

In the ensuing empirical evaluation, the data generation scheme is as follows. First, we generate n data points $(z_{ij}, x_{ik})_{i=1}^n$ from a standard bivariate normal with correlation Σ_{jk}^*. Second, we apply the same transformation $f_t^{-1}(x) = 5x^5$ for $t \in \{j, k\}$ to all the data points. Third, we generate binary data x_{ij}^r by randomly choosing $f_j^{-1}(z_{ij})$-thresholds (guaranteeing relatively balanced classes) and then applying inversion sampling.

Computing the transformation-based MLE for *Case II* can be achieved in several ways. Consider the plugged-in log-likelihood function, i.e.,

$$
\begin{aligned}
\ell_{jk}^{(n)}(\Sigma_{jk}, x_j^r, \hat{f}_k(x_k)) &= \sum_{i=1}^{n} \big[\log(p(\hat{f}_k(x_{ik}))) + \log(p(x_j^r \mid \hat{f}_k(x_{ik}), \Sigma_{jk})) \big] \\
&= \sum_{i=1}^{n} \big[\log(p(\hat{f}_k(x_{ik}))) + \Phi(\tilde{\Gamma}_j^r(\hat{f}_k)) - \Phi(\tilde{\Gamma}_j^{r-1}(\hat{f}_k)) \big].
\end{aligned}
$$

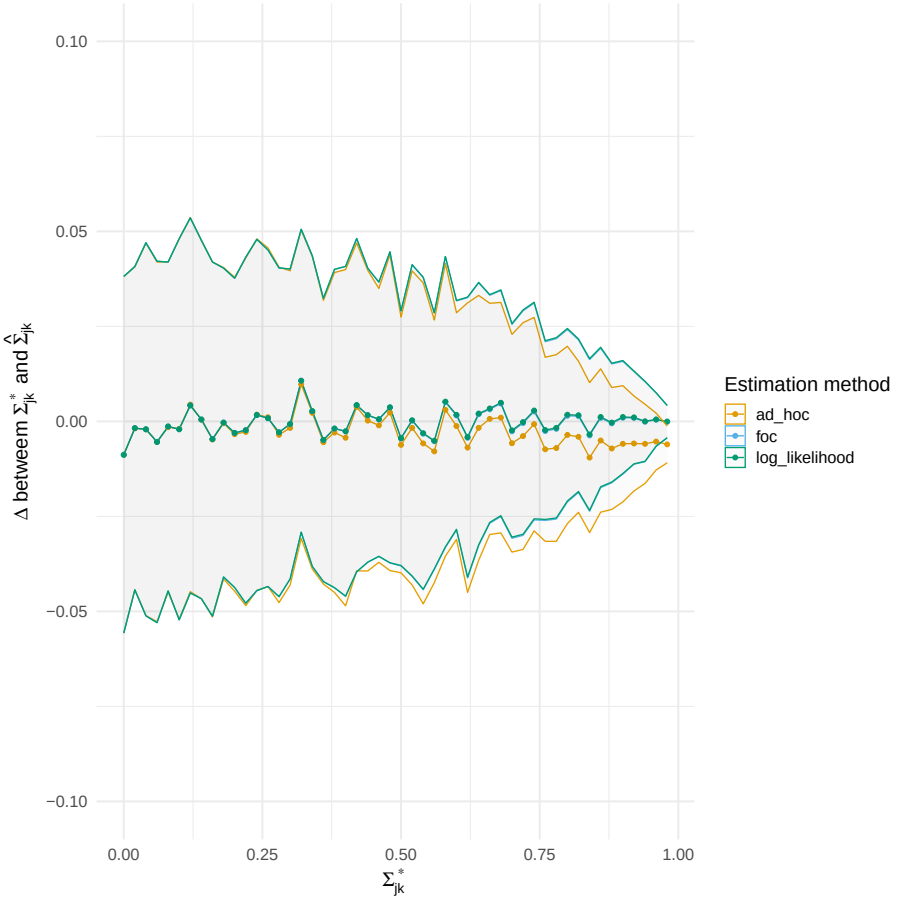

FIG 1. Comparison of the Case II MLE under the LGCM and the ad hoc estimator.

One strategy to optimize the function is direct maximization with a quasi-Newton optimization procedure to determine the optimal values for $\hat{\Sigma}_{jk}$. This strategy is used, for instance, in the R package `polycor` [14]. Alternatively, another approach involves utilizing the FOC and solving for $\hat{\Sigma}_{jk}$ through a nonlinear root-finding method. To do this, we employ Broyden's method [5].

In Figure 1, we generate data according to the scheme above for $n = 1000$ and a grid of true correlation values $\Sigma^*_{jk} \in [0, 0.98]$ with a step size of $s = 0.02$. Due to symmetry, taking only positive correlations is sufficient for comparison purposes. For each correlation value along the grid, we generate 100 mixed binary-continuous data pairs and compute the MLE (using the abovementioned strategies) and the ad hoc estimator from Section 3.2 in the main text. We plot the true correlation values against the absolute difference between estimates and true correlation and the corresponding Monte-Carlo standard error for the MLE and the ad hoc estimator.

As expected, both strategies for attaining the MLE yield the same results. The

ad hoc estimator's bias becomes noticeable only when the underlying correlation Σ^*_{jk} exceeds 0.75 and it remains at such a mild level that we consider it negligible; see [29] for a similar observation. The strength of the ad hoc estimator lies in its simplicity and computational efficiency. The left panel of Figure 2 shows the median computation time surrounded by the first and third quartiles. We compare the two MLE optimization strategies and the ad hoc estimator for a grid of sample sizes $n \in [50, 10000]$ with a step size of $s_t = 50$. Here we fix $\Sigma^*_{jk} = .87$ and repeat each calculation 100 times recording the time elapsed.

The right panel of Figure 2 demonstrates computation time across a grid of length 200 of values for $\Sigma^*_{jk} \in [-.98, 98]$. The sample size is, in this case, fixed at $n = 1000$. The ad hoc estimator is consistently and considerably faster than the MLE, regardless of the strategy used. The difference in computation time is especially pronounced for large sample sizes and correlation values approaching the endpoints of the $[-1, 1]$-interval. Setting the FOC to zero and solving for Σ_{jk} is computationally more efficient than directly maximizing the log-likelihood function. The time difference in MLE strategies is more pronounced at the endpoints of the $[-1, 1]$ interval. The ad hoc estimator is not affected by this issue. Therefore, in the high-dimensional setting we consider in this paper, the ad hoc estimator is preferable to the MLE due to (1) its computational efficiency, (2) its simplicity, which allows us to form concentration inequalities, and (3) its robustness to the underlying correlation value.

Reviewer Point P 1.3 — Theorem 3.2 seems confusing to me. First, it would be more clear if the authors can clarify that this result is for latent Gaussian model (if I am correct). Second, if the right hand side of (14) depends on α, I think the probability of that event on the left hand side should also depend on α, but is not the case in the current formulation. Please clarify this point.

***Reply**: We apologize for the confusion. We clarified the structure of the section and fixed the statement of Theorem 3.2.*

Changed: The subsequent theorem, drawing on Mei, Bai and Montanari [24], hinges on four conditions, all substantiated in Section 2 of the Supplementary Materials. This concentration result specifically pertains to the MLEs introduced in Section 2.1 within the framework of the latent Gaussian model. We remark that related methodology has been applied by Anne, Aurélie and Clémence [1] in addressing zero-inflated Gaussian data under double truncation.

Theorem. *Suppose that Assumptions 3.1–3.3 hold, and let $j \in [d_1]$ and $k \in [d_2]$ for Case II and $j, k \in [d_1]$ for Case III. Let $\alpha \in (0, 1)$, and let $n \geq 4C \log(n) \log\left(\frac{B}{\alpha}\right)$ for known constants B, C, and D depending on cases II and*

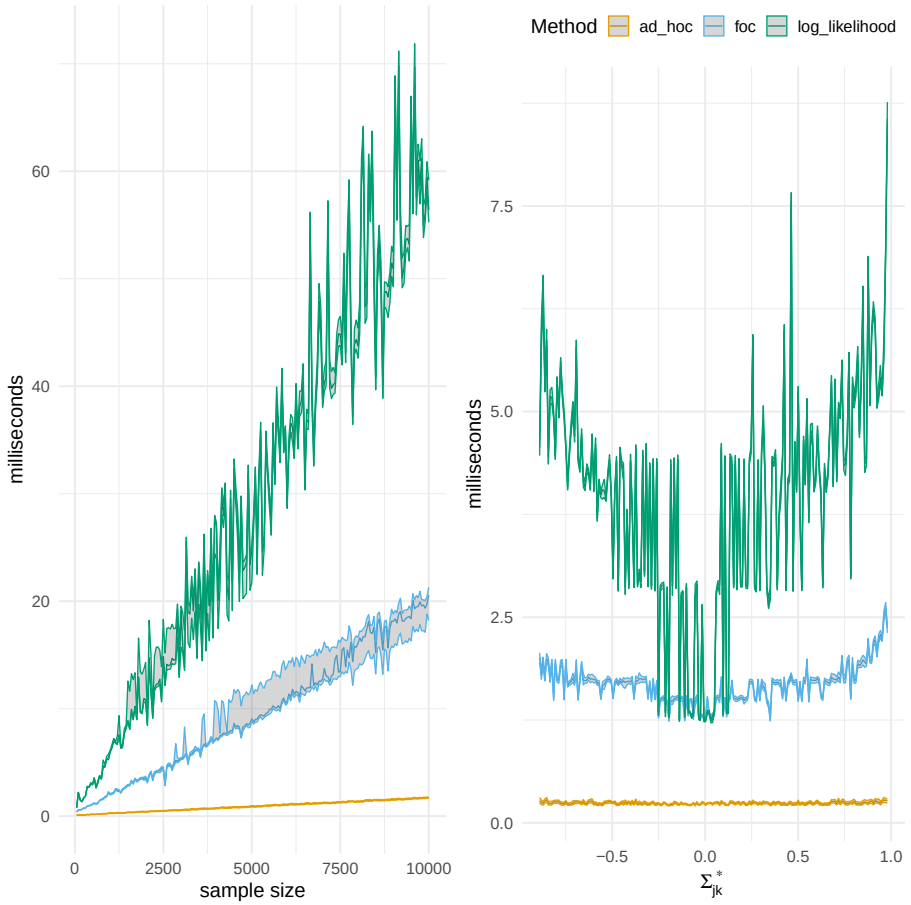

FIG 2. Computation time in milliseconds for the Case II MLE and ad hoc estimators. We report the median (solid line) and the first and third quartile (shaded area) of recorded computation time. In the left panel, we compare computation time against a grid of sample sizes $n \in [50, 10000]$ *with a step size of* $s_t = 50$*. In the right panel, we compare computation time against a grid of true correlation values* $\Sigma^*_{jk} \in [-.98, .98]$*.*

III but independent of (n, d)*. Then, it holds that*

$$P\left(\max_{j,k}\left|\hat{\Sigma}_{jk}^{(n)} - \Sigma_{jk}^{*}\right| \geq D\sqrt{\frac{\log(n)}{n}\log\left(\frac{B}{\alpha}\right)}\right) \leq \frac{d(d-1)}{2}\alpha.$$

Case I of the latent Gaussian model addresses the well-understood scenario involving observed Gaussian variables, with concentration results and rates of convergence readily available -see, for example, Lemma 1 in Ravikumar et al. [35]. Consequently, the MLEs converge to $\mathbf{\Sigma}^{*}$ at the optimal rate of $n^{-1/2}$, mirroring the convergence rate as if the underlying latent variables were directly observed.

