Electronic Journal of Statistics
ISSN: 1935-7524

High-Dimensional Undirected Graphical Models for Arbitrary Mixed Data

Konstantin Göbler

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

Anne Miloschewski

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

arXiv: 2010.00000

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. 11, 14, 29] and as a component within other analyses, including two-sample testing, unsupervised learning, hidden Markov modeling, and more [examples include 44, 42, 38, 39, 34].

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 27, 17, 2, 21, 49, 37, 7]. 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 [28, 24, 14]. In the latter case, discrete graphical models – related to Ising-type models in statistical physics – have also been extensively studied [see, e.g. 43, 36].

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 [22] as treated by Cheng et al. [9], Lee and Hastie [23]. 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. 8, 47, 41]. A third approach uses latent Gaussian models, with a key recent reference being the paper of Fan et al. [12], 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 [12], 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 [32, 33] 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 [24] 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) [17], CLIME [7], or the graphical Dantzig selector [49]. 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 [24, 25, 46], 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 [24], 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. [12], 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 [24]. 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 [35] and of Feng and Ning [13]. The problem phrased by Fan et al. [12] 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 [37]. 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 [31]. 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 [10].

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

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

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 [30] 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 $\hat{\mathbf{\Sigma}}^{(n)}$ of $\mathbf{\Sigma}$; *Case III*)**.** *Recall the log-likelihood in Eq.* (8)*. The estimator* $\hat{\mathbf{\Sigma}}^{(n)} = (\hat{\Sigma}_{jk}^{(n)})_{1 \leq j < k \leq d_1}$ *of* $\mathbf{\Sigma}$ *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 [20].

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

3. Latent Gaussian Copula Models

Fan et al. [12] 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 [25], 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. [12] lies in deriving corresponding bridge functions for cases II and III. To reduce computational burden Yoon, Müller and Gaynanova [48] propose a hybrid multilinear interpolation and optimization scheme of the underlying latent correlation.

When considering the general mixed case, Fan et al. [12] advocate for binarizing all ordinal variables. This concept has been embraced by Feng and Ning [13], 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 [35] 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 [24]. 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 [25]. 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 [24] 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 [24] 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:*

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

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 $|\Gamma_j^r| \leq G$ 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 [26], 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*

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

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. [37]. 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. [25] 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* $C_\Gamma = \sum_{r=1}^{l_j}\phi(\bar{\Gamma}_j^r)(x_j^{r+1} - x_j^r)$ *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 [46] and Liu et al. [25] 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. [12], 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 [17], 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 [24], convergence rate and graph selection results follow immediately.

We do not penalize diagonal entries of $\boldsymbol{\Omega}$ and therefore have to make sure that $\hat{\boldsymbol{\Sigma}}^{(n)}$ is at least positive semidefinite to establish convergence in Eq. (20). Hence, we need to project $\hat{\boldsymbol{\Sigma}}^{(n)}$ into the cone of positive semidefinite matrices; see also [25, 12]. In practice, we use an efficient implementation of the alternating projections method proposed by Higham [18].

To select the tuning parameter in Eq. (20) Foygel and Drton [16] introduce an extended BIC (eBIC) in particular for Gaussian graphical models establishing consistency in higher dimensions under mild asymptotic assumptions. We consider

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

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 $\ell^{(n)}(\hat{\boldsymbol{\Omega}}(E))$ 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 $\boldsymbol{\Omega}^*$ whose zero pattern encodes the undirected graph. We set $\boldsymbol{\Omega}^*_{jj} = 1$ and $\boldsymbol{\Omega}^*_{jk} = s \cdot b_{jk}$ 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 $\boldsymbol{\Sigma}^* = (\boldsymbol{\Omega}^*)^{-1}$ be rescaled such that all diagonal elements are equal to 1. Given $\boldsymbol{\Sigma}^*$, we first obtain the partially latent continuous data $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{X}_2)$ where $\mathbf{Z} \sim \mathrm{NPN}_d(\mathbf{0}, \boldsymbol{\Sigma}^*, f)$. In practice, we draw n i.i.d. samples from $\mathrm{N}_d(\mathbf{0}, \boldsymbol{\Sigma}^*)$ 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. [12].

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. [12]. For this purpose, Figure 1 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. [12]. (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.

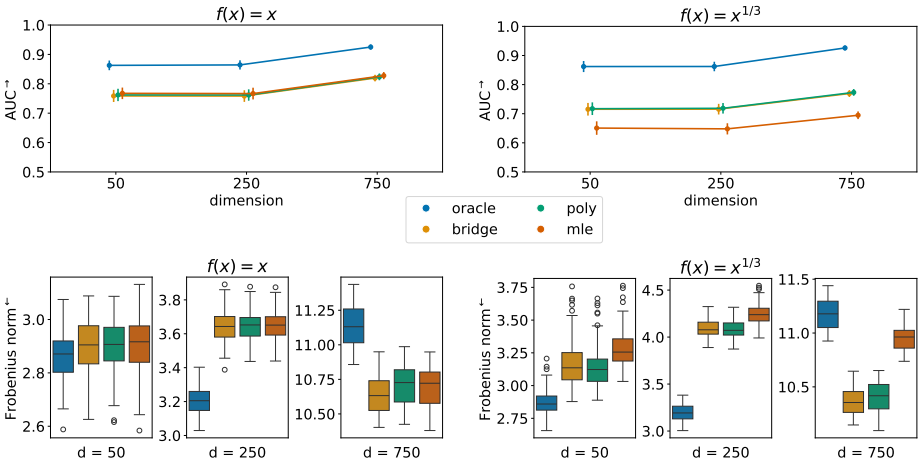

Fig 1. Simulation results for the binary-continuous data setting based on 100 *simulation runs. The left column corresponds to the latent Gaussian model, where the transformation function is the identity. The right colum depicts results for the LGCM with* $f_j(x) = x^{1/3}$ *for all* j*. The top row 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 1 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 1, 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. [12] does not extend beyond the binary-continuous mix, we can still compare our approach to the ensemble method developed by Feng and Ning [13]. Due to its close connection to the original method, we continue to denote the proposed ensemble estimator `bridge`.

Similar to above, Figure 2 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

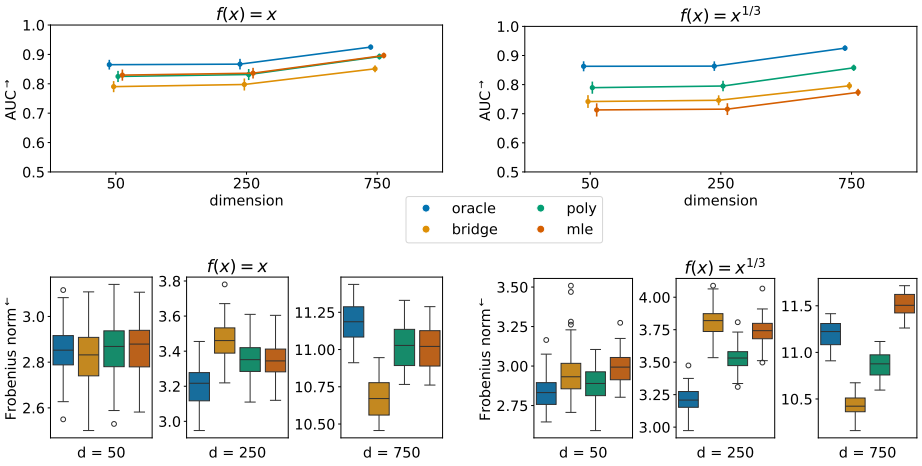

FIG 2. Simulation results for the general mixed data setting based on 100 *simulation runs. The left column corresponds to the latent Gaussian model, where the transformation function is the identity. The right colum depicts results for the LGCM with* $f_j(x) = x^{1/3}$ *for all* j*. The top row reports 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. [12], followed by Quan, Booth and Wells [35] and Feng and Ning [13]. 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`.

Appendix A: Methodology

In the following sections, we derive the MLEs for the latent Gaussian model.

A.1. Case II MLE derivation

In *Case II*, we consider the instance where X_j is ordinal and X_k is continuous. Recall that in the latent Gaussian model, we take all $\{f_k\}_{k=1}^d$ to be the identity. Consequently, X_k is Gaussian and $\Gamma_j^r = f_j(\gamma_j^r) = \gamma_j^r$ for all $j \in [d_1]$ and $r \in [l_j]$. Recall that 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 (Eq. (2)) are given. The likelihood of the n-sample is defined by:

$$\begin{aligned} L_{jk}^{(n)}(\Sigma_{jk}, x_j^r, x_k) &= \prod_{i=1}^{n} p(x_{ij}^r, x_{ik}, \Sigma_{jk}) \\ &= \prod_{i=1}^{n} p(x_{ik}) p(x_{ij}^r \mid x_{ik}, \Sigma_{jk}), \end{aligned} \tag{22}$$

where $p(x_{ij}^r, x_{ik}, \Sigma_{jk})$ denotes the joint probability of X_j and X_k and $p(x_{ik})$ the marginal density of the Gaussian variable X_k, i.e.

$$p(x_{ik}) = \left(2\pi\sigma\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(\frac{x_{ik} - \mu}{\sigma_{ik}}\right)^2 \right].$$

Furthermore, the conditional probability of X_j in Eq. (5) can be written as:

$$\begin{aligned} p(X_j = x_j^r \mid X_k, \Sigma_{jk}) &= p(\Gamma_j^{r-1} \leq Z_j < \Gamma_j^r \mid X_k, \Sigma_{jk}) \\ &= p(Z_j \leq \Gamma_j^r \mid X_k, \Sigma_{jk}) - p(Z_j \leq \Gamma_j^{r-1} \mid X_k, \Sigma_{jk}) \\ &\quad \Phi(\tilde{\Gamma}_j^r) - \Phi(\tilde{\Gamma}_j^{r-1}), \quad r = 1, \ldots, l_j, \end{aligned} \tag{23}$$

where

$$\tilde{\Gamma}_j^r = \frac{\Gamma_j^r - \Sigma_{jk}\tilde{X}_k}{\sqrt{1 - (\Sigma_{jk})^2}},$$

with $\tilde{X}_k = \frac{X_k - \mu_k}{\sigma_k}$ and $\Phi(t)$ denoting the standard normal distribution function. This follows straight from the fact that the conditional distribution of Z_j is Gaussian with mean $\Sigma_{jk}\tilde{X}_k$ and variance $(1 - (\Sigma_{jk})^2)$. The log-likelihood is then $\ell_{jk}^{(n)}(\Sigma_{jk}, x_j^r, x_k)$ with

$$\ell_{jk}^{(n)}(\Sigma_{jk}, x_j^r, x_k) = \sum_{i=1}^{n} \left[\log(p(x_{ik})) + \log(p(x_j^r \mid x_{ik}, \Sigma_{jk})) \right]. \tag{24}$$

Due to the heavy computational burden involved when estimating all parameters simultaneously, a *two-step estimator* has been proposed [31]. That is, in a first step μ_k, σ_k^2 are estimated by $\bar{X}_k$ and s_k^2, respectively. Moreover, the thresholds $\Gamma_j^r, r = 1, \ldots, l_j$ are estimated by the quantile function of the standard normal distribution evaluated at the cumulative marginal proportions of x_j^r just as described in Section 3.4.