Reviewer Point P 1.4 — It seems that the concentration in Theorem 3.3 is relatively slow, so that one can only obtain the rate $n^{-1/4}$ rather than the standard $n^{-1/2}$ (ignoring all log factors). I can expect that the technical analysis is very involved, but it is unclear to me which terms (from estimating f or other quantities) make the rate slow. Please provide some discussion along this line.

Reply*: Thanks for raising this issue. We added a discussion on the convergence rate below Corollary 3.4.:*

Changed: The nonparanormal estimator for *Case II* converges to Σ_{jk}^{*} at rate $n^{-1/4}$, which is slower than the optimal parametric rate of $n^{-1/2}$. This stems not from the presence of the discrete variable but from the direct estimation of the transformation functions f_j and the corresponding truncation constant δ_n. Both Xue and Zou [42] and Liu et al. [23] discuss room for improvement of the estimator for f_j to get a rate closer to the optimal one. Improvements depend strongly on the choice of truncation constant, striking a bias-variance balance in high-dimensional settings. In all of the numerical experiments below, we find that Theorem 3.3 gives a worst-case rate that does not appear to negatively impact performance compared to estimators that attain the optimal rate.

Reviewer 2

Reviewer Point P 2.1 — **Threshold estimation**: As the authors noted, the two-step estimation procedure relies crucially on the estimation of the unknown thresholds which map the latent Gaussian variable to the observed ordinal variable. The quantile estimator in 3.4 makes sense but I cannot see how Lemma 3.1 "assures that these threshold estimates can be formed with high accuracy." Lemma 3.1 seems to say that the estimators are bounded away from infinity

with high probability. Maybe I'm missing some important connections here. In addition, I assume A^{cj} stands for the complement of A^j, i.e. $(A^j)^c$.

***Reply**: You are right; the connection between Lemma 3.1 and the threshold estimation was unclear. We fixed the statement accordingly.*

Changed: The following lemma assures that the threshold estimates can be formed with high accuracy.

Lemma 0.1. *Suppose the estimated thresholds are bounded away from infinity, i.e., $|\hat{\Gamma}_j^r| \leq G$ for all $j \in [d_1]$ and $r = 1, \ldots, l_j$ and some G. The following bound holds for all $t > 0$ with Lipschitz constant $L_1 = 1/(\sqrt{\frac{2}{\pi}} \min\{\hat{\pi}_j^r, 1 - \hat{\pi}_j^r\})$:*

$$P\Big(|\hat{\Gamma}_j^r - \Gamma_j^r| \geq t\Big) \leq 2 \exp\Big(-\frac{2t^2 n}{L_1^2}\Big).$$

The proof of Lemma 3.1 is given in Section 5 of the Supplementary Materials. The requirement that the estimated thresholds are bounded away from infinity typically does not pose any restriction in finite samples. All herein-developed methods are applied in a two-step fashion. In the ensuing theoretical results, we stress this by denoting the estimated thresholds as $\bar{\Gamma}_j^r$.

Reviewer Point P 2.2 — Concentration results: Given that this manuscript is proposing a unifying model for mixed graphical models, it might help the readers if the authors can compare the rates in Theorems 3.2 and 3.3 with existing rates in special cases. In particular, (i) the concentration inequality (14) should match existing rates for Gaussian graphical models when plugging in proper α, and (ii) the five terms in the probability bound in Theorem 3.3 can use some explanation.

***Reply**: Thanks for the suggestion. We added a comparison of the concentration results to the existing rates for Gaussian graphical models. We also discuss the suboptimal rate of convergence in Theorem 3.3. As we mention below, there is room for improvement of the estimator for f_j to get a rate closer to the optimal one. This is something we are currently investigating further.*

Changed:

Theorem. *Suppose that Assumptions 3.1–3.3 hold, and let $j \in [d_1]$ and $k \in [d_2]$ for Case II and $j, k \in [d_1]$ for Case III. Let $\alpha \in (0, 1)$, and let $n \geq 4C \log(n) \log\left(\frac{B}{\alpha}\right)$ for known constants B, C, and D depending on cases II and III but independent of (n, d). Then, it holds that*

$$P\left(\max_{j,k}\left|\hat{\Sigma}_{jk}^{(n)} - \Sigma_{jk}^*\right| \geq D\sqrt{\frac{\log(n)}{n}\log\left(\frac{B}{\alpha}\right)}\right) \leq \frac{d(d-1)}{2}\alpha.$$

Case I of the latent Gaussian model addresses the well-understood scenario involving observed Gaussian variables, with concentration results and rates of convergence readily available -see, for example, Lemma 1 in Ravikumar et al. [35]. Consequently, the MLEs converge to $\mathbf{\Sigma}^*$ at the optimal rate of $n^{-1/2}$, mirroring the convergence rate as if the underlying latent variables were directly observed.

Changed: The first four terms in the probability bound stem from finding bounds to different regions of the support of the transformed continuous variable. The last term is a consequence of the fact that we estimate the transform directly.

The nonparanormal estimator for *Case II* converges to Σ^*_{jk} at rate $n^{-1/4}$, which is slower than the optimal parametric rate of $n^{-1/2}$. This stems not from the presence of the discrete variable but from the direct estimation of the transformation function f_j and the corresponding truncation constant δ_n. Both Xue and Zou [42] and Liu et al. [23] discuss room for improvement of the estimator for f_j to get a rate closer to the optimal one. Improvements depend strongly on the choice of truncation constant, striking a bias-variance balance in high-dimensional settings. In all of the numerical experiments below, we find that Theorem 3.3 gives a worst-case rate that does not appear to negatively impact performance compared to estimators that attain the optimal rate.

Reviewer Point P 2.3 — Numerical Experiments: It seems that too many details are deferred to the supplementary materials that a reader cannot understand how the data are simulated by reading the main text. If space is of concern, the two tables can be condensed into two plots. Furthermore, some rows (e.g., TPR and FPR) can be deferred to the supplementary materials as there are no additional insights to learn from these numbers. In addition, the real data analysis can also be deferred to the supplementary materials. I fail to gain any new insights about Covid from this analysis, and the texts in Fig. 1 and 2 are too small to read. The data analysis can be improved, but, in my opinion, it serves only as a distraction to the paper considering that the main contribution of this paper is a general model. Lastly, the low number of replicates in simulation (100 in Tables 1 and 2) makes me concern about the computational feasibility of the proposed method. Maybe the authors can report the computing time, as this might give more insight into the proposal's computational feasibility.

***Reply**: Thanks again for the helpful suggestions. We agree with your points and thus condensed the tables into plots and moved some of the details to the Supplementary Materials, including the real data analysis. We added the simulation setup back to the main text. The low number of replicates in the simulation study has been set due to the $d = 750$ scenario. Calculating $\binom{750}{2}$ correlation pairs, plugging the corresponding correlation matrix into the glasso routine for each penalty parameter λ along a grid of length 30 for four different estimators is computationally expensive. We added a microbenchmark of the*

correlation matrix estimation to give an idea of the computational feasibility of the proposed and existing methods. When applied in practice, one typically only runs the procedure once for a given dataset.

Changed: Below, we provide a small microbenchmark study regarding the computational feasibility of the proposed and existing methods. We compare the computation time of the correlation matrix estimation for the `oracle` estimator [23], our proposed polyserial/polychoric method, and the `bridge` ensemble estimator [12]. The results are displayed in Table 1.

Method	Dimension	Time (s)
oracle	50	0.2947
poly	50	58.4005
bridge	50	181.6449
oracle	250	6.6614
poly	250	1487.8901
bridge	250	4761.2884
oracle	750	66.9916
poly	750	9102.1628
bridge	750	27789.9450

TABLE 1. *Microbenchmark of the correlation matrix estimation. Displayed are mean compute times of the correlation matrix of the general mixed simulation setup described in Section 4.1 for the* `oracle` *estimator, our proposed polyserial/polychoric method, and the* `bridge` *ensemble estimator.*

The general mixed data simulation setup we use here is detailed in Section 4.1. We report the mean computation time in seconds over 20 repetitions. The `oracle` estimator is the fastest, followed by the polyserial/polychoric method and the `bridge` ensemble estimator. Computing polyserial/polychoric correlations is about three times faster than the bridge function approach. In the $d = 750$ setting, our polyserial/polychoric method takes, on average, around 2.5 hours to compute the correlation matrix. The `bridge` ensemble estimator requires about 7.7 hours.

Electronic Journal of Statistics
Vol. 0 (0000) 1
ISSN: 1935-7524
DOI: 10.1214/154957804100000000

High-Dimensional Undirected Graphical Models for Arbitrary Mixed Data

Konstantin Göbler

Technical University of Munich,
Robert Bosch GmbH
e-mail: konstantin.goebler@tum.de

Anne Miloschewski

German Center for Neurodegenerative Diseases (DZNE)
Bonn, Germany
e-mail: anne.miloschwski@dzne.de

Mathias Drton

Munich Center for Machine Learning,
Technical University of Munich
e-mail: mathias.drton@tum.de

Sach Mukherjee

German Center for Neurodegenerative Diseases (DZNE)
Bonn, Germany,

University of Cambridge, MRC Biostatistics Unit
e-mail: sach.mukherjee@dzne.de

Abstract:
Graphical models are an important tool in exploring relationships between variables in complex, multivariate data. Methods for learning such graphical models are well-developed in the case where all variables are either continuous or discrete, including in high dimensions. However, in many applications, data span variables of different types (e.g., continuous, count, binary, ordinal, etc.), whose principled joint analysis is nontrivial. Latent Gaussian copula models, in which all variables are modeled as transformations of underlying jointly Gaussian variables, represent a useful approach. Recent advances have shown how the binary-continuous case can be tackled, but the general mixed variable type regime remains challenging. In this work, we make the simple but useful observation that classical ideas concerning polychoric and polyserial correlations can be leveraged in a latent Gaussian copula framework. Building on this observation, we propose a flexible and scalable methodology for data with variables of entirely general mixed type. We study the key properties of the approaches theoretically and empirically.

Keywords and phrases: Generalized correlation, high-dimensional statistics, latent Gaussian copula, mixed data, polychoric/polyserial correlation, undirected graphical models.

1. Introduction

Graphical models are widely used in the analysis of multivariate data, providing a convenient and interpretable way to study relationships among potentially large numbers of variables. They are key tools in modern statistics and machine learning and play an important role in diverse applications. Undirected graphical models are used in a wide range of settings, including, among others, systems biology, omics, deep phenotyping [see, e.g. 10, 13, 27] and as a component within other analyses, including two-sample testing, unsupervised learning, hidden Markov modeling, and more [examples include 41, 39, 36, 37, 32].

A significant portion of the literature on graphical models has concentrated on scenarios where either only continuous variables or only discrete variables are present. Regarding the former case, Gaussian graphical models have been extensively studied, including in the high-dimensional regime [see among others 25, 16, 2, 19, 45, 35, 6]. In such models, it is assumed that the observed random vector follows a multivariate Gaussian distribution, and the graph structure of the model is given by the zero pattern in the inverse covariance matrix. Generalizations for continuous, non-Gaussian data have also been studied [26, 22, 13]. In the latter case, discrete graphical models – related to Ising-type models in statistical physics – have also been extensively studied [see, e.g. 40, 34].

However, in many applications, it is common to encounter data that entail *mixed* variable types, i.e., where the data vector includes components of different types (e.g., continuous-Gaussian, continuous-non-Gaussian, count, binary, etc.). Such "column heterogeneity" (from the usual convention of samples in rows and variables in columns) is the rule rather than the exception. For instance, in statistical genetics, the construction of regulatory networks using expression profiling of genes may involve jointly analyzing gene expression levels alongside categorical phenotypes. Similarly, diagnostic data in many medical applications may contain continuous measurements such as blood pressure and discrete information about disease status or pain levels.

In analyzing such data, estimating a joint multivariate graphical model spanning the various variable types is often of interest. In practice, this is sometimes done using *ad hoc* pipelines and data transformations. However, in graphical modeling, since the model output is intended to be scientifically interpretable and involves statements about properties such as conditional independence between variables, the use of *ad hoc* workflows without an understanding of the resulting estimation properties is arguably problematic.

There have been three main lines of work that tackle high-dimensional graphical modeling for mixed data. The earliest approach is conditional Gaussian modeling of a mix of categorical and continuous data [20] as treated by Cheng et al. [8], Lee and Hastie [21]. A second approach is to employ neighborhood selection, which amounts to separate modeling of conditional distributions for each variable given all others [see, e.g. 7, 43, 38]. A third approach uses latent Gaussian models, with a key recent reference being the paper of Fan et al. [11], who proposed a latent Gaussian copula model for mixed data. The generative structure in their work posits that the discrete data is obtained from latent

continuous variables thresholded at certain (unknown) levels. However, in [11], only a mix of binary and continuous data is considered. Their setting does not allow for more general combinations (including counts or ordinal variables) as found in many real-world applications.

This third approach will be the focus of this paper, which aims to provide a simple framework for working with latent Gaussian copula models to analyze general mixed data. To do so, we combine classical ideas concerning polychoric and polyserial correlations with approaches from the high-dimensional graphical models and copula literature. As we discuss below, this provides an overall framework that is scalable, general, and straightforward from the user's point of view.

Already in the early 1900s, Pearson [30, 31] worked on the foundations of these ideas in the form of the tetrachoric and biserial correlation coefficients. From these arose the maximum likelihood estimators (MLEs) for the general version of these early ideas, namely the polychoric and the polyserial correlation coefficients. One drawback of these original measures is that they have been proposed in the context of latent Gaussian variables. A richer distributional family is the nonparanormal proposed by Liu, Lafferty and Wasserman [22] as a nonparametric extension to the Gaussian family. A random vector $\boldsymbol{X} \in \mathbb{R}^d$ is a member of the nonparanormal family when $f(\boldsymbol{X}) = (f_1(X_1), \ldots, f_d(X_d))^T$ is Gaussian, where $\{f_k\}_{k=1}^d$ is a set of univariate monotone transformation functions. Moreover, if the f_j's are monotone and differentiable, the nonparanormal family is equivalent to the Gaussian copula family. As the polychoric and polyserial correlation assumes that observed discrete data are generated from latent continuous variables, they adhere to a latent copula approach.

We propose two estimators of the latent correlation matrix, which can subsequently be plugged into existing precision matrix estimation routines, such as the graphical lasso (glasso) [16], CLIME [6], or the graphical Dantzig selector [45]. The first is appropriate under a latent Gaussian model and unifies the aforementioned MLEs. The second is more general and is applicable under the latent Gaussian copula model. Both approaches can deal with discrete variables with arbitrarily many levels. We that both estimators exhibit favorable theoretical properties and include empirical results based on real and simulated data. The main contributions of the paper are as follows:

- We posit that integrating polychoric and polyserial correlations into the latent Gaussian copula framework offers an elegant, straightforward, and highly effective approach to graphical modeling for comprehensively diverse mixed data sets.
- We present theoretical findings on the performance of the proposed estimators, encompassing their behavior in high-dimensional scenarios. The concentration results underscore the statistical validity of the introduced procedures.
- We empirically examine the estimators through a series of simulations and a practical example involving real phenotyping data of mixed types sourced from the UK Biobank. Our findings illustrate the practical utility of the

proposed methods, demonstrating that their performance often closely aligns with an oracle model granted access to true latent data.

Our proposed procedure provides users with a method for conducting statistically sound graphical modeling of mixed data that is both straightforward to implement and carries no more overhead than conventional high-dimensional Gaussian graphical modeling approaches. Our procedure requires no manual specification of variable-type-specific model components, such as bridge functions.

The remainder of this paper is organized as follows. In Sections 2 and 3, we present the estimators based on polychoric and polyserial correlations, including theoretical guarantees in terms of concentration inequalities. In Section 4, we describe the experimental setup used to test the proposed approaches on simulated data together with the results themselves. We conclude with a summary of our findings in Section 5 and point towards our `R` package **hume**, providing users with a convenient implementation of the methods developed in this study.

2. Background and model set-up

The objective of this paper is to learn the structure of undirected graphical models applicable to a wide range of mixed and high-dimensional data. To achieve this, we extend the Gaussian copula model [22, 23, 42], enabling the incorporation of both discrete and continuous data of any nature.

Definition 2.1 (The nonparanormal model)**.** *A random vector of continuous variables* $\mathbf{X} = (X_1, \ldots, X_d)$ *follows a* d*-dimensional nonparanormal distribution if there exists a set of monotone and differentiable univariate functions* $f = \{f_1, \ldots, f_d\}$ *such that the transformed vector* $f(\mathbf{X}) = (f_1(X)_1, \ldots, f_d(X)_d)$ *is multivariate Gaussian with mean* 0 *and covariance matrix* Σ*, i.e.* $f(\mathbf{X}) \sim N(0, \Sigma)$*. We write*

$$\mathbf{X} \sim NPN(0, \Sigma, f), \tag{1}$$

where without loss of generality, the diagonal entries in Σ *are equal to one.*

As demonstrated by Liu, Lafferty and Wasserman [22], the model in Eq. (1) is a semiparametric Gaussian copula model. The following definition indicates how to extend this model to the presence of general mixed data.

Definition 2.2 (latent Gaussian copula model for general mixed data)**.** *Let* $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ *be a* d*-dimensional random vector with* $\mathbf{X}_1$ *a* d_1*-dimensional vector of possibly ordered discrete variables, and* $\mathbf{X}_2$ *a* d_2*-dimensional vector of continuous variables with* $d = d_1 + d_2$*. Suppose there exists a* d_1*-dimensional random vector of latent continuous variables* $\mathbf{Z}_1 = (Z_1, \ldots, Z_{d_1})^T$ *such that the following relation holds:*

$$X_j = x_j^r \quad if \quad \gamma_j^{r-1} \leq Z_j < \gamma_j^r \quad for\ all\ j = 1, \ldots d_1 \ and\ r = 1, \ldots, l_j + 1, \tag{2}$$

where γ_j^r *represents some unknown thresholds with* $\gamma_j^0 = -\infty$ *and* $\gamma_j^{l_j+1} = +\infty$*,* $x_j^r \in \mathbb{N}_0$ *and* $l_j + 1$ *the number of discrete levels of* X_j *for all* $j \in 1, \ldots, d_1$*.*

Then, $\mathbf{X}$ *satisfies the latent Gaussian copula model if* $\mathbf{Z} \coloneqq (\mathbf{Z}_1, \mathbf{X}_2) \sim NPN(0, \mathbf{\Sigma}, f)$. *We write*

$$\mathbf{X} \sim LNPN(0, \mathbf{\Sigma}, f, \gamma), \tag{3}$$

where $\gamma = \cup_{j=1}^{d_1} \{\gamma_j^r, r = 0, \ldots, l_j + 1\}$.

Note that Definition 2.2 entails the class of Gaussian copula models if no discrete variables are present and the class of latent Gaussian models if $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{X}_2) \sim \mathrm{N}(0, \mathbf{\Sigma})$. As shown by Fan et al. [11], the latent Gaussian copula model (LGCM) is invariant concerning any re-ordering of the discrete variables.

We denote $[d] = \{1, \ldots, d\}$, $[d_1] = \{1, \ldots, d_1\}$, and $[d_2] = \{d_1 + 1, \ldots, d_2\}$, respectively. Several identifiability issues arise in the latent Gaussian copula class. First, the mean and the variances are not identifiable unless the monotone transformations f were restricted to preserve them. Note that this only affects the diagonal entries in $\mathbf{\Sigma}$, not the full covariance matrix. Therefore, without loss of generality, we assume the mean to be the zero vector and $\Sigma_{jj} = 1$ for all $j \in [d]$. Another identifiability issue relates to the unknown threshold parameters. To ease notation, let $\Gamma_j^r \equiv f_j(\gamma_j^r)$ and $\Gamma_j \equiv \{f_j(\gamma_j^r)\}_{r=0}^{l_j+1}$. In the LGCM, only the transformed thresholds Γ_j rather than the original thresholds are identifiable from the discrete variables. We assume, without loss of generality, that the transformed thresholds retain the limiting behavior of the original thresholds, i.e., $\Gamma_j^0 = -\infty$ and $\Gamma_j^{l_j+1} = \infty$.

Let $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ denote the latent precision matrix. Then, the zero-pattern of $\mathbf{\Omega}$ under the LGCM still encodes the conditional independencies of the latent continuous variables [22]. Thus, the underlying undirected graph is represented by $\mathbf{\Omega}$ just as for the parametric normal. Note that the LGCM for general mixed data in Definition 2.2 agrees with that of Quan, Booth and Wells [33] and of Feng and Ning [12]. The problem phrased by Fan et al. [11] is a special case of Definition 2.2. A more detailed comparison between both approaches can be found in Section 3. Nominal discrete variables need to be transformed into a dummy system.