In a second step, all that remains is obtaining the MLE for Σ_{jk} now with the readily computed estimates from the first step:

$$\frac{\partial \ell_{jk}^{(n)}(\Sigma_{jk}, x_j^r, x_k)}{\partial \Sigma_{jk}} = \sum_{i=1}^{n} \frac{1}{p(x_{ij}^r \mid x_{ik}, \Sigma_{jk})} \frac{\partial p(x_{ij}^r \mid x_{ik}, \Sigma_{jk})}{\partial \Sigma_{jk}}. \tag{25}$$

Let us take a closer look at the partial derivative of the conditional probability in Eq. (25):

$$\begin{aligned}
&\frac{\partial p(x_{ij}^r \mid x_{ik}, \Sigma_{jk})}{\partial \Sigma_{jk}} \\
&= \frac{\partial \Phi(\tilde{\Gamma}_j^r)}{\partial \Sigma_{jk}} - \frac{\partial \Phi(\tilde{\Gamma}_j^{r-1})}{\partial \Sigma_{jk}} \\
&= \phi(\tilde{\Gamma}_j^r)\frac{\partial \tilde{\Gamma}_j^r}{\partial \Sigma_{jk}} - \phi(\tilde{\Gamma}_j^{r-1})\frac{\partial \tilde{\Gamma}_j^r}{\partial \Sigma_{jk}} \\
&= (1-(\Sigma_{jk})^2)^{-\frac{3}{2}}\Big[\phi(\tilde{\Gamma}_j^r)(\Gamma_j^r\Sigma_{jk} - \tilde{x}_{ik}) - \phi(\tilde{\Gamma}_j^{r-1})(\Gamma_j^{r-1}\Sigma_{jk} - \tilde{x}_{ik})\Big],
\end{aligned} \tag{26}$$

where

$$\tilde{X}_k = \frac{X_k - \bar{X}_k}{\sqrt{s_k^2}} \quad \text{and} \quad \tilde{\Gamma}_j^r = \frac{\Gamma_j^r - \Sigma_{jk}\tilde{X}_k}{\sqrt{1-(\Sigma_{jk})^2}}.$$

The last equality in Eq. (26) follows from taking the derivative and applying the chain rule. Putting all the pieces together, we obtain

$$\begin{aligned}
\frac{\partial \ell_{jk}^{(n)}(\Sigma_{jk}, x_j^r, x_k)}{\partial \Sigma_{jk}} = \sum_{i=1}^{n}\Bigg[&\frac{1}{p(x_{ij}^r \mid x_{ik}, \Sigma_{jk})}(1-(\Sigma_{jk})^2)^{-\frac{3}{2}} \\
&\Big[\phi(\tilde{\Gamma}_j^r)(\Gamma_j^r\Sigma_{jk} - \tilde{x}_{ik}) - \phi(\tilde{\Gamma}_j^{r-1})(\Gamma_j^{r-1}\Sigma_{jk} - \tilde{x}_{ik})\Big]\Bigg].
\end{aligned} \tag{27}$$

Setting Eq. (27) to zero and solving for Σ_{jk} yields the *Case II* two-step MLE for Σ_{jk}. Note that this is a nonlinear optimization problem, which can efficiently be solved utilizing some quasi-Newton method.

A.2. Case III MLE derivation

Turning to *Case III*, where both X_j and X_k are ordinal variables, we argue in Section 3.3 in the manuscript that the MLE for the polychoric correlation remains valid for the LGCM. The probability of an observation taking values $X_j = x_j^r$ and $X_k = x_k^s$ is

$$\begin{aligned}
\pi_{rs} &:= p(X_j = x_j^r, X_k = x_k^s) \\
&= p(\Gamma_j^{r-1} \leq Z_j < \Gamma_j^r, \Gamma_k^{s-1} \leq Z_k < \Gamma_k^s) \\
&= p(\Gamma_j^{r-1} \leq f_j(Z_j) < \Gamma_j^r, \Gamma_k^{s-1} \leq f_k(Z_k) < \Gamma_k^s) \\
&= \int_{\Gamma_j^{r-1}}^{\Gamma_j^r}\int_{\Gamma_k^{s-1}}^{\Gamma_s^k} \phi_2(z_j, z_k, \Sigma_{jk}) dz_j dz_k,
\end{aligned} \tag{28}$$

where $r \in [l_j]$ and $s \in [l_k]$ and $\phi_2(x, y, \rho)$ denotes the standard bivariate density with correlation ρ. Then, as in the manuscript, the likelihood and log-likelihood of the n-sample are defined as:

$$\begin{aligned} L_{jk}^{(n)}(\Sigma_{jk}, x_j^r, x_k^s) &= C \prod_{r=1}^{l_j} \prod_{s=1}^{l_k} \pi_{rs}^{n_{rs}}, \\ \ell_{jk}^{(n)}(\Sigma_{jk}, x_j^r, x_k^s) &= \log(C) + \sum_{r=1}^{l_j} \sum_{s=1}^{l_k} n_{rs} \log(\pi_{rs}). \end{aligned} \tag{29}$$

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 $n = \sum_{r=1}^{l_j} \sum_{s=1}^{l_k} n_{rs}$. Similar to *Case II* above, we employ the *two-step estimator* for the polychoric correlation. Given the threshold estimates from the first step, let us state the derivative of $\ell_{jk}^{(n)}(\Sigma_{jk}, x_j^r, x_k^s)$ with respect to Σ_{jk} explicitly. First, recall that from Eq. (28)

$$\begin{aligned} \pi_{rs} &= \int_{\Gamma_j^{r-1}}^{\Gamma_j^r} \int_{\Gamma_k^{s-1}}^{\Gamma_s^k} \phi_2(z_j, z_k, \Sigma_{jk}) dz_j dz_k \\ &= \Phi_2(\Gamma_j^r, \Gamma_k^s, \Sigma_{jk}) - \Phi_2(\Gamma_j^{r-1}, \Gamma_k^s, \Sigma_{jk}) \\ &\quad - \Phi_2(\Gamma_j^r, \Gamma_k^{s-1}, \Sigma_{jk}) + \Phi_2(\Gamma_j^{r-1}, \Gamma_k^{s-1}, \Sigma_{jk}), \end{aligned} \tag{30}$$

where $\Phi_2(u, v, \rho)$ is the standard bivariate normal distribution function with correlation parameter ρ. Note also that we have $\frac{\partial \Phi_2(u,v,\rho)}{\partial \rho} = \phi_2(u, v, \rho)$; see [40]. Taking the derivative of $\ell_{jk}^{(n)}(\Sigma_{jk}, x_j^r, x_k^s)$ with respect to Σ_{jk} yields

$$\begin{aligned} \frac{\partial \ell^{(n)}(\Sigma_{jk}, x_j^r, x_k^s)}{\partial \Sigma_{jk}} &= \sum_{r=1}^{l_{X_j}} \sum_{s=1}^{l_{X_k}} \frac{n_{rs}}{\pi_{rs}} \frac{\partial \pi_{rs}}{\partial \Sigma_{jk}} \\ &= \sum_{r=1}^{l_{X_j}} \sum_{s=1}^{l_{X_k}} \frac{n_{rs}}{\pi_{rs}} \Big[\phi_2(\Gamma_j^r, \Gamma_k^s, \Sigma_{jk}) - \phi_2(\Gamma_j^{r-1}, \Gamma_k^s, \Sigma_{jk}) - \\ &\qquad \phi_2(\Gamma_j^r, \Gamma_k^{s-1}, \Sigma_{jk}) + \phi_2(\Gamma_j^{r-1}, \Gamma_k^{s-1}, \Sigma_{jk}) \Big]. \end{aligned}$$

Again, setting the derivative to zero and solving for Σ_{jk} using some quasi-Newton method yields the *Case III* two-step MLE for Σ_{jk}.

A.3. Comparison of Case II estimators under the LGCM

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, which involves incorporating the estimated transformations $\hat{f}$ at appropriate locations.

Furthermore, we investigate the ad hoc estimator presented in Section 3.2 in more detail.