For the remainder of the paper, assume we observe an independent n-sample of the d-dimensional vector $\mathbf{X}$ which is assumed to follow an LGCM of the form $\mathrm{LNPN}(0, \mathbf{\Sigma}, f, \Gamma)$, where $\Gamma = \cup_{j=1}^{d_1} \Gamma_j$. We estimate $\mathbf{\Sigma}$ by considering the corresponding entries separately i.e. the couples (X_j, X_k) for $j, k \in [d]$. Consequently, we have to keep in view three possible cases depending on the couple's variable types, respectively:

Case I: Both X_j and X_k are continuous, i.e. $j, k \in [d_2]$.
Case II: X_j is discrete and X_k is continuous, i.e. $j \in [d_1], k \in [d_2]$ and vice versa.
Case III: Both X_j and X_k are discrete, i.e. $j, k \in [d_1]$.

2.1. Maximum-likelihood estimation under the latent Gaussian model

At the outset, we examine each of the three cases under the latent Gaussian model, a special case of the LGCM where all transformations are identity functions. Consider *Case I*, where both X_j and X_k are continuous. This corresponds to the regular Gaussian graphical model set-up discussed thoroughly, for instance, in [35]. Hence, the estimator for $\mathbf{\Sigma}$ when both X_j and X_k are continuous is:

Definition 2.3 (MLE $\hat{\mathbf{\Sigma}}^{(n)}$ of $\mathbf{\Sigma}$; *Case I*)**.** *Let $\bar{x}_j$ denote the sample mean of X_j. The estimator $\hat{\mathbf{\Sigma}}^{(n)} = (\hat{\Sigma}^{(n)}_{jk})_{d_1<j<k\leq d_2}$ of the correlation matrix $\mathbf{\Sigma}$ is defined by:*

$$\hat{\Sigma}^{(n)}_{jk} = \frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}\sqrt{\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}} \tag{4}$$

for all $d_1 < j < k \leq d_2$.

This is the Pearson product-moment correlation coefficient, which, of course, coincides with the maximum likelihood estimator (MLE) for the bivariate normal couple $\{(X_j, X_k)\}_{i=1}^{n}$.

Turning to *Case II*, let X_j be ordinal and X_k be continuous. We are interested in the product-moment correlation Σ_{jk} between two jointly Gaussian variables, where X_j is not directly observed but only the ordered categories (see Eq. (2)). This is called the *polyserial* correlation [29]. The likelihood and log-likelihood of the n-sample are defined by:

$$\begin{aligned} L^{(n)}_{jk}(\Sigma_{jk}, x^r_j, x_k) &= \prod_{i=1}^{n} p(x^r_{ij}, x_{ik}, \Sigma_{jk}) = \prod_{i=1}^{n} p(x_{ik})p(x^r_{ij} \mid x_{ik}, \Sigma_{jk}) \\ \ell^{(n)}_{jk}(\Sigma_{jk}, x^r_j, x_k) &= \sum_{i=1}^{n} \big[\log(p(x_{ik})) + \log(p(x^r_{ij} \mid x_{ik}, \Sigma_{jk}))\big], \end{aligned} \tag{5}$$

where $p(x^r_{ij}, x_{ik}, \Sigma_{jk})$ denotes the joint probability of X_j and X_k and $p(x_{ik})$ the marginal density of the Gaussian variable X_k. MLEs are obtained by differentiating the log of the likelihood in Eq. (5) with respect to the unknown parameters, setting the partial derivatives to zero, and solving the system of equations for $\Sigma_{jk}, \mu, \sigma^2$, and Γ^r_j for $r \in [l_j]$. Under the latent Gaussian model, we have the special case that the thresholds are identifiable from the observed data as $\Gamma^r_j = \gamma^r_j$.

Definition 2.4 (MLE $\hat{\mathbf{\Sigma}}^{(n)}$ of $\mathbf{\Sigma}$; *Case II*)**.** *Recall the log-likelihood in Eq. (5). The estimator $\hat{\mathbf{\Sigma}}^{(n)} = (\hat{\Sigma}^{(n)}_{jk})_{1<j\leq d_1<k\leq d_2}$ is defined by:*

$$\begin{aligned} \hat{\Sigma}^{(n)}_{jk} &= \operatorname*{arg\,max}_{|\Sigma_{jk}|\leq 1} \ell^{(n)}_{jk}(\Sigma_{jk}, x^r_j, x_k) \\ &= \operatorname*{arg\,max}_{|\Sigma_{jk}|\leq 1} \frac{1}{n}\ell^{(n)}_{jk}(\Sigma_{jk}, x^r_j, x_k) \end{aligned} \tag{6}$$

for all $1 < j \leq d_1 < k \leq d_2$.

Regularity conditions ensuring consistency and asymptotic efficiency, as well as asymptotic normality, can be verified to hold here [9].

Lastly, consider *Case III*, where both X_j and X_k are ordinal. The probability of an observation with $X_j = x_j^r$ and $X_k = x_k^s$ is given by

$$\begin{aligned}\pi_{rs} &:= p(X_j = x_j^r, X_k = x_k^s) \\ &= p(\Gamma_j^{r-1} \leq Z_j < \Gamma_j^r, \Gamma_k^{s-1} \leq Z_k < \Gamma_k^s) \\ &= \int_{\Gamma_j^{r-1}}^{\Gamma_j^r} \int_{\Gamma_k^{s-1}}^{\Gamma_k^s} \phi(z_j, z_k, \Sigma_{jk}) dz_j dz_k,\end{aligned} \tag{7}$$

where $r = 1, \ldots, l_j$ and $s = 1, \ldots, l_k$ and $\phi(x, y, \rho)$ denotes the standard bivariate density with correlation ρ. Then, as outlined by Olsson [28] the likelihood and log-likelihood of the n-sample are defined as:

$$\begin{aligned}L_{jk}^{(n)}(\Sigma_{jk}, x_j^r, x_k^s) &= C \prod_{r=1}^{l_j} \prod_{s=1}^{l_k} \pi_{rs}^{n_{rs}}, \\ \ell_{jk}^{(n)}(\Sigma_{jk}, x_j^r, x_k^s) &= \log(C) + \sum_{r=1}^{l_j} \sum_{s=1}^{l_k} n_{rs} \log(\pi_{rs}),\end{aligned} \tag{8}$$

where C is a constant and n_{rs} denotes the observed frequency of $X_j = x_j^r$ and $X_k = x_k^s$ in a sample of size $n = \sum_{r=1}^{l_j} \sum_{s=1}^{l_k} n_{rs}$. Differentiating the log-likelihood, setting it to zero, and solving for the unknown parameters yields the estimator for Σ for *Case III*:

Definition 2.5 (MLE $\hat{\boldsymbol{\Sigma}}^{(n)}$ of $\boldsymbol{\Sigma}$; *Case III*)**.** *Recall the log-likelihood in Eq.* (8)*. The estimator* $\hat{\boldsymbol{\Sigma}}^{(n)} = (\hat{\Sigma}_{jk}^{(n)})_{1 \leq j < k \leq d_1}$ *of* $\boldsymbol{\Sigma}$ *is defined by:*

$$\begin{aligned}\hat{\Sigma}_{jk}^{(n)} &= \underset{|\Sigma_{jk}| \leq 1}{\arg\max}\, \ell_{jk}^{(n)}(\Sigma_{jk}, x_j^r, x_k^s) \\ &= \underset{|\Sigma_{jk}| \leq 1}{\arg\max}\, \frac{1}{n} \ell_{jk}^{(n)}(\Sigma_{jk}, x_j^r, x_k^s),\end{aligned} \tag{9}$$

for all $1 < j < k \leq d_1$.

Regularity conditions ensuring consistency and asymptotic efficiency, as well as asymptotic normality, can again be verified to hold here [18].

Summing up, under the latent Gaussian model, a special case of the LGCM, $\hat{\boldsymbol{\Sigma}}^{(n)}$ is a consistent and asymptotically efficient estimator for the underlying latent correlation matrix $\boldsymbol{\Sigma}$. Corresponding concentration results are derived in Section 3.5.

3. Latent Gaussian Copula Models

Fan et al. [11] propose the binary LGCM, a special case of the LGCM allowing for the presence of binary and continuous variables. Following the approach

of the nonparanormal SKEPTIC [23], they circumvent the direct estimation of monotone transformation functions $\{f_j\}_{j=1}^d$ by employing rank correlation measures, such as Kendall's tau or Spearman's rho. These measures remain invariant under monotone transformations. Notably, for *Case I*, a well-known mapping exists between Kendall's tau, Spearman's rho, and the underlying Pearson correlation coefficient Σ_{jk}. As a result, the primary contribution of Fan et al. [11] lies in deriving corresponding bridge functions for cases II and III. To reduce computational burden Yoon, Müller and Gaynanova [44] propose a hybrid multilinear interpolation and optimization scheme of the underlying latent correlation.

When considering the general mixed case, Fan et al. [11] advocate for binarizing all ordinal variables. This concept has been embraced by Feng and Ning [12], who suggest an initial step of binarizing all ordinal variables to create preliminary estimators. Subsequently, these estimators are meaningfully combined using a weighted aggregate. To extend the binary latent Gaussian copula model and explore generalizations regarding bridge functions, Quan, Booth and Wells [33] ventured into scenarios where a combination of continuous, binary, and ternary variables is present. However, a notable drawback of this approach becomes evident. Dealing with a mix of binary and continuous variables requires three bridge functions – one for each case. The complexity grows as discrete variables introduce distinct state spaces. In fact, a combination of continuous variables and discrete variables with k different state spaces necessitates $\binom{k+2}{2}$ bridge functions.

For this reason, we adopt an alternative approach to the latent Gaussian copula model when dealing with general mixed data, allowing discrete variables to possess any number of states. In this strategy, the number of cases to be considered remains consistent at three, as already introduced in the preceding section.

3.1. Nonparanormal Case I

For *Case I*, the mapping between Σ_{jk} and the population versions of Spearman's rho and Kendall's tau is well known [22]. Here we make use of Spearman's rho $\rho_{jk}^{Sp} = corr(F_j(X_j), F_k(X_k))$ with F_j and F_k denoting the cumulative distribution functions (CDFs) of X_j and X_k, respectively. Then $\Sigma_{jk} = 2\sin\frac{\pi}{6}\rho_{jk}^{Sp}$ for $d_1 < j < k \leq d_2$. In practice, we use the sample estimate

$$\hat{\rho}_{jk}^{Sp} = \frac{\sum_{i=1}^n (R_{ji} - \bar{R}_j)(R_{ki} - \bar{R}_k)}{\sqrt{\sum_{i=1}^n (R_{ji} - \bar{R}_j)^2 \sum_{i=1}^n (R_{ki} - \bar{R}_k)^2}},$$

with R_{ji} corresponding to the rank of X_{ji} among $X_{j1}, \ldots, X_{jn}$ and $\bar{R}_j = 1/n \sum_{i=1}^n R_{ji} = (n+1)/2$; compare [23]. From this, we obtain the following estimator:

Definition 3.1 (Nonparanormal estimator $\hat{\mathbf{\Sigma}}^{(n)}$ of $\mathbf{\Sigma}$; *Case I*)**.** *The estimator* $\hat{\mathbf{\Sigma}}^{(n)} = (\hat{\Sigma}^{(n)}_{jk})_{d_1<j<k\leq d_2}$ *of the correlation matrix* $\mathbf{\Sigma}$ *is defined by:*

$$\hat{\Sigma}^{(n)}_{jk} = 2\sin\frac{\pi}{6}\hat{\rho}^{Sp}_{jk}, \tag{10}$$

for all $d_1 < j < k \leq d_2$.

3.2. Nonparanormal Case II

In *Case II*, the complexity increases. Employing a rank-based approach for the nonparanormal model makes direct application of the ML procedure unfeasible, given that the continuous variable is not observed in its Gaussian form. Nevertheless, a two-step approach remains viable. First, an estimate of f_j must be formulated and subsequently employed in Definition 2.4. Yet, scrutinizing convergence rates for this procedure poses challenges, as the estimated transformation appears in multiple instances within the first-order condition of the MLE. Section 2 in the Supplementary Materials provides further details and compares the two-step likelihood approach to the one we propose below.

Instead, we will proceed by suitably modifying other approaches that address the Gaussian case through a more direct *ad hoc* examination of the relationship between Σ_{jk} and the point polyserial correlation [3, 4]. Section 2 of the Supplementary Materials compares the nonparanormal *Case II* estimation strategies.

In what follows, in the interest of readability, we omit the index in the monotone transformation functions but explicitly allow them to vary among the $\mathbf{Z}$. According to Definition 2.3, we have the following Gaussian conditional expectation

$$E[f(X_k) \mid f(Z_j)] = \mu_{f(X_k)} + \Sigma_{jk}\sigma_{f(X_k)}f(Z_j), \quad \text{for } 1 \leq j \leq d_1 < k \leq d_2, \tag{11}$$

where we can assume w.l.o.g. that $\mu_{f(X_k)} = 0$. After multiplying both sides with the discrete variable X_j, we move it into the expectation on the left-hand side of the equation. This is permissible as X_j is a function of $f(Z_j)$, i.e.

$$E[f(X_k)X_j \mid f(Z_j)] = \Sigma_{jk}\sigma_{f(X_k)}f(Z_j)X_j.$$

Now let us take again the expectation on both sides, rearrange and expand by σ_{X_j}, yielding

$$\Sigma_{jk} = \frac{E[f(X_k)X_j]}{\sigma_{f(X_k)}E[f(Z_j)X_j]} = \frac{r_{f(X_k)X_j}\sigma_{X_j}}{E[f(Z_j)X_j]}, \tag{12}$$

where $r_{f(X_k)X_j}$ is the product-moment correlation between the Gaussian (unobserved) variable $f(X_k)$ and the observed discretized variable X_j.

All that remains is to find sample versions of each of the three components in Eq. (12). Let us start with the expectation in the denominator $E[f(Z_j)X_j]$. By assumption $f(\mathbf{Z}) \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma})$ and therefore w.l.o.g. $f(Z_j) \sim \mathrm{N}(0, 1)$ for all

$j \in 1, \ldots, d_1$. Consequently, we have:

$$\begin{aligned} E[f(Z_j)X_j] &= \sum_{r=1}^{l_{j+1}} x_j^r \int_{\Gamma_j^{r-1}}^{\Gamma_j^r} f(z_j) dF(f(z_j)) = \sum_{r=1}^{l_{j+1}} x_j^r \int_{\Gamma_j^{r-1}}^{\Gamma_j^r} f(z_j)\phi(f(z_j))dz_j \\ &= \sum_{r=1}^{l_{j+1}} x_j^r \Big(\phi(\Gamma_j^r) - \phi(\Gamma_j^{r-1}) \Big) = \sum_{r=1}^{l_j} (x_j^{r+1} - x_j^r)\phi(\Gamma_j^r), \end{aligned} \tag{13}$$

where $\phi(t)$ denotes the standard normal density. Whenever the ordinal states are consecutive integers we have $\sum_{r=1}^{l_j}(x_j^{r+1} - x_j^r)\phi(\Gamma_j^r) = \sum_{r=1}^{l_j}\phi(\Gamma_j^r)$. Based on this derivation, it is straightforward to give an estimate of $E[f(Z_j)X_j]$ once estimates of the thresholds Γ_j have been formed (see Section 3.4 for more details). Let us turn to the numerator of Eq. (12). The standard deviation of X_j does not require any special treatment, and we simply use $\sigma_{X_j}^{(n)} = \sqrt{1/n \sum_{i=1}^n (X_{ij} - \bar{X}_j)^2}$ to be able to treat discrete variables with a general number of states. However, the product-moment correlation $r_{f(X_k),X_j}$ is inherently more challenging as it involves the (unobserved) transformed version of the continuous variables. Therefore, we proceed to estimate the transformation.

To this end, consider the marginal distribution function of X_k, namely

$$F_{X_k}(x) = P(X_k \leq x) = P(f(X_k) \leq f(x)) = \Phi(f(x)),$$

such that $f(x) = \Phi^{-1}(F_{X_k}(x))$. In this setting, Liu, Lafferty and Wasserman [22] propose to evaluate the quantile function of the standard normal at a Winsorized version of the empirical distribution function. This is necessary as the standard Gaussian quantile function $\Phi^{-1}(\cdot)$ diverges when evaluated at the boundaries of the $[0, 1]$ interval. More precisely, consider $\hat{f}(u) = \Phi^{-1}(W_{\delta_n}[\hat{F}_{X_k}(u)])$,where W_{δ_n} is a Winsorization operator, i.e.

$$W_{\delta_n}(u) \equiv \delta_n I(u < \delta_n) + uI(\delta_n \leq u \leq (1 - \delta_n)) + (1 - \delta_n)I(u > (1 - \delta_n)).$$

The truncation constant δ_n can be chosen in several ways. Liu, Lafferty and Wasserman [22] propose to use $\delta_n = 1/(4n^{1/4}\sqrt{\pi \log n})$ in order to control the bias-variance trade-off. Thus, equipped with an estimator for the transformation functions, the product-moment correlation is obtained the usual way, i.e.

$$r_{\hat{f}(X_k),X_j}^{(n)} = \frac{\sum_{i=1}^n (\hat{f}(X_{ik}) - \mu(\hat{f}))(X_{ij} - \mu(X_j)}{\sqrt{\sum_{i=1}^n \Big(\hat{f}(X_{ik}) - \mu(\hat{f})\Big)^2} \sqrt{\sum_{i=1}^n \Big(X_{ij} - \mu(X_j)\Big)^2}},$$

where $\mu(\hat{f}) \equiv 1/n \sum_{i=1}^n \hat{f}(X_{ik})$ and $\mu(X_j) \equiv 1/n \sum_{i=1}^n X_{ij}$. The resulting estimator is a double-two-step estimator of the mixed couple X_j and X_k.

Definition 3.2 (Estimator $\hat{\mathbf{\Sigma}}^{(n)}$ of $\mathbf{\Sigma}$; *Case II* nonparanormal)**.** *The estimator*

$\hat{\mathbf{\Sigma}}^{(n)} = (\hat{\Sigma}^{(n)}_{jk})_{1<j\leq d_1<k\leq d_2}$ *of the correlation matrix* $\mathbf{\Sigma}$ *is defined by:*

$$\hat{\Sigma}^{(n)}_{jk} = \frac{r^{(n)}_{\hat{f}(X_k),X_j}\,\sigma^{(n)}_{X_j}}{\sum_{r=1}^{l_j}\phi(\hat{\Gamma}^r_j)(x_j^{r+1}-x_j^r)} \tag{14}$$

for all $1 < j \leq d_1 < k \leq d_2$.

3.3. Nonparanormal Case III

Lastly, let us turn to *Case III* where both X_j and X_k are discrete, but they might differ in their respective state spaces. In the previous section, the ML procedure could no longer be applied directly because we do not observe the continuous variable in its Gaussian form. In *Case III*, however, we only observe the discrete variables generated by the latent scheme outlined in Definition 2.3. Due to the monotonicity of the transformation functions, the ML procedure for *Case III* from Section 2.1 can still be applied, i.e.

Definition 3.3 (Nonparanormal estimator $\hat{\mathbf{\Sigma}}^{(n)}$ of $\mathbf{\Sigma}$; *Case III*)**.** *The estimator* $\hat{\mathbf{\Sigma}}^{(n)} = (\hat{\Sigma}^{(n)}_{jk})_{1\leq j<k\leq d_1}$ *of the correlation matrix* $\mathbf{\Sigma}$ *is defined by:*

$$\hat{\Sigma}^{(n)}_{jk} = \underset{|\Sigma_{jk}|\leq 1}{\arg\max}\,\frac{1}{n}\ell^{(n)}_{jk}(\Sigma_{jk}, x_j^r, x_k^s) \tag{15}$$

for all $1 < j < k \leq d_1$.

In summary, the estimator $\hat{\mathbf{\Sigma}}^{(n)}$ under the latent Gaussian copula model is a simple but important tool for flexible mixed graph learning. By using ideas from polyserial and polychoric correlation measures, we not only have an easy-to-calculate estimator but also overcome the issue of finding bridge functions between all different kinds of discrete variables.

3.4. Threshold estimation

The unknown threshold parameters Γ_j for $j \in [d_1]$ play a key role in linking the observed discrete to the latent continuous variables. Therefore, being able to form accurate estimates of the Γ_j is crucial for both the likelihood-based procedures and the nonparanormal estimators outlined above.