We start by rewriting Eq. (27), 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:

$$\frac{\partial \ell_{jk}^{(n)}(\Sigma_{jk}, x_j^r, \hat{f}_k(x_k))}{\partial \Sigma_{jk}} \tag{31}$$

$$= \sum_{i=1}^{n} \Bigg[\frac{1}{\Phi(\tilde{\Gamma}_j^r(\hat{f}_k)) - \Phi(\tilde{\Gamma}_j^{r-1}(\hat{f}_k))} (1 - (\Sigma_{jk})^2)^{-\frac{3}{2}} \tag{32}$$

$$\Big[\phi(\tilde{\Gamma}_j^r(\hat{f}_k))(\Gamma_j^r \Sigma_{jk} - \hat{f}_k(\tilde{x}_{ik})) - \phi(\tilde{\Gamma}_j^{r-1}(\hat{f}_k))(\Gamma_j^{r-1} \Sigma_{jk} - \hat{f}_k(\tilde{x}_{ik})) \Big] \Bigg], \tag{33}$$

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.

$$\begin{aligned} \ell_{jk}^{(n)}(\Sigma_{jk}, x_j^r, \hat{f}_k(x_k)) &= \sum_{i=1}^{n} \big[\log(p(\hat{f}_k(x_{ik}))) + \log(p(x_j^r \mid \hat{f}_k(x_{ik}), \Sigma_{jk})) \big] \\ &= \sum_{i=1}^{n} \big[\log(p(\hat{f}_k(x_{ik}))) + \Phi(\tilde{\Gamma}_j^r(\hat{f}_k)) - \Phi(\tilde{\Gamma}_j^{r-1}(\hat{f}_k)) \big]. \end{aligned} \tag{34}$$

One strategy to optimize Eq. (34) 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` [15]. Alternatively, another approach involves utilizing the Eq. (31) and solving for $\hat{\Sigma}_{jk}$ through a nonlinear root-finding method. To do this, we employ Broyden's method [6].

In Figure A.3, 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

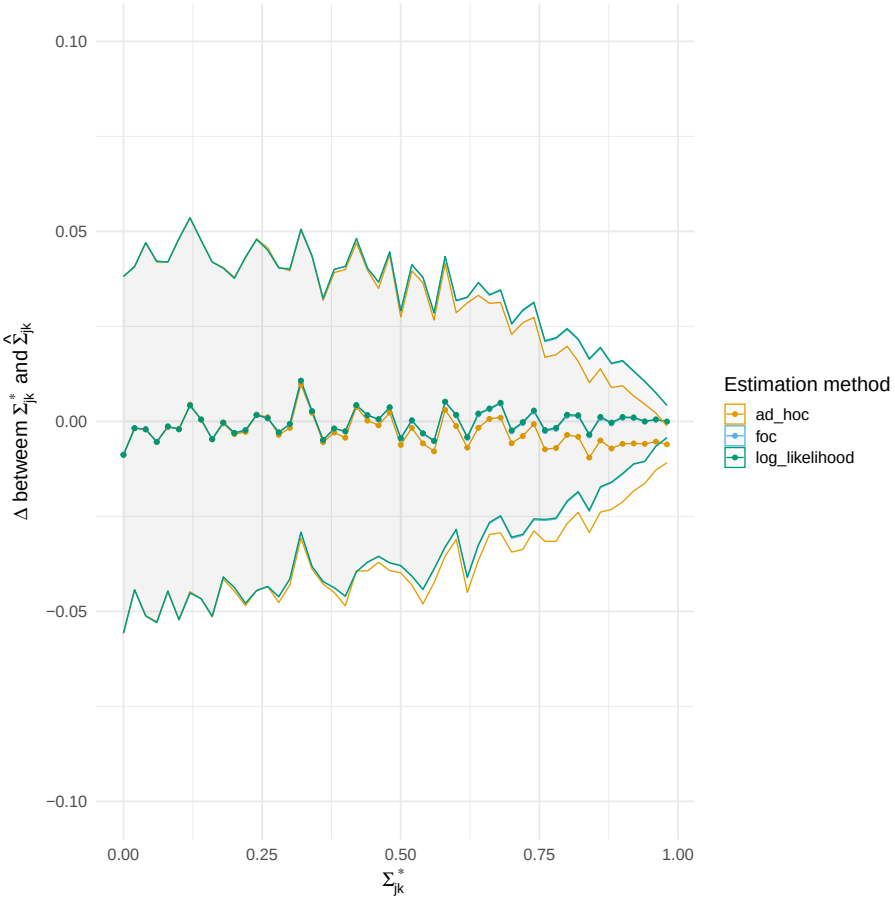

Fig 3. Comparison of the Case II MLE under the LGCM and the ad hoc estimator.

strategies) and the ad hoc estimator from Section 3.2. 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 31, for a similar observation]. The strength of the ad hoc estimator lies in its simplicity and computational efficiency. The left panel of Figure A.3 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 A.3 demonstrates computation time across a grid of

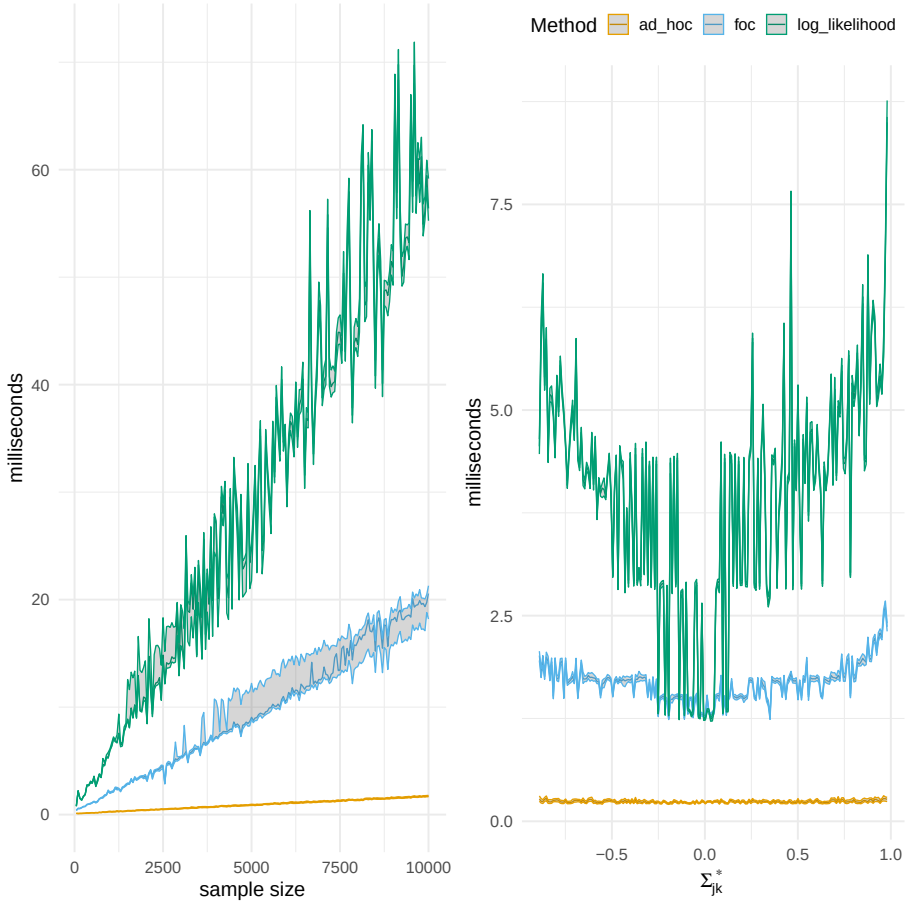

Fig 4. Computation time in milliseconds for the Case II MLE and ad hoc estimators. We report the median (solid line) and the first and third quartile (shaded area) of recorded computation time. In the left panel, we compare computation time against a grid of sample sizes $n \in [50, 10000]$ *with a step size of* $s_t = 50$*. In the right panel, we compare computation time against a grid of true correlation values* $\Sigma^*_{jk} \in [-.98, .98]$*.*

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.

Appendix B: Proofs

B.1. Proof of Theorem 3.2

Condition B.1 (Gradient statistical noise)**.** *The gradient of the log-likelihood function is τ^2-sub-Gaussian. That is, for any $\lambda \in \mathbb{R}$ and for all $\Sigma_{jk} \in [-1+\delta, 1-\delta]$ for j, k according to Case II or Case III.*

$$\mathbb{E}\left[\exp\left(\lambda\Big(\frac{\partial \ell_{jk}}{\partial \Sigma_{jk}} - \mathbb{E}\frac{\partial \ell_{jk}}{\partial \Sigma_{jk}}\Big)\right)\right] \leq \exp\Big(\frac{\tau^2\lambda^2}{2}\Big), \tag{35}$$

where ℓ_{jk} corresponds to the log-likelihood functions in Definitions 2.4 and 2.5, respectively.

Case II *: Recall that*

$$\frac{\partial \ell_{jk}(\Sigma_{jk}, x^r_j, x_k)}{\partial \Sigma_{jk}} = \frac{1}{p(x^r_{ij} \mid x_{ik}, \Sigma_{jk})} \frac{\partial p(x^r_{ij} \mid x_{ik}, \Sigma_{jk})}{\partial \Sigma_{jk}}.$$

Replacing these with the derivations made in Eq. (27), we write

$$\frac{\partial \ell_{jk}(\Sigma_{jk}, x^r_j, x_k)}{\partial \Sigma_{jk}} = \frac{(1 - (\Sigma_{jk})^2)^{-\frac{3}{2}}}{\Phi(\tilde{\Gamma}^r_j) - \Phi(\tilde{\Gamma}^{r-1}_j)} \left[\phi(\tilde{\Gamma}^r_j)(\Gamma^r_j \Sigma_{jk} - \tilde{x}_k) - \phi(\tilde{\Gamma}^{r-1}_j)(\Gamma^{r-1}_j \Sigma_{jk} - \tilde{x}_k)\right].$$

It is easy to see that $p(x^r_{ij} \mid x_{ik}, \Sigma_{jk}) \in (0, 1)$ almost surely. Assumption 3.3 makes sure that we exclude impossible events where $p(x^r_{ij} \mid x_{ik}, \Sigma_{jk}) = 0$. Moreover, we require that $\Gamma^r_j > \Gamma^{r-1}_j, \forall j \in 1, \ldots, d_1$ this implies that $\Phi(\tilde{\Gamma}^r_j) > \Phi(\tilde{\Gamma}^{r-1}_j)$. In other words, there exists a $\kappa > 0$ such that $p(x^r_{ij} \mid x_{ik}, \Sigma_{jk}) \leq \frac{1}{\kappa}$.

Let us now turn to $\partial p(x_{ij}^r \mid x_{ik}, \Sigma_{jk})/\partial\Sigma_{jk}$. *First, for all* $\Sigma_{jk} \in [-1+\delta, 1-\delta]$ *we clearly have* $1 \leq (1-(\Sigma_{jk})^2)^{-\frac{3}{2}} \leq \varpi$ *for* $\varpi > 1$. *What's more, the density of the standard normal is bounded, i.e.,* $\left|\phi(t)\right| \leq (2\pi)^{-\frac{1}{2}}$ *for all* $t \in \mathbb{R}$. *Similarly,*

$$\left|\phi(\tilde{\Gamma}_j^r)(\Gamma_j^r\Sigma_{jk} - \tilde{x}_k) - \phi(\tilde{\Gamma}_j^{r-1})(\Gamma_j^{r-1}\Sigma_{jk} - \tilde{x}_k)\right| \leq \left|\phi(\tilde{\Gamma}_j^r)(\Gamma_j^r\Sigma_{jk} - \tilde{x}_k)\right| \leq L_1,$$

due to Assumption 3.2. Therefore,

$$\left|\frac{\partial \ell_{jk}(\Sigma_{jk}, x_j^r, x_k)}{\partial \Sigma_{jk}}\right| \leq \kappa L_1,$$

and $\left(\frac{\partial \ell_{jk}}{\partial \Sigma_{jk}} - \mathbb{E}\frac{\partial \ell_{jk}}{\partial \Sigma_{jk}}\right)$ *is zero-mean and bounded. Then by Hoeffding's [19] lemma, the gradient of the log-likelihood function is* τ^2*-sub-Gaussian with* $\tau = 2\kappa L_1$

Case III *: Recall that we have*

$$\frac{\partial \ell_{jk}(\Sigma_{jk}, x_j^r, x_k^s)}{\partial \Sigma_{jk}} = \frac{1}{\pi_{rs}}\frac{\partial \pi_{rs}}{\partial \Sigma_{jk}}, \quad \text{for some } j < k \in [d_1].$$

Considering

$$\begin{aligned}\pi_{rs} =& \Phi_2(\Gamma_j^r, \Gamma_k^s, \Sigma_{jk}) - \Phi_2(\Gamma_j^{r-1}, \Gamma_k^s, \Sigma_{jk}) \\ & -\Phi_2(\Gamma_j^r, \Gamma_k^{s-1}, \Sigma_{jk}) + \Phi_2(\Gamma_j^{r-1}, \Gamma_k^{s-1}, \Sigma_{jk}),\end{aligned}$$

we note that this again has to be in $(0,1)$ *due to Assumptions 3.1 and 3.2. such that* $\pi_{rs} \leq \frac{1}{\xi}$.

Now let us show that

$$\begin{aligned}\frac{\partial \pi_{rs}}{\partial \Sigma_{jk}} = \Big[&\phi_2(\Gamma_j^r, \Gamma_k^s, \Sigma_{jk}) - \phi_2(\Gamma_j^{r-1}, \Gamma_k^s, \Sigma_{jk}) \\ & -\phi_2(\Gamma_j^r, \Gamma_k^{s-1}, \Sigma_{jk}) + \phi_2(\Gamma_j^{r-1}, \Gamma_k^{s-1}, \Sigma_{jk})\Big]\end{aligned}$$

is bounded. Indeed, the density of the standard bivariate normal random variable is of the form $\phi_2(x,y) = ce^{-q(x,y)}$. *Since* $q(x,y)$ *is a quadratic function of* x, y *it follows that* $\left|\phi_2(x,y)\right| \leq c$. *Therefore, every element in* $\frac{\partial \pi_{rs}}{\partial \Sigma_{jk}}$ *is bounded and thus* $\left|\frac{\partial \pi_{rs}}{\partial \Sigma_{jk}}\right| \leq K_1$. *By the same argument as for Case II* $\left(\frac{\partial \ell_{jk}}{\partial \Sigma_{jk}} - \mathbb{E}\frac{\partial \ell_{jk}}{\partial \Sigma_{jk}}\right)$ *is zero-mean and bounded and by Hoeffding's lemma the gradient of the log-likelihood function is* τ^2*-sub-Gaussian with* $\tau = 2\xi K_1$. *Based on these arguments, we can conclude that the condition for gradient statistical noise is satisfied.*

Condition B.2 (Hessian statistical noise)**.** *The Hessian of the log-likelihood function is* τ^2*-sub-exponential, i.e. for all* $\Sigma_{jk} \in [-1+\delta, 1-\delta]$ *and for* j, k *according to Case II or Case III we have*

$$\left\|\frac{\partial^2 \ell_{jk}}{\partial \Sigma_{jk}^2}\right\|_{\psi_1} \leq \tau^2, \tag{36}$$

where $\|\cdot\|_{\psi_1}$ *denotes the Orlicz* ψ_1*-norm, defined as*

$$\|X\|_{\psi_1} \coloneqq \sup_{p\geq 1} \frac{1}{p}\mathbb{E}\Big(\big|X-\mathbb{E}(X)\big|^p\Big)^{\frac{1}{p}}.$$

Note that ℓ_{jk} *corresponds to the respective log-likelihood functions in Definitions 2.4 and 2.5.*

Case II *: We have*

$$\begin{aligned}
\frac{\partial^2 \ell_{jk}}{\partial \Sigma_{jk}^2} &= \frac{\partial^2 p(x_j^r \mid x_k, \Sigma_{jk})/\partial \Sigma_{jk}^2}{p(x_j^r \mid x_k, \Sigma_{jk})} - \left(\frac{\partial p(x_j^r \mid x_k, \Sigma_{jk})/\partial \Sigma_{jk}}{p(x_j^r \mid x_k, \Sigma_{jk})}\right)^2.\\
\textit{Clearly, } \left|\frac{\partial^2 \ell_{jk}}{\partial \Sigma_{jk}^2}\right| &\leq \frac{\left|\partial^2 p(x_j^r \mid x_k, \Sigma_{jk})/\partial \Sigma_{jk}^2\right|}{\left|p(x_j^r \mid x_k, \Sigma_{jk})\right|} + \left(\frac{\left|\partial p(x_j^r \mid x_k, \Sigma_{jk})/\partial \Sigma_{jk}\right|}{\left|p(x_j^r \mid x_k, \Sigma_{jk})\right|}\right)^2\\
&\leq \kappa L_2 + \kappa^2 L_1^2,
\end{aligned} \tag{37}$$

where it remains to show that $\left|\partial^2 p(x_{ij}^r \mid x_{ik}, \Sigma_{jk})/\partial \Sigma_{jk}^2\right| \leq L_2$*. Indeed, we can rewrite our objective as follows:*

$$\begin{aligned}
&\frac{\partial}{\partial \Sigma_{jk}}\left(\frac{\partial \ell_{jk}(\Sigma_{jk}, x_j^r, x_k)}{\partial \Sigma_{jk}}\right)\\
&= \frac{\partial}{\partial \Sigma_{jk}}\left((1-(\Sigma_{jk})^2)^{-\frac{3}{2}}\left[\phi(\tilde{\Gamma}_j^r)(\Gamma_j^r \Sigma_{jk} - \tilde{x}_k) - \phi(\tilde{\Gamma}_j^{r-1})(\Gamma_j^{r-1}\Sigma_{jk} - \tilde{x}_k)\right]\right)\\
&= \frac{3\Sigma_{jk}}{1-\Sigma_{jk}^2}(1-(\Sigma_{jk})^2)^{-\frac{3}{2}}\phi(\tilde{\Gamma}_j^r)(\Gamma_j^r\Sigma_{jk}-\tilde{x}_k) + \frac{\phi'(\tilde{\Gamma}_j^r)(\Gamma_j^r\Sigma_{jk}-\tilde{x}_k)^2}{(1-\Sigma_{jk}^2)^3}\\
&+ \frac{\phi(\tilde{\Gamma}_j^r)\Gamma_j^r}{(1-\Sigma_{jk}^2)^{-\frac{3}{2}}} - \frac{3\Sigma_{jk}}{1-\Sigma_{jk}^2}(1-(\Sigma_{jk})^2)^{-\frac{3}{2}}\phi(\tilde{\Gamma}_j^{r-1})(\Gamma_j^{r-1}\Sigma_{jk}-\tilde{x}_k)\\
&- \frac{\phi'(\tilde{\Gamma}_j^{r-1})(\Gamma_j^{r-1}\Sigma_{jk}-\tilde{x}_k)^2}{(1-\Sigma_{jk}^2)^3} - \frac{\phi(\tilde{\Gamma}_j^{r-1})\Gamma_j^{r-1}}{(1-\Sigma_{jk}^2)^{-\frac{3}{2}}}.
\end{aligned} \tag{38}$$

Thus,

$$\begin{aligned}
\left|\frac{\partial}{\partial \Sigma_{jk}}\left(\frac{\partial \ell_{jk}(\Sigma_{jk}, x_j^r, x_k)}{\partial \Sigma_{jk}}\right)\right| &\leq \left|\frac{3\Sigma_{jk}}{1-\Sigma_{jk}^2}(1-(\Sigma_{jk})^2)^{-\frac{3}{2}}\phi(\tilde{\Gamma}_j^r)(\Gamma_j^r\Sigma_{jk}-\tilde{x}_k)\right|\\
&\quad + \left|\frac{\phi'(\tilde{\Gamma}_j^r)(\Gamma_j^r\Sigma_{jk}-\tilde{x}_k)^2}{(1-\Sigma_{jk}^2)^3} + \frac{\phi(\tilde{\Gamma}_j^r)\Gamma_j^r}{(1-\Sigma_{jk}^2)^{-\frac{3}{2}}}\right|\\
&\leq L_2,
\end{aligned}$$

due to Assumptions 3.1 and 3.2 and because both $\phi(t)$ and $\phi'(t)$ are bounded for all $t \in \mathbb{R}$. Therefore, the inequality in Eq. (37) follows and $\frac{\partial^2 \ell_{jk}}{\partial \Sigma_{jk}^2} - \mathbb{E}\left(\frac{\partial^2 \ell_{jk}}{\partial \Sigma_{jk}^2}\right)$ is bounded by $2(\kappa L_2 + \kappa^2 L_1^2)$. Hence, for all $p \geq 1$

$$\frac{1}{p}\mathbb{E}\left[\left|\partial^2 \ell_{jk}/\partial \Sigma_{jk}^2 - \mathbb{E}\big(\partial^2 \ell_{jk}/\partial \Sigma_{jk}^2\big)\right|^p\right]^{\frac{1}{p}} \leq \frac{2}{p}\big(\kappa L_2 + \kappa^2 L_1^2\big). \tag{39}$$

Finally, for $\tau = 2\kappa L_1$ we can choose L_1 and κ such that

$$2(\kappa L_2 + \kappa^2 L_1^2) \leq \tau^2 = 4\kappa^2 L_1^2.$$

Thus, the Hessian statistical noise-condition for Case II is satisfied.