We start by highlighting that we set the LGCM model up such that for each Γ_j, there exists a constant G such that $|\Gamma^r_j| \leq G$ for all $r \in [l_j]$, i.e., the estimable thresholds are bounded away from infinity. Let us define the cumulative probability vector $\pi_j = (\pi_j^1, \ldots, \pi_j^{l_j})$. Then, by Eq. (2), it is easy to see that

$$\begin{aligned}\pi_j^r &= \sum_{i=1}^{r} P(X_j = x_j^i) = P(X_j \leq x_j^r)\\ &= P(Z_j \leq \gamma_j^r) = P(f_j(Z_j) \leq f_j(\gamma_j^r)) = \Phi(\Gamma_j^r).\end{aligned} \tag{16}$$

From this equation, it is immediately clear that the thresholds satisfy $\Gamma_j^r = \Phi^{-1}(\pi_j^r)$. Consequently, when forming sample estimates of the unknown thresholds, we replace the cumulative probability vector with its sample equivalent, namely

$$\hat{\pi}_j^r = \sum_{k=1}^{r} \Big[\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_{ij} = x_j^k) \Big] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_{ij} \leq x_j^r), \tag{17}$$

and plug it into the identity, i.e. $\hat{\Gamma}_j^r = \Phi^{-1}\big(\hat{\pi}_j^r\big)$ for $j \in [d_1]$. The following lemma assures that these threshold estimates can be formed with high accuracy.

Lemma 3.1. *Suppose the estimated thresholds are bounded away from infinity, i.e.,* $|\hat{\Gamma}_j^r| \leq G$ *for all* $j \in [d_1]$ *and* $r = 1, \ldots, l_j$ *and some* G*. The following bound holds for all* $t > 0$ *with Lipschitz constant* $L_1 = 1/(\sqrt{\frac{2}{\pi}} \min\{\hat{\pi}_j^r, 1 - \hat{\pi}_j^r\})$*:*

$$P\Big(|\hat{\Gamma}_j^r - \Gamma_j^r| \geq t\Big) \leq 2 \exp\Big(- \frac{2t^2 n}{L_1^2} \Big).$$

The proof of Lemma 3.1 is given in Section 5 of the Supplementary Materials. The requirement that the estimated thresholds are bounded away from infinity typically does not pose any restriction in finite samples. All herein-developed methods are applied in a two-step fashion. In the ensuing theoretical results, we stress this by denoting the estimated thresholds as $\bar{\Gamma}_j^r$.

3.5. Concentration results

Define $\mathbf{\Sigma}^*$ and $\mathbf{\Omega}^*$ as the true covariance matrix and its inverse, respectively. We start by stating the following assumptions:

Assumption 3.1. *For all* $1 \leq j < k \leq d$, $|\Sigma_{jk}^*| \neq 1$*. In other words, there exists a constant* $\delta > 0$ *such that* $|\Sigma_{jk}^*| \leq 1 - \delta$*.*

Assumption 3.2. *For any* Γ_j^r *with* $j \in [d_1]$ *and* $r \in [l_j]$ *there exists a constant* G *such that* $|\Gamma_j^r| \leq G$*.*

Assumption 3.3. *Let* $j < k$ *and consider the log-likelihood functions in Definition 2.4 and in Definition 2.5. We assume that with probability one,*

- $\{-1+\delta, 1-\delta\}$ *are not critical points of the respective log-likelihood functions.*
- *The log-likelihood functions have a finite number of critical points.*
- *Every critical point that is different from* Σ_{jk}^* *is non-degenerate.*
- *All joint and conditional states of the discrete variables have positive probability.*

Assumptions 3.1 and 3.2 ensure that $f(X_j)$ and $f(X_k)$ are not perfectly linearly dependent and that the thresholds are bounded away from infinity, respectively. Importantly, these constraints impose minimal restrictions in practice. Assumption 3.3 guarantees that the likelihood functions in Section 2.2 exhibit a "nice" behavior, representing a mild technical requirement.

Convergence results for latent Gaussian models The subsequent theorem, drawing on Mei, Bai and Montanari [24], hinges on four conditions, all substantiated in Section 3 of the Supplementary Materials. This concentration result specifically pertains to the MLEs introduced in Section 2.1 within the framework of the latent Gaussian model. We remark that related methodology has been applied by Anne, Aurélie and Clémence [1] in addressing zero-inflated Gaussian data under double truncation.

Theorem 3.2. *Suppose that Assumptions 3.1–3.3 hold, and let* $j \in [d_1]$ *and* $k \in [d_2]$ *for Case II and* $j, k \in [d_1]$ *for Case III. Let* $\alpha \in (0,1)$*, and let* $n \geq 4C\log(n)\log\left(\frac{B}{\alpha}\right)$ *with some known constants* B*,* C*, and* D *depending on cases II and III but independent of* (n, d)*. Then, it holds that*

$$P\left(\max_{j,k}\left|\hat{\Sigma}^{(n)}_{jk} - \Sigma^*_{jk}\right| \geq D\sqrt{\frac{\log(n)}{n}\log\left(\frac{B}{\alpha}\right)}\right) \leq \frac{d(d-1)}{2}\alpha. \tag{18}$$

Case I of the latent Gaussian model addresses the well-understood scenario involving observed Gaussian variables, with concentration results and rates of convergence readily available -see, for example, Lemma 1 in Ravikumar et al. [35]. Consequently, the MLEs converge to $\mathbf{\Sigma}^*$ at the optimal rate of $n^{-1/2}$, mirroring the convergence rate as if the underlying latent variables were directly observed.

Convergence of nonparanormal estimators. Recall the three cases, which, in principle, will have to be considered again.

Case I: When both random variables are continuous, concentration results follow immediately from Liu et al. [23] who make use of Hoeffding's inequalities for U-statistics.

Case II: For the case where one variable is discrete and the other one continuous, we present concentration results below.

Case III: When both variables are discrete, we make an important observation that Theorem 3.2 above still applies and needs not to be altered. We do not observe the continuous variables directly but only their discretized versions. Consequently, the threshold estimates remain valid under the monotone transformation functions, and so does the polychoric correlation.

The following theorem provides concentration properties for *Case II* under the LGCM.

Theorem 3.3. *Suppose that Assumptions 3.1 and 3.2 hold and* $j \in [d_1]$ *and* $k \in [d_2]$*. Then for any* $\epsilon \in \left[C_M\sqrt{\frac{\log d \log^2 n}{\sqrt{n}}}, 8(1+4c^2)\right]$*, with sub-Gaussian parameter* c*, generic constants* $k_i, i = 1,2,3$ *and constant* $C_M = \frac{48}{\sqrt{\pi}}\left(\sqrt{2M}-1\right)(M+2)$ *for some* $M \geq 2\left(\frac{\log d_2}{\log n}+1\right)$ *with* $C_\Gamma = \sum_{r=1}^{l_j}\phi(\bar{\Gamma}^r_j)(x^{r+1}_j - x^r_j)$ *and Lipschitz*

constant L the following probability bound holds

$$
\begin{aligned}
&P\left(\max_{jk}\left|\hat{\Sigma}_{jk}^{(n)}-\Sigma_{jk}^{*}\right|\geq\epsilon\right)\\
&\quad\leq 8\exp\left(2\log d-\frac{\sqrt{n}\epsilon^{2}}{(64\ L\ C_{\gamma}\ l_{\max}\ \pi)^{2}\log n}\right)\\
&\quad+8\exp\left(2\log d-\frac{n\epsilon^{2}}{(4L\ C_{\gamma})^{2}\ 128(1+4c^{2})^{2}}\right)\\
&\quad+8\exp\left(2\log d-\frac{\sqrt{n}}{8\pi\log n}\right)+4\exp\left(-\frac{k_{1}n^{3/4}\sqrt{\log n}}{k_{2}+k_{3}}\right)+\frac{2}{\sqrt{\pi\log(nd_{2})}}.
\end{aligned}
$$

The proof of the theorem is given in Section 6 of the Supplementary Materials. The first four terms in the probability bound stem from finding bounds to different regions of the support of the transformed continuous variable. The last term is a consequence of the fact that we estimate the transform directly.

Regarding the scaling of the dimension in terms of sample size, the ensuing corollary follows immediately.

Corollary 3.4. *For some known constant K_{Σ} independent of d and n we have*

$$
P\left(\max_{j,k}\left|\hat{\Sigma}_{jk}^{(n)}-\Sigma_{jk}^{*}\right|>K_{\Sigma}\sqrt{\frac{\log d\log n}{\sqrt{n}}}\right)=o(1). \tag{19}
$$

The nonparanormal estimator for *Case II* converges to Σ_{jk}^{*} at rate $n^{-1/4}$, which is slower than the optimal parametric rate of $n^{-1/2}$. This stems not from the presence of the discrete variable but from the direct estimation of the transformation function f_j and the corresponding truncation constant δ_n. Both Xue and Zou [42] and Liu et al. [23] discuss room for improvement of the estimator for f_j to get a rate closer to the optimal one. Improvements depend strongly on the choice of truncation constant, striking a bias-variance balance in high-dimensional settings. In all of the numerical experiments below, we find that Theorem 3.3 gives a worst-case rate that does not appear to negatively impact performance compared to estimators that attain the optimal rate.

3.6. Estimating the precision matrix

Similar to Fan et al. [11], we plug our estimate of the sample correlation matrix into existing routines for estimating $\mathbf{\Omega}^{*}$. In particular, we employ the graphical lasso (glasso) estimator [16], i.e.

$$
\hat{\mathbf{\Omega}}=\underset{\mathbf{\Omega}\succeq 0}{\arg\min}\left[\operatorname{tr}(\hat{\mathbf{\Sigma}}^{(n)}\mathbf{\Omega})-\log|\mathbf{\Omega}|+\lambda\sum_{j\neq k}|\Omega_{jk}|\right], \tag{20}
$$

where $\lambda>0$ is a regularization parameter. As $\hat{\mathbf{\Sigma}}^{(n)}$ exhibits at worst the same theoretical properties as established in Liu, Lafferty and Wasserman [22], convergence rate and graph selection results follow immediately.

We do not penalize diagonal entries of $\mathbf{\Omega}$ and therefore have to make sure that $\hat{\mathbf{\Sigma}}^{(n)}$ is at least positive semidefinite to establish convergence in Eq. (20). Hence, we need to project $\hat{\mathbf{\Sigma}}^{(n)}$ into the cone of positive semidefinite matrices; see also [23, 11]. In practice, we use an efficient implementation of the alternating projections method proposed by Higham [17].

To select the tuning parameter in Eq. (20) Foygel and Drton [15] introduce an extended BIC (eBIC) in particular for Gaussian graphical models establishing consistency in higher dimensions under mild asymptotic assumptions. We consider

$$eBIC_\theta = -2\ell^{(n)}(\hat{\mathbf{\Omega}}(E)) + |E|\log(n) + 4|E|\theta\log(d), \tag{21}$$

where $\theta \in [0, 1]$ governs penalization of large graphs. Furthermore, $|E|$ represents the cardinality of the edge set of a candidate graph on d nodes and $\ell^{(n)}(\hat{\mathbf{\Omega}}(E))$ denotes the corresponding maximized log-likelihood which in turn depends on λ from Eq. (20). In practice, first, one retrieves a small set of models over a range of penalty parameters $\lambda > 0$ (called *glasso path*). Then, we calculate the eBIC for each model in the path and select the one with the minimal value.