Case III: *Let us start with the Hessian of ℓ_{jk} in the polychoric case:*

$$\begin{aligned} \frac{\partial^2 \ell_{jk}(\Sigma_{jk}, x_j^r, x_k^s)}{\partial \Sigma_{jk}^2} &= \frac{\partial^2 \pi_{rs}/\partial \Sigma_{jk}^2}{\pi_{rs}} - \left(\frac{\partial \pi_{rs}/\partial \Sigma_{jk}}{\pi_{rs}}\right)^2 \\ \textit{Thus: } \left|\frac{\partial^2 \ell_{jk}(\Sigma_{jk}, x_j^r, x_k^s)}{\partial \Sigma_{jk}^2}\right| &\leq \frac{\left|\partial^2 \pi_{rs}/\partial \Sigma_{jk}^2\right|}{|\pi_{rs}|} + \left(\frac{\left|\partial \pi_{rs}/\partial \Sigma_{jk}\right|}{|\pi_{rs}|}\right)^2 \\ &\leq \xi K_2 + \xi^2 K_1^2. \end{aligned} \tag{40}$$

Again it remains to show that $\partial^2 \pi_{rs}/\partial \Sigma_{jk}^2 \leq K_2$. Consider

$$\begin{aligned} &\left|\frac{\partial}{\partial \Sigma_{jk}}\left(\frac{\partial \pi_{rs}}{\partial \Sigma_{jk}}\right)\right| \\ &= \Big|\frac{\partial}{\partial \Sigma_{jk}}\phi_2(\Gamma_j^r, \Gamma_k^s, \Sigma_{jk}) - \phi_2(\Gamma_j^{r-1}, \Gamma_k^s, \Sigma_{jk}) \\ &- \phi_2(\Gamma_j^r, \Gamma_k^{s-1}, \Sigma_{jk}) + \phi_2(\Gamma_j^{r-1}, \Gamma_k^{s-1}, \Sigma_{jk})\Big| \\ &\leq \left|\frac{\partial}{\partial \Sigma_{jk}}\phi_2(\Gamma_j^r, \Gamma_k^s, \Sigma_{jk})\right| + \left|\frac{\partial}{\partial \Sigma_{jk}}\phi_2(\Gamma_j^{r-1}, \Gamma_k^s, \Sigma_{jk})\right| \\ &+ \left|\frac{\partial}{\partial \Sigma_{jk}}\phi_2(\Gamma_j^r, \Gamma_k^{s-1}, \Sigma_{jk})\right| + \left|\frac{\partial}{\partial \Sigma_{jk}}\phi_2(\Gamma_j^{r-1}, \Gamma_k^{s-1}, \Sigma_{jk})\right| \\ &\leq K_2, \end{aligned}$$

where each of the derivatives of the bivariate density is bounded, since we assume that the correlation is bounded away from one and minus one, i.e., $\Sigma_{jk} \in [-1+\delta, 1-\delta]$.

Similar to Case II, for $\tau = 2\xi K_1$ we can choose K_1 and ξ such that

$$2(\xi k_2 + \xi^2 k_1^2) \leq \tau^2 = 4\xi^2 K_1^2.$$

Consequently, the Hessian statistical noise-condition for Case III is satisfied, which concludes the proof of the Hessian statistical noise-condition.

Concerning the third condition, we introduce some additional notation. Denote the sample risk by $\hat{R}_{jk}^{(n)}(\Sigma_{jk})$ for j,k according to *Case II* and *Case III*, i.e.,

$$\textit{Case II:}\quad \hat{R}_{jk}^{(n)}(\Sigma_{jk}) = \frac{1}{n}\sum_{i=1}^{n}\big[\log(p(x_{ik})) + \log(p(x_{ij}^r \mid x_{ik}, \Sigma_{jk}))\big],$$

and

$$\textit{Case III:}\quad \hat{R}_{jk}^{(n)}(\Sigma_{jk}) = \frac{1}{n}\sum_{r=1}^{l_{X_j}}\sum_{s=1}^{l_{X_k}} n_{rs}\log(\pi_{rs}).$$

Lastly, we define $R_{jk}(\Sigma_{jk}) = \mathbb{E}_{\Sigma_{jk}^*}\hat{R}_{jk}^{(n)}(\Sigma_{jk})$ to be the population risk for each of the respective cases.

Condition B.3 (Hessian regularity)**.** *The Hessian regularity condition consists of three parts:*

1. *The second derivative of the population risk $R_{jk}(\Sigma_{jk})$ is bounded at one point. That is, there exists one $|\bar{\Sigma}_{jk}| \le 1-\delta$ and $H > 0$ such that $|R''_{jk}(\bar{\Sigma}_{jk})| \le H$.*
2. *The second derivative of the log-likelihood with respect to Σ_{jk} is Lipschitz continuous with integrable Lipschitz constant, i.e. there exists a $M^* > 0$ such that $\mathbb{E}[M] \le M^*$, where*

$$M = \sup_{\substack{|\Sigma_{jk}^{(1)}|,\ |\Sigma_{jk}^{(2)}|\le 1-\delta,\\ \Sigma_{jk}^{(1)}\neq\Sigma_{jk}^{(2)}}} \frac{\left|\ell''_{jk}(\Sigma_{jk}^{(1)}) - \ell''_{jk}(\Sigma_{jk}^{(2)})\right|}{\left|\Sigma_{jk}^{(1)} - \Sigma_{jk}^{(2)}\right|}.$$

3. *The constants H and M^* are such that $H \le \tau^2$ and $M^* \le \tau^3$.*

We need some intermediate results that make dealing with $R_{jk}(\Sigma_{jk})$ easier. First, note that $\mathbb{E}_{\Sigma_{jk}^*}\hat{R}_{jk}^{(n)}(\Sigma_{jk}) = \mathbb{E}_{\Sigma_{jk}^*}\ell_{jk}(\Sigma_{jk})$. Second, for all $\Sigma_{jk} \in [-1+\delta, 1-\delta]$ and $m \in \{1,2\}$

$$R_{jk}^m(\Sigma_{jk}) = \frac{\partial^m}{\partial\Sigma_{jk}^m}\mathbb{E}_{\Sigma_{jk}^*}\ell_{jk}(\Sigma_{jk}) = \mathbb{E}_{\Sigma_{jk}^*}\frac{\partial^m}{\partial\Sigma_{jk}^m}\ell_{jk}(\Sigma_{jk}),$$

by Lemma B.7 and Corollary B.8.

1. *Recall the first part of the Hessian regularity condition, whereby Eq. (37) and Eq. (40) for all $\Sigma_{jk} \in [-1+\delta, 1-\delta]$ we have*

$$\textit{Case II:}\quad \left|\frac{\partial^2}{\partial\Sigma_{jk}^2}\ell_{jk}(\Sigma_{jk})\right| \le \kappa L_2 + \kappa^2 L_1^2$$

and

$$\textit{Case III:}\quad \left|\frac{\partial^2}{\partial\Sigma_{jk}^2}\ell_{jk}(\Sigma_{jk})\right| \le \xi K_2 + \xi^2 K_1^2$$

for cases II and III, respectively. Consequently, the claim in the first part of the Hessian regularity condition holds for Case II and Case III for any $|\bar{\Sigma}_{jk}| \leq 1-\delta$ *with* $H_1 = \kappa L_2 + \kappa^2 L_1^2$ *and* $H_2 = \xi K_2 + \xi^2 K_1^2$*. Moreover, we also have* $H_1 \leq \tau^2 = 4\kappa^2 L_1^2$ *and* $H_2 \leq \tau^2 = 4\xi^2 K_1^2$*.*

2. *The second part of the Hessian regularity condition requires that the second derivative of the log-likelihood with respect to* Σ_{jk} *is Lipschitz continuous with integrable Lipschitz constant. By the mean-value-theorem, all we need to show is that we can find a bound on the third derivative of the log-likelihood function.*

Case II: *Note that we have*

$$\begin{aligned}
&\frac{\partial^3}{\partial \Sigma_{jk}^3}\ell_{jk}(\Sigma_{jk}) \\
&\quad= \frac{\partial}{\partial \Sigma_{jk}}\left[\frac{\partial^2 \ell_{jk}}{\partial \Sigma_{jk}^2}\right] \\
&\quad= \frac{\partial}{\partial \Sigma_{jk}}\left[\frac{\partial^2 p(x_j^r \mid x_k, \Sigma_{jk})/\partial \Sigma_{jk}^2}{p(x_j^r \mid x_k, \Sigma_{jk})} - \left(\frac{\partial p(x_j^r \mid x_k, \Sigma_{jk})/\partial \Sigma_{jk}}{p(x_j^r \mid x_k, \Sigma_{jk})}\right)^2\right] \\
&\quad= \frac{\partial^3 p(x_j^r \mid x_k, \Sigma_{jk})/\partial \Sigma_{jk}^3}{p(x_j^r \mid x_k, \Sigma_{jk})} \\
&\quad- 3\frac{\big(\partial p(x_j^r \mid x_k, \Sigma_{jk})/\partial \Sigma_{jk}\big)\big(\partial^2 p(x_j^r \mid x_k, \Sigma_{jk})/\partial \Sigma_{jk}^2\big)}{(p(x_j^r \mid x_k, \Sigma_{jk}))^2} \\
&\quad+ 2\left(\frac{\partial p(x_j^r \mid x_k, \Sigma_{jk})/\partial \Sigma_{jk}}{p(x_j^r \mid x_k, \Sigma_{jk})}\right)^3.
\end{aligned}$$

Hence

$$\left|\frac{\partial^3}{\partial \Sigma_{jk}^3}\ell_{jk}(\Sigma_{jk})\right| \leq \kappa L_3 + 3\kappa^2 L_2 L_1 + 2\kappa^3 L_1^3.$$

It remains to show therefore, that

$$\left|\partial^3 p(x_j^r \mid x_k, \Sigma_{jk})/\partial \Sigma_{jk}^3\right| \leq L_3.$$

When taking the derivative of Eq. (38), it is obvious that the resulting statement is bounded due to Assumptions 3.1 and 3.2 and the fact that $\phi(t), \phi'(t), \phi''(t)$ *are all bounded for all* $t \in \mathbb{R}$*.*
Therefore, by applying the mean-value-theorem we get

$$M_1 \leq \kappa L_3 + 3\kappa^2 L_2 L_1 + 2\kappa^3 L_1^3,$$

and the natural choice for

$$M_1^* = \kappa L_3 + 3\kappa^2 L_2 L_1 + 2\kappa^3 L_1^3,$$

where it follows that

$$M_1^* \leq \tau^3 = 8\kappa^3 L_1^3.$$

Case III: *We proceed similarly and first consider*

$$\begin{aligned}
&\frac{\partial^3}{\partial\Sigma_{jk}^3}\ell_{jk}(\Sigma_{jk}) \\
&\quad= \frac{\partial}{\partial\Sigma_{jk}}\left[\frac{\partial^2\ell_{jk}(\Sigma_{jk}, x_j^r, x_k^s)}{\partial\Sigma_{jk}^2}\right] \\
&\quad= \frac{\partial}{\partial\Sigma_{jk}}\left[\frac{\partial^2\pi_{rs}/\partial\Sigma_{jk}^2}{\pi_{rs}} - \left(\frac{\partial\pi_{rs}/\partial\Sigma_{jk}}{\pi_{rs}}\right)^2\right] \\
&\quad= \frac{\partial^3\pi_{rs}/\partial\Sigma_{jk}^3}{\pi_{rs}} - 3\frac{\big(\partial\pi_{rs}/\partial\Sigma_{jk}\big)\big(\partial^2\pi_{rs}/\partial\Sigma_{jk}^2\big)}{(\pi_{rs})^2} \\
&\quad+ 2\left(\frac{\partial\pi_{rs}/\partial\Sigma_{jk}}{\pi_{rs}}\right)^3.
\end{aligned} \tag{41}$$

Hence

$$\left|\frac{\partial^3}{\partial\Sigma_{jk}^3}\ell_{jk}(\Sigma_{jk})\right| \leq \xi K_3 + 3\xi^2 K_2 K_1 + 2\xi^3 K_1^3.$$

Taking a closer look at $\left|\partial^3\pi_{rs}/\partial\Sigma_{jk}^3\right|$*, boundedness again follows from the fact that the quadratic function in the exponential of the bivariate normal density does not vanish. Thus, we have*

$$M_2 \leq \xi K_3 + 3\xi^2 K_2 K_1 + 2\xi^3 K_1^3,$$

and the natural choice for

$$M_2^* = \xi K_3 + 3\xi^2 K_2 K_1 + 2\xi^3 K_1^3,$$

where we have

$$M_2^* \leq \tau^3 = 8\xi^3 K_1^3.$$

These steps validate the Hessian regularity condition.

Condition B.4 (Population risk is strongly Morse)**.** *There exist* $\epsilon > 0$ *and* $\eta > 0$ *such that* $R_{jk}(\Sigma_{jk})$ *is* (ϵ, η)*-strongly Morse, i.e.*

1. *For all* Σ_{jk} *such that* $|\Sigma_{jk}| = 1 - \delta$ *we have that* $|R'_{jk}(\Sigma_{jk})| > \epsilon$.
2. *For all* Σ_{jk} *such that* $|\Sigma_{jk}| \leq 1 - \delta$*:*

$$|R'_{jk}(\Sigma_{jk})| \leq \epsilon \implies |R''_{jk}(\Sigma_{jk})| \geq \eta.$$

Put differently, $R_{jk}(\Sigma_{jk})$ *is* (ϵ, η)*-strongly Morse if the boundaries* $\{-1+\delta, 1-\delta\}$ *are not critical points of* $R_{jk}(\Sigma_{jk})$ *and if* $R_{jk}(\Sigma_{jk})$ *only has finitely many critical points that are all non-degenerate.*

Let us verify that $R''(\Sigma_{jk}) \neq 0$ *for cases II and III. Indeed by Lemma B.7 and Corollary B.8 we can rewrite* $R''_{jk}(\Sigma_{jk})$ *and obtain*

$$
\begin{aligned}
&R''(\Sigma_{jk}) \\
&\quad = \mathbb{E}_{\Sigma_{jk}}\left[\frac{\partial^2 \ell_{jk}(\Sigma_{jk})}{\partial \Sigma_{jk}^2}\right] \\
&\quad = \mathbb{E}_{\Sigma^*_{jk}}\begin{cases} \dfrac{\partial^2 p(x_j^r \mid x_k, \Sigma_{jk})/\partial \Sigma_{jk}^2}{p(x_j^r \mid x_k, \Sigma_{jk})} - \left(\dfrac{\partial p(x_j^r \mid x_k, \Sigma_{jk})/\partial \Sigma_{jk}}{p(x_j^r \mid x_k, \Sigma_{jk})}\right)^2 & \textit{Case II,} \\ \dfrac{\partial^2 \pi(\Sigma_{jk})_{rs}/\partial \Sigma_{jk}^2}{\pi(\Sigma_{jk})_{rs}} - \left(\dfrac{\partial \pi(\Sigma_{jk})_{rs}/\partial \Sigma_{jk}}{\pi(\Sigma_{jk})_{rs}}\right)^2 & \textit{Case III,} \end{cases}
\end{aligned}
$$

where in $\pi(\Sigma_{jk})_{rs}$ *we made the dependence on* Σ_{jk} *explicit.*

Case II: *Recall that we have*

$$
\begin{aligned}
&\mathbb{E}_{\Sigma^*_{jk}}\left[\frac{\partial^2 p(x_j^r \mid x_k, \Sigma^*_{jk})/\partial \Sigma_{jk}^2}{p(x_j^r \mid x_k, \Sigma^*_{jk})}\right] \\
&\quad = \int_{-\infty}^{\infty} \sum_{r=1}^{l_j+1} \frac{\partial^2 p(x_j^r \mid x_k, \Sigma^*_{jk})/\partial \Sigma_{jk}^2}{p(x_j^r \mid x_k, \Sigma^*_{jk})} p(x_j^r, x_k; \Sigma^*_{jk}) dx_k \\
&\quad = \int_{-\infty}^{\infty} \sum_{r=1}^{l_j+1} \frac{\partial^2 p(x_j^r \mid x_k, \Sigma^*_{jk})/\partial \Sigma_{jk}^2}{p(x_j^r \mid x_k, \Sigma^*_{jk})} p(x_j^r \mid x_k, \Sigma^*_{jk}) p(x_k) dx_k \\
&\quad = \sum_{r=1}^{l_j+1} \partial^2 p(x_j^r \mid x_k, \Sigma^*_{jk})/\partial \Sigma_{jk}^2,
\end{aligned}
$$

with

$$
\begin{aligned}
&\sum_{r=1}^{l_j+1} \partial^2 p(x_j^r \mid x_k, \Sigma^*_{jk})/\partial \Sigma_{jk}^2 \\
&\quad = \sum_{r=1}^{l_j+1} \Bigg[\frac{3\Sigma_{jk}}{1-\Sigma_{jk}^2}(1-(\Sigma_{jk})^2)^{-\frac{3}{2}} \phi(\tilde{\Gamma}_j^r)(\Gamma_j^r \Sigma_{jk} - \tilde{x}_k) \\
&\quad + \frac{\phi'(\tilde{\Gamma}_j^r)(\Gamma_j^r \Sigma_{jk} - \tilde{x}_k)^2}{(1-\Sigma_{jk}^2)^3} + \frac{\phi(\tilde{\Gamma}_j^r)\Gamma_j^r}{(1-\Sigma_{jk}^2)^{-\frac{3}{2}}} \\
&\quad - \frac{3\Sigma_{jk}}{1-\Sigma_{jk}^2}(1-(\Sigma_{jk})^2)^{-\frac{3}{2}} \phi(\tilde{\Gamma}_j^{r-1})(\Gamma_j^{r-1} \Sigma_{jk} - \tilde{x}_k) \\
&\quad - \frac{\phi'(\tilde{\Gamma}_j^{r-1})(\Gamma_j^{r-1} \Sigma_{jk} - \tilde{x}_k)^2}{(1-\Sigma_{jk}^2)^3} - \frac{\phi(\tilde{\Gamma}_j^{r-1})\Gamma_j^{r-1}}{(1-\Sigma_{jk}^2)^{-\frac{3}{2}}} \Bigg] \\
&\quad = 0,
\end{aligned}
$$

since all terms except the ones involving $\phi(\tilde{\Gamma}_j^0)$ *and* $\phi(\tilde{\Gamma}_{l_j+1}^j)$ *cancel and furthermore* $\lim_{t\to\pm\infty}\phi(t) = \lim_{t\to\pm\infty}\phi'(t) = 0$.