4. Numerical results

To numerically assess the accuracy of our mixed graph estimation approach, we commence with a simulation study in which the estimators are rigorously evaluated in a gold-standard fashion and compared against oracles.

4.1. Simulation setup

We start by constructing the underlying precision matrix $\mathbf{\Omega}^*$ whose zero pattern encodes the undirected graph. We set $\mathbf{\Omega}^*_{jj} = 1$ and $\mathbf{\Omega}^*_{jk} = s \cdot b_{jk}$ if $j \neq k$, where s is a constant signal strength chosen to assure positive definiteness. Furthermore, b_{jk} are realizations of a Bernoulli random variable with corresponding success probability $p_{jk} = (2\pi)^{-1/2} \exp\left[\|v_j - v_k\|_2/(2c)\right]$. In particular, $v_j = (v_j^{(1)}, v_j^{(2)})$ are independent realizations of a bivariate uniform $[0, 1]$ distribution and c controls the sparsity of the graph.

Throughout the simulation, we set $s = 0.15$ and incrementally increase the dimensionality s.t. $d \in \{50, 250, 750\}$, representing a transition from small to large-scale graphs. We let $\mathbf{\Sigma}^* = (\mathbf{\Omega}^*)^{-1}$ be rescaled such that all diagonal elements are equal to 1. Given $\mathbf{\Sigma}^*$, we first obtain the partially latent continuous data $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{X}_2)$ where $\mathbf{Z} \sim \mathrm{NPN}_d(\mathbf{0}, \mathbf{\Sigma}^*, f)$. In practice, we draw n i.i.d. samples from $\mathrm{N}_d(\mathbf{0}, \mathbf{\Sigma}^*)$ and apply the back-transform f^{-1} to each individual variable.

To generate general mixed data $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ according to the LGCM we need to appropriately threshold $\mathbf{Z}_1$. Let $\mathbf{X}_1$ be partitioned into equally sized collections of binary, ordinal, and Poisson distributed random variables, i.e., $\mathbf{X}_1 = (\mathbf{X}_1^{\mathrm{bin}}, \mathbf{X}_1^{\mathrm{ord}}, \mathbf{X}_1^{\mathrm{pois}})$. We use the inverse probability integral transform (IPT) to generate random samples from the respective cumulative distribution

functions, corresponding to the relationship described in Eq. (2). For $\mathbf{X}_1^{\text{bin}}$ IPT is employed with success probability drawn from Uniform$[0.4, 0.6]$ for 80% of $\mathbf{X}_1^{\text{bin}}$. We assign unbalanced classes to the remaining 20%, where the success probability is drawn from Uniform$[0.05, 0.1]$. Regarding $\mathbf{X}_1^{\text{ord}}$, IPT is used to generate samples from the multinomial distribution. To that end, we draw the number of categories from Uniform$[3, 7]$ and round it to the nearest integer. We set the probability of falling into one of these categories to be proportional to their number. Lastly, $\mathbf{X}_1^{\text{pois}}$ is generated using IPT with the rate parameter set to 6. In case we only need a mix of binary and continuous data, we set $\mathbf{X}_1 = \mathbf{X}_1^{\text{bin}}$.

Throughout the experiments, $\hat{\Omega}$ is chosen by minimizing the eBIC according to the procedure outlined in Section 3.6 with $\theta = 0.1$ for the low and medium, $\theta = 0.5$ for the high dimensional graphs. The sample size n is set to 200 for $d \in \{50, 250\}$ and 300 for $d = 750$. We set the number of simulation runs to 100. Lastly, we choose the sparsity parameter c such that the number of edges aligns roughly with the dimension – except for $d = 50$, where we allow for 200 edges following Fan et al. [11].

Performance metrics. To evaluate performance, we report the mean estimation error $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_F$ using the Frobenius norm. Additionally, we employ graph recovery metrics. For this purpose, we calculate the number of true positives $\text{TP}(\lambda)$ and false positives $\text{FP}(\lambda)$ based on the *glasso path*. $\text{TP}(\lambda)$ represents the count of non-zero lower off-diagonal elements that are consistent both in $\mathbf{\Omega}^*$ and $\hat{\mathbf{\Omega}}$, while $\text{FP}(\lambda)$ denotes the count of non-zero lower off-diagonal elements in $\hat{\mathbf{\Omega}}$ that are zero in $\mathbf{\Omega}^*$.

The true positive rate $\text{TPR}(\lambda)$ and false positive rate $\text{FPR}(\lambda)$ are defined as $\text{TPR} = \frac{\text{TP}(\lambda)}{|E|}$ and $\text{FPR} = \frac{\text{FP}(\lambda)}{d(d-1)/2-|E|}$, respectively. Finally, we consider the area under the curve (AUC), where a value of 0.5 corresponds to random guessing of edge presence and a value of 1 indicates perfect error-free recovery of the underlying latent graph (in the rank sense of ROC analysis).

4.2. Simulation results

Binary-continuous data We start by considering a mix of binary and continuous variables generated as outlined in Section 4.1 to compare our methods against the bridge function approach of Fan et al. [11]. For this purpose, Figure 3 depicts the mean estimation error $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_F$ and the AUC for the different estimators under the different (d, n) regimes. We include the following estimators for $\mathbf{\Omega}^*$: (1) An oracle estimator (`oracle`) that corresponds to estimating $\hat{\mathbf{\Sigma}}^{(n)}$ using the mapping between Spearman's rho and $\hat{\Sigma}^*_{jk}$ (Eq. (10)) based on realization of the (partially) latent continuous data $(\mathbf{Z_1}, \mathbf{X_2})$. (2) The bridge function based estimator (`bridge`) proposed by Fan et al. [11]. (3) The polychhoric and polyserial MLE estimator (`mle`) proposed in Section 2.1. (4) The general mixed estimator (`poly`) proposed in Section 3. In the left column, we set $f_j(x) = x$ for all j, i.e., we recover the latent Gaussian model. In the right column, we set $f_j(x) = x^{1/3}$ for all j to recover the LGCM.

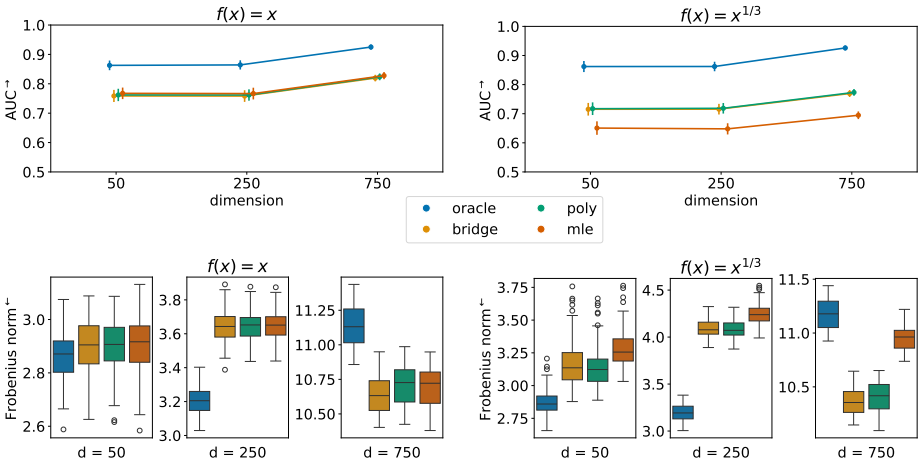

Fig 3. Simulation results for the binary-continuous data setting based on 100 *simulation runs. The left column corresponds to the latent Gaussian model, where the transformation function is the identity. The right colum depicts results for the LGCM with* $f_j(x) = x^{1/3}$ *for all* j*. The top row reports mean and standard deviation of the AUC along simulation runs, and the bottom row depicts boxplots of the estimation error* $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_F$*. The y-axis labels have superscript arrows attached to indicate the direction of improvement:* $\rightarrow$ *implies that larger values are better, and* $\leftarrow$ *implies that smaller values are better.*

Figure 3 suggests that under the latent Gaussian model (left column), there are virtually no differences between all non-oracle estimators in graph recovery or estimation error. As expected, the `oracle` has the highest AUC and lowest estimation error across scenarios. The only exception is the estimation error when the dimension is $d = 750$. This surprising result stems from an increased FPR for `oracle` (see Figure 3 in the Supplementary Materials) when minimizing the eBIC with the additional penalty set to $\theta = 0.5$. A higher penalty seems appropriate in this case. Meanwhile, the non-oracle estimators are more conservative, and the additional penalty appears to be chosen correctly in these cases.

In the right column of Figure 3, binary-continuous mixed data is generated from the LGCM. The *Case I* and *Case II* MLEs are misspecified in this case, which translates to lower AUC and higher estimation error. The remaining estimators are unaffected by the transformation. Encouragingly, we find no substantial performance difference between the bridge function approach and our procedure in any of the metrics considered, including the TPR and FPR results in Figure 3 in the Supplementary Materials.

General mixed data Let us turn to the general mixed setting. While the bridge function approach by Fan et al. [11] does not extend beyond the binary-continuous mix, we can still compare our approach to the ensemble method developed by Feng and Ning [12]. Due to its close connection to the original method, we continue to denote the proposed ensemble estimator `bridge`.

Similar to above, Figure 4 depicts the mean estimation error $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_F$ and the AUC and left and right columns correspond to latent Gaussian und LGCM

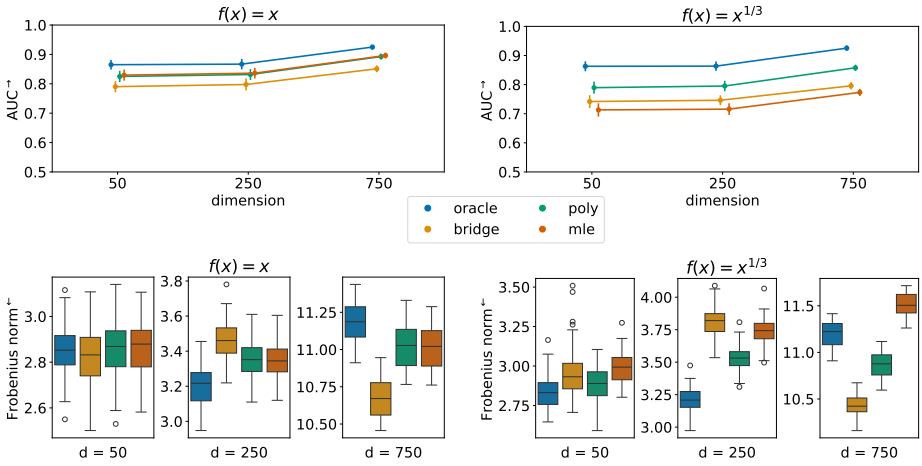

FIG 4. Simulation results for the general mixed data setting based on 100 *simulation runs. The left column corresponds to the latent Gaussian model, where the transformation function is the identity. The right colum depicts results for the LGCM with* $f_j(x) = x^{1/3}$ *for all* j*. The top row reports mean and standard deviation of the AUC along simulation runs, and the bottom row depicts boxplots of the estimation error* $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_F$*. The y-axis labels have superscript arrows attached to indicate the direction of improvement:* $\rightarrow$ *implies that larger values are better, and* $\leftarrow$ *implies that smaller values are better.*

settings, respectively. This time, given general mixed data, differences in terms of AUC between the estimators are noticeable. When the transformations are the identity, the MLE is correctly specified, and it is tied with `poly` in terms of AUC. The `bridge` estimator performs worse than the other two estimators. When the transformations are $f_j(x) = x^{1/3}$, the MLE is misspecified, and our `poly` estimator performs best among non-oracle estimators in terms of AUC. The `bridge` estimator performs only marginally better than the misspecified `mle`.