Case III: *Similarly, consider*

$$
\begin{aligned}
&\mathbb{E}_{\Sigma_{jk}^*}\left[\frac{\partial^2\pi(x_j^r, x_s^k; \Sigma_{jk}^*)/\partial\Sigma_{jk}^2}{\pi(x_j^r, x_s^k; \Sigma_{jk}^*)}\right] \\
&= \sum_r\sum_s\left[\frac{\partial^2\pi(x_j^r, x_s^k; \Sigma_{jk}^*)/\partial\Sigma_{jk}^2}{\pi(x_j^r, x_s^k; \Sigma_{jk}^*)}P(X_j = x_j^r, X_k = x_k^s)\right] \\
&= \sum_r\sum_s\left[\partial^2\pi(x_j^r, x_s^k; \Sigma_{jk}^*)/\partial\Sigma_{jk}^2\right] \\
&= \sum_r\sum_s\Big[q(\Gamma_j^r, \Gamma_k^s, \Sigma_{jk}^*)\phi_2(\Gamma_j^r, \Gamma_k^s, \Sigma_{jk}^*) \\
&- q(\Gamma_j^{r-1}, \Gamma_k^s, \Sigma_{jk}^*)\phi_2(\Gamma_j^{r-1}, \Gamma_k^s, \Sigma_{jk}^*) \\
&- q(\Gamma_j^r, \Gamma_k^{s-1}, \Sigma_{jk}^*)\phi_2(\Gamma_j^r, \Gamma_k^{s-1}, \Sigma_{jk}^*) \\
&+ q(\Gamma_j^{r-1}, \Gamma_k^{s-1}, \Sigma_{jk}^*)\phi_2(\Gamma_j^{r-1}, \Gamma_k^{s-1}, \Sigma_{jk}^*)\Big] \\
&= q(\Gamma_j^{l_j+1}, \Gamma_k^{l_k+1}, \Sigma_{jk}^*)\phi_2(\Gamma_j^{l_j+1}, \Gamma_k^{l_k+1}, \Sigma_{jk}^*) \\
&- q(\Gamma_j^{l_j+1}, \Gamma_k^0, \Sigma_{jk}^*)\phi_2(\Gamma_j^{l_j+1}, \Gamma_k^0, \Sigma_{jk}^*) \\
&- q(\Gamma_j^0, \Gamma_k^{l_k+1}, \Sigma_{jk}^*)\phi_2(\Gamma_j^0, \Gamma_k^{l_k+1}, \Sigma_{jk}^*) \\
&+ q(\Gamma_j^0, \Gamma_k^0, \Sigma_{jk}^*)\phi_2(\Gamma_j^0, \Gamma_k^0, \Sigma_{jk}^*) \\
&= 0,
\end{aligned}
$$

with $q(s, t, \Sigma_{jk}^*))$ *denoting the corresponding quadratic function from the derivative of the bivariate normal density. As above,* $\Gamma_k^{l_k+1} = \infty$ *and* $\Gamma_k^0 = -\infty$ *for all* $k \in 1, \dots d_1$. *This, together with the fact that all other terms cancel when summing over* r, s, $\phi(\cdot)$ *is zero in all points containing* $\Gamma_k^{l_k+1}, \Gamma_k^0$ *and so the last equality follows.*

From this it follows, that $R''_{jk}(\Sigma_{jk}^*)$ *can only be zero if* $\partial p(x_j^r \mid x_k, \Sigma_{jk}^*)/\partial\Sigma_{jk}$ *for Case II and* $\partial\pi(\Sigma_{jk}^*)_{rs}/\partial\Sigma_{jk}$ *for Case III are zero. However, this is not possible due to Assumptions 3.2 and 3.3. To see this note that in Eq. (26)* $\partial p(x_j^r \mid x_k, \Sigma_{jk}^*)/\partial\Sigma_{jk}$ *can only be zero if either* $\Gamma_j^r = \Gamma_j^{r-1}$ *which we ruled out in the definition of the LGCM, or if* $|\Gamma_j^r| = |\Gamma_j^{r-1}| = \infty$ *which is ruled out by Assumption 3.2.We would not observe any discrete states in the first place if we had* $r = \{0, l_j + 1\}$. *Assumption 3.3 rules this case out. Consequently, there exist* $\epsilon > 0$ *and* $\eta > 0$ *such that* $R_{jk}(\Sigma_{jk})$ *is* (ϵ, η)*-strongly Morse*

With these considerations, we have verified the required four conditions to hold such that Theorem 2 in Mei, Bai and Montanari [26] holds for each couple (j, k) according to cases II and III. More precisely, let $\alpha \in (0, 1)$. Now, letting $n \geq 4C\log(n)\log(\frac{B}{\alpha})$ where $C = C_0\left(\frac{\tau^2}{\epsilon^2} \vee \frac{\tau^4}{\eta^2} \vee \frac{\tau^2 L^2}{\eta^4}\right)$ and $B = \tau(1-\delta)$

with $\tau = 2[\kappa L_1 \vee \xi K_1]$ and C_0 denoting a universal constant. Letting further $L = \sup_{\Sigma_{jk}:|\Sigma_{jk}|\leq 1-\delta} |R'''_{jk}(\Sigma_{jk})|$ we obtain

$$\mathbb{P}\left(\left|\hat{\Sigma}^{(n)}_{jk} - \Sigma^*_{jk}\right| \geq \frac{2\tau}{\eta}\sqrt{C_0 \frac{\log(n)}{n}\left[\log\left(\frac{B}{\alpha}\right) \vee 1\right]}\right) \leq \alpha, \tag{42}$$

and consequently, the result in Theorem 3.2 follows.

B.2. Proof of Lemmas B.5 to B.8

Lemma B.5. *For all $|\Sigma_{jk}| \leq 1-\delta$ and all $j \in [d_1], k \in [d_2]$ we have*

$$\int_S \frac{\partial}{\partial \Sigma_{jk}} p(x^r_j \mid x_k, \Sigma_{jk}) d\mu(x^r_j) = \frac{\partial}{\partial \Sigma_{jk}} \int_S p(x^r_j \mid x_k, \Sigma_{jk}) d\mu(x^r_j),$$

where μ is the counting measure on S, the corresponding discrete space.

Proof. Clearly, from Eq. (26) we have

$$\begin{aligned}
&\frac{\partial}{\partial \Sigma_{jk}} p(x^r_j \mid x_k, \Sigma_{jk}) \\
&\qquad = (1-(\Sigma_{jk})^2)^{-\frac{3}{2}}\Big[\phi(\tilde{\Gamma}^r_j)(\Gamma^r_j \Sigma_{jk} - \tilde{x}_k) - \phi(\tilde{\Gamma}^{r-1}_j)(\Gamma^{r-1}_j \Sigma_{jk} - \tilde{x}_k)\Big],
\end{aligned}$$

and therefore

$$\begin{aligned}
&\int_S (1-(\Sigma_{jk})^2)^{-\frac{3}{2}}\Big[\phi(\tilde{\Gamma}^r_j)(\Gamma^r_j \Sigma_{jk} - \tilde{x}_k) - \phi(\tilde{\Gamma}^{r-1}_j)(\Gamma^{r-1}_j \Sigma_{jk} - \tilde{x}_k)\Big] d\mu(x^r_j) \\
&\quad = (1-(\Sigma_{jk})^2)^{-\frac{3}{2}} \sum_{r=1}^{l_j} \Big[\phi(\tilde{\Gamma}^r_j)(\Gamma^r_j \Sigma_{jk} - \tilde{x}_k) - \phi(\tilde{\Gamma}^{r-1}_j)(\Gamma^{r-1}_j \Sigma_{jk} - \tilde{x}_k)\Big] \\
&\quad = 0 \\
&\quad = \frac{\partial}{\partial \Sigma_{jk}} \sum_{r=1}^{l_j} p(x^r_j \mid x_k, \Sigma_{jk}) \\
&\quad = \frac{\partial}{\partial \Sigma_{jk}} 1,
\end{aligned}$$

since all terms except the ones involving $\phi(\tilde{\Gamma}^0_j)$ and $\phi(\tilde{\Gamma}^{l_j+1}_j)$ cancel and

$$\lim_{t\to\pm\infty} \phi(t) = \lim_{t\to\pm\infty} \phi'(t) = 0,$$

and probabilities associated with all possible values must sum up to one. □

Corollary B.6. *For all $|\Sigma_{jk}| \leq 1-\delta$ and all $j \in [d_1], k \in [d_2]$ we have*

$$\int_S \frac{\partial^2}{\partial \Sigma^2_{jk}} p(x^r_j \mid x_k, \Sigma_{jk}) d\mu(x^r_j) = \frac{\partial^2}{\partial \Sigma^2_{jk}} \int_S p(x^r_j \mid x_k, \Sigma_{jk}) d\mu(x^r_j),$$

where μ is the counting measure on S, the corresponding discrete space.