Turning to estimation error results, when $f_j(x) = x$, the `poly` and `mle` estimators perform similarly across dimensions. As the `oracle` estimator is formed on the latent continuous data, it is unaffected by any discretization and behaves the same as in the binary-continuous case above. The `bridge` ensemble estimator accounts for a slightly higher estimation error when $d = 250$ and a lower one when $d = 750$. This pattern can be explained by the FPR of the `bridge` estimator as illustrated in Figure 4 in the Supplementary Materials. While the FPR is slightly higher in the $d = 250$ case, it is lower in the $d = 750$ case. Similar to the `oracle` results, this appears to be a consequence of the additional penalty term in the eBIC. Considering the estimation error when $f_j(x) = x^{1/3}$, the `poly` and `bridge` estimators retain their performance. As before, the `mle` estimator is misspecified and performs worse than the other two estimators.

Overall, the simulation results suggest that our proposed `poly` estimator performs similarly (binary-continuous data setting) or better (general mixed setting) than the current state-of-the-art. In particular, the `poly` estimator

achieves good performance scores regarding recovery of the graph structure in the general mixed setting. Estimation error results are more sensitive to the choice of the additional high-dimensional penalty. These empirical results suggest that despite the theoretically slower convergence rate, the `poly` estimator is competitive regarding graph recovery and estimation error.

5. Conclusion

Estimating high-dimensional undirected graphs from general mixed data is a challenging task. We propose an innovative approach that blends classical generalized correlation measures, specifically polychoric and polyserial correlations, with recent concepts from high-dimensional graphical modeling and copulas.

A pivotal insight guiding our approach is recognizing that polychoric and polyserial correlations can be effectively modeled through a latent Gaussian copula. Although adapting polyserial correlation to the nonparanormal case demands careful consideration, the polychoric correlation requires no adjustments. The resulting estimators exhibit favorable theoretical properties, even in high dimensions, and demonstrate robust empirical performance in our simulation study.

Our advocated framework builds on prior work extending the graphical lasso for Gaussian observations to nonparanormal models and subsequently to mixed data, as seen in the contributions of Fan et al. [11], followed by Quan, Booth and Wells [33] and Feng and Ning [12]. A key distinction in our approach is the absence of the need to specify bridge functions. Indeed, our method seamlessly handles various types of mixed data without requiring additional user effort.

6. Software

Software in the form of the `R` package **hume** is available on the corresponding author's GitHub page (`https://github.com/konstantingoe/hume`). The `R`-code to reproduce the simulation study conducted in the paper is available under `https://github.com/konstantingoe/mixed_hidim_graphs`.

Supplementary Materials

We refer the reader to the Supplementary Materials for technical appendices, including proofs of the theorems and lemmas in the main manuscript. Additionally, we present further simulation results and an analysis of real phenotyping data concerning severe COVID-19 outcomes from the UK Biobank.

Funding

This work was supported by the Helmholtz AI project "Scalable and Interpretable Models for Complex And Structured Data" (SIMCARD), the Medical Research

Council [programme number MC UU 00002/17] and the National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre.

This project also received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme [grant agreement No 883818].

Acknowledgments

We thank the anonymous reviewers whose insightful comments and constructive feedback significantly enhanced the quality and clarity of this paper. We further want to thank Hongjian Shi for his helpful insights on the subject.

Conflict of Interest: None declared.

References

[1] Anne, G.-P., Aurélie, G.-M. and Clémence, K. (2019). Graph estimation for Gaussian data zero-inflated by double truncation. arXiv:1911.07694.
[2] 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. MR2417243
[3] Bedrick, E. J. (1992). A comparison of generalized and modified sample biserial correlation estimators. *Psychometrika* **57** 183–201. MR1173589
[4] Bedrick, E. J. and Breslin, F. C. (1996). Estimating the polyserial correlation coefficient. *Psychometrika* **61** 427–443. MR1424910
[5] Broyden, C. G. (1965). A class of methods for solving nonlinear simultaneous equations. *Math. Comp.* **19** 577–593. MR198670
[6] Cai, T., Liu, W. and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973
[7] Chen, S., Witten, D. M. and Shojaie, A. (2015). Selection and estimation for mixed graphical models. *Biometrika* **102** 47–64. MR3335095
[8] Cheng, J., Li, T., Levina, E. and Zhu, J. (2017). High-dimensional mixed graphical models. *J. Comput. Graph. Statist.* **26** 367–378. MR3640193
[9] Cox, N. R. (1974). Estimation of the correlation between a continuous and a discrete variable. *Biometrics* **30** 171–178. MR334376
[10] Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G. and West, M. (2004). Sparse graphical models for exploring gene expression data. *J. Multivariate Anal.* **90** 196–212. MR2064941
[11] Fan, J., Liu, H., Ning, Y. and Zou, H. (2017). High dimensional semiparametric latent graphical model for mixed data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 405–421. MR3611752

[12] Feng, H. and Ning, Y. (2019). High-dimensional Mixed Graphical Model with Ordinal Data: Parameter Estimation and Statistical Inference. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics* (K. Chaudhuri and M. Sugiyama, eds.). *Proceedings of Machine Learning Research* **89** 654–663. PMLR.

[13] Finegold, M. and Drton, M. (2011). Robust graphical modeling of gene networks using classical and alternative t-distributions. *Ann. Appl. Stat.* **5** 1057–1080. MR2840186

[14] Fox, J. (2022). polycor: Polychoric and Polyserial Correlations R package version 0.8-1.

[15] Foygel, R. and Drton, M. (2010). Extended Bayesian Information Criteria for Gaussian Graphical Models. In *Advances in Neural Information Processing Systems* (J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel and A. Culotta, eds.) **23** 604–612. Curran Associates, Inc.

[16] Friedman, J., Hastie, T. and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

[17] Higham, N. J. (1988). Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra Appl.* **103** 103–118. MR943997

[18] Jin, S. and Yang-Wallentin, F. (2017). Asymptotic robustness study of the polychoric correlation estimation. *Psychometrika* **82** 67–85. MR3614808

[19] Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. MR2572459

[20] Lauritzen, S. L. (1996). *Graphical models. Oxford Statistical Science Series* **17**. The Clarendon Press, Oxford University Press, New York Oxford Science Publications. MR1419991

[21] Lee, J. D. and Hastie, T. J. (2015). Learning the structure of mixed graphical models. *J. Comput. Graph. Statist.* **24** 230–253. MR3328255

[22] Liu, H., Lafferty, J. and Wasserman, L. (2009). The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* **10** 2295–2328. MR2563983

[23] Liu, H., Han, F., Yuan, M., Lafferty, J. and Wasserman, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.* **40** 2293–2326. MR3059084

[24] Mei, S., Bai, Y. and Montanari, A. (2018). The landscape of empirical risk for nonconvex losses. *Ann. Statist.* **46** 2747–2774. MR3851754

[25] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

[26] Miyamura, M. and Kano, Y. (2006). Robust Gaussian graphical modeling. *J. Multivariate Anal.* **97** 1525–1550. MR2275418

[27] Monti, R. P., Hellyer, P., Sharp, D., Leech, R., Anagnostopoulos, C. and Montana, G. (2014). Estimating time-varying brain connectivity networks from functional MRI time series. *NeuroImage* **103** 427–443.

[28] Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika* **44** 443–460. MR554892

[29] Olsson, U., Drasgow, F. and Dorans, N. J. (1982). The polyserial correlation coefficient. *Psychometrika* **47** 337–347. MR678066

[30] Pearson, K. (1900). I. Mathematical contributions to the theory of evolution.—VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* **195** 1–47.
[31] Pearson, K. (1913). On the measurement of the influence of "broad categories" on correlation. *Biometrika* **9** 116–139.
[32] Perrakis, K., Lartigue, T., Dondelinger, F. and Mukherjee, S. (2019). Regularized joint mixture models. arXiv:1908.07869.
[33] Quan, X., Booth, J. G. and Wells, M. T. (2018). Rank-based approach for estimating correlations in mixed ordinal data. arXiv: 1809.06255.
[34] Ravikumar, P., Wainwright, M. J. and Lafferty, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.* **38** 1287–1319. MR2662343
[35] Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980. MR2836766
[36] Städler, N. and Mukherjee, S. (2013). Penalized estimation in high-dimensional hidden Markov models with state-specific graphical models. *Ann. Appl. Stat.* **7** 2157–2179. MR3161717
[37] Städler, N. and Mukherjee, S. (2015). Multivariate gene-set testing based on graphical models. *Biostatistics* **16** 47–59. MR3365410
[38] Yang, Z., Ning, Y. and Liu, H. (2018). On semiparametric exponential family graphical models. *J. Mach. Learn. Res.* **19** Paper No. 57, 59. MR3899759
[39] Verzelen, N. and Villers, F. (2009). Tests for Gaussian graphical models. *Comput. Statist. Data Anal.* **53** 1894–1905. MR2649554
[40] Wainwright, M. J. and Jordan, M. I. (2006). Log-determinant relaxation for approximate inference in discrete Markov random fields. *IEEE Transactions on Signal Processing* **54** 2099–2109.
[41] Wei, Z. and Li, H. (2007). A Markov random field model for network-based analysis of genomic data. *Bioinformatics* **23** 1537–1544.
[42] Xue, L. and Zou, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann. Statist.* **40** 2541–2571. MR3097612
[43] Yang, E., Baker, Y., Ravikumar, P., Allen, G. and Liu, Z. (2014). Mixed Graphical Models via Exponential Families. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics* (S. Kaski and J. Corander, eds.). *Proceedings of Machine Learning Research* **33** 1042–1050. PMLR, Reykjavik, Iceland.
[44] Yoon, G., Müller, C. L. and Gaynanova, I. (2021). Fast Computation of Latent Correlations. *Journal of Computational and Graphical Statistics* **30** 1249-1256.
[45] Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **11** 2261–2286. MR2719856