Proof. From Eq. (38) we obtain

$$\begin{aligned}
&\frac{\partial^2}{\partial\Sigma_{jk}^2} p(x_j^r \mid x_k, \Sigma_{jk}) \\
&\quad= \frac{3\Sigma_{jk}}{1-\Sigma_{jk}^2}(1-(\Sigma_{jk})^2)^{-\frac{3}{2}}\phi(\tilde{\Gamma}_j^r)(\Gamma_j^r\Sigma_{jk}-\tilde{x}_k) \\
&\quad+ \frac{\phi'(\tilde{\Gamma}_j^r)(\Gamma_j^r\Sigma_{jk}-\tilde{x}_k)^2}{(1-\Sigma_{jk}^2)^3} + \frac{\phi(\tilde{\Gamma}_j^r)\Gamma_j^r}{(1-\Sigma_{jk}^2)^{-\frac{3}{2}}} \\
&\quad- \frac{3\Sigma_{jk}}{1-\Sigma_{jk}^2}(1-(\Sigma_{jk})^2)^{-\frac{3}{2}}\phi(\tilde{\Gamma}_j^{r-1})(\Gamma_j^{r-1}\Sigma_{jk}-\tilde{x}_k) \\
&\quad- \frac{\phi'(\tilde{\Gamma}_j^{r-1})(\Gamma_j^{r-1}\Sigma_{jk}-\tilde{x}_k)^2}{(1-\Sigma_{jk}^2)^3} - \frac{\phi(\tilde{\Gamma}_j^{r-1})\Gamma_j^{r-1}}{(1-\Sigma_{jk}^2)^{-\frac{3}{2}}}.
\end{aligned}$$

By similar arguments to Lemma B.5, when taking the sum over all possible states, all terms except the ones involving $\phi(\tilde{\Gamma}_j^0)$ and $\phi(\tilde{\Gamma}_j^{l_j+1})$ still cancel as they appear in every additive term in the above equation – recall that $\phi'(t) = -t\phi(t)$ – and equality then follows immediately. □

Lemma B.7. *For all $|\Sigma_{jk}| \leq 1-\delta$ we have*

1. $$\frac{\partial}{\partial\Sigma_{jk}}\mathbb{E}_{\Sigma_{jk}^*}\left[\ell_{jk}(\Sigma_{jk}, x_j^r, x_k)\right] = \mathbb{E}_{\Sigma_{jk}^*}\left[\frac{\partial}{\partial\Sigma_{jk}}\ell_{jk}(\Sigma_{jk}, x_j^r, x_k)\right],$$

   i.e.

   $$\begin{aligned}
   &\frac{\partial}{\partial\Sigma_{jk}}\int_{S\times\mathbb{R}} \log L_{jk}(\Sigma_{jk}, x_j^r, x_k) L_{jk}(\Sigma_{jk}^*, x_j^r, x_k) d\varepsilon(x_j^r, x_k) \\
   &\quad= \int_{S\times\mathbb{R}} \frac{\partial}{\partial\Sigma_{jk}} \log L_{jk}(\Sigma_{jk}, x_j^r, x_k) L_{jk}(\Sigma_{jk}^*, x_j^r, x_k) d\varepsilon(x_j^r, x_k)
   \end{aligned}$$

   where ε is the product measure on $S\times\mathbb{R}$ defined by

   $$\varepsilon \coloneqq \mu \otimes \lambda$$

   with μ denoting the counting measure on the corresponding discrete space S and λ the Lebesgue measure on the corresponding Euclidean space.

2. $$\frac{\partial}{\partial\Sigma_{jk}}\mathbb{E}_{\Sigma_{jk}^*}\left[\ell_{jk}(\Sigma_{jk}, x_j^r, x_k^s)\right] = \mathbb{E}_{\Sigma_{jk}^*}\left[\frac{\partial}{\partial\Sigma_{jk}}\ell_{jk}(\Sigma_{jk}, x_j^r, x_k^s)\right],$$

   i.e.

   $$\begin{aligned}
   &\frac{\partial}{\partial\Sigma_{jk}}\int_{S\times S'} \log L_{jk}(\Sigma_{jk}, x_j^r, x_k^s) L_{jk}(\Sigma_{jk}^*, x_j^r, x_k^s) d\varpi(x_j^r, x_k^s) \\
   &\quad= \int_{S\times S'} \frac{\partial}{\partial\Sigma_{jk}} \log L_{jk}(\Sigma_{jk}, x_j^r, x_k^s) L_{jk}(\Sigma_{jk}^*, x_j^r, x_k^s) d\varpi(x_j^r, x_k^s)
   \end{aligned}$$

where ϖ is the product measure on $S \times S'$ defined by

$$\varpi := \mu \otimes \mu'$$

with μ and μ' denoting the counting measure on the corresponding discrete space S and S', respectively.

Proof. Let us start with 1. and rewrite the right-hand side:

$$\begin{aligned}\int_{S\times\mathbb{R}} \frac{\partial}{\partial\Sigma_{jk}} \log L_{jk}(\Sigma_{jk}, x_j^r, x_k) L_{jk}(\Sigma_{jk}^*, x_j^r, x_k) d\varepsilon(x_j^r, x_k) \\ = \int_{\mathbb{R}} \sum_{r=1}^{l_j} \frac{\partial}{\partial\Sigma_{jk}} \log p(x_j^r, x_k, \Sigma_{jk}) p(x_j^r, x_k, \Sigma_{jk}^*) dx_k.\end{aligned}$$

The left-hand side corresponds to

$$\begin{aligned}\frac{\partial}{\partial\Sigma_{jk}} \int_{S\times\mathbb{R}} \log L_{jk}(\Sigma_{jk}, x_j^r, x_k) L_{jk}(\Sigma_{jk}^*, x_j^r, x_k) d\varepsilon(x_j^r, x_k) \\ = \frac{\partial}{\partial\Sigma_{jk}} \int_{\mathbb{R}} \sum_{r=1}^{l_j} \log p(x_j^r, x_k, \Sigma_{jk}) p(x_j^r, x_k, \Sigma_{jk}^*) dx_k.\end{aligned}$$

By Lebesgue's Dominated Convergence Theorem, we can interchange integration and differentiation as $\log p(x_j^r, x_k, \Sigma_{jk})$ is absolutely continuous s.t. its derivative exists almost everywhere and

$$\left| \frac{\partial \log p(x_j^r, x_k, \Sigma_{jk})}{\partial\Sigma_{jk}} \right|$$

is upper bounded by some integrable function. Indeed, the latter requirement has already been shown in Condition B.1. The second point follows by the same arguments where $\log p(x_j^r, x_k^s, \Sigma_{jk}) = \log(C) + \log(\pi_{rs})$ is absolutely continuous and bounded as shown in Condition B.1. This concludes the proof. □

Corollary B.8. *For all $|\Sigma_{jk}| \leq 1 - \delta$ we have*

$$\frac{\partial^2}{\partial\Sigma_{jk}^2} \mathbb{E}_{\Sigma_{jk}^*}\left[\ell_{jk}(\Sigma_{jk}, x_j^r, x_k)\right] = \mathbb{E}_{\Sigma_{jk}^*}\left[\frac{\partial^2}{\partial\Sigma_{jk}^2} \ell_{jk}(\Sigma_{jk}, x_j^r, x_k)\right], \text{ for Case II and}$$

$$\frac{\partial^2}{\partial\Sigma_{jk}^2} \mathbb{E}_{\Sigma_{jk}^*}\left[\ell_{jk}(\Sigma_{jk}, x_j^r, x_k^s)\right] = \mathbb{E}_{\Sigma_{jk}^*}\left[\frac{\partial^2}{\partial\Sigma_{jk}^2} \ell_{jk}(\Sigma_{jk}, x_j^r, x_k^s)\right], \text{ for case III.}$$

Proof. The claim follows immediately by the same arguments as in Lemma B.7 and the bound on the second derivative of the log-likelihood functions in Condition B.2, respectively. □

B.3. Proof of Lemma 3.1

First, note that $\Phi^{-1}(\cdot)$ is Lipschitz on $[\Phi(-G), \Phi(G)]$ with a Lipschitz constant $L_1 = \nabla\Phi^{-1}(\hat{\pi}_j^r) = 1/(\sqrt{\frac{2}{\pi}}\min\{\hat{\pi}_j^r, 1-\hat{\pi}_j^r\})$. Then we have

$$\begin{aligned}|\hat{\Gamma}_j^r - \Gamma_j^r| &= \Big|\Phi^{-1}\big(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(X_{ij} \le x_j^r)\big) - \Phi^{-1}\big(\Phi(\Gamma_j^r)\big)\Big| \\ &\le L_1\Big|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(X_{ij} \le x_j^r) - \Phi(\Gamma_j^r)\Big|.\end{aligned}$$

Applying Hoeffding's inequality, we obtain for some $t > 0$

$$\begin{aligned}P\Big(|\hat{\Gamma}_j^r - \Gamma_j^r| \ge t\Big) &\le P\Big(L_1\Big|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(X_{ij} \le x_j^r) - \Phi(\Gamma_j^r)\Big| \ge t\Big) \\ &\le 2\exp\Big(-\frac{2t^2 n}{L_1^2}\Big).\end{aligned}$$

This concludes the proof.

B.4. Proof of Theorem 3.3

In what follows, the proof of Theorem 3.3 revolves largely around the Winsorized estimator introduced in Section 3.2. Recall that

$$\hat{f}(x) = \Phi^{-1}(W_{\delta_n}[\hat{F}_{X_k}(x)])$$

where

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

with truncation constant $\delta_n = 1/(4n^{1/4}\sqrt{\pi \log n})$. Recall $f(x) = \Phi^{-1}(F_{X_k}(x))$, and let $g = f^{-1}$.

Assume w.l.o.g. that we have consecutive integer scoring in our discrete variable X_j such that the polyserial ad hoc estimator simplifies as

$$\hat{\Sigma}_{jk}^{(n)} = \frac{r_{\hat{f}(X_k),X_j}^{(n)}\sigma_{X_j}^{(n)}}{\sum_{r=1}^{l_j}\phi(\bar{\Gamma}_j^r)(x_j^{r+1} - x_j^r)} = \frac{r_{\hat{f}(X_k),X_j}^{(n)}\sigma_{X_j}^{(n)}}{\sum_{r=1}^{l_j}\phi(\bar{\Gamma}_j^r)} = \frac{S_{\hat{f}(X_k)X_j}}{\sigma_{\hat{f}(X_k)}^{(n)}\sum_{r=1}^{l_j}\phi(\bar{\Gamma}_j^r)}, \quad (43)$$

for all $j \in [d_1], k \in [d_2]$. $S_{\hat{f}(X_k)X_j}$ denotes the sample covariance between the $\hat{f}(X_k)$ and the X_j, i.e.

$$S_{\hat{f}(X_k)X_j} = \frac{1}{n}\sum_{i=1}^{n}\Big(\hat{f}(X_{ik}) - \mu_n(\hat{f})\Big)\Big(X_{ij} - \mu_n(X_j)\Big),$$

where $\mu_n(\hat{f}) = 1/n \sum_{i=1}^n \hat{f}(X_{ik})$ and $\mu_n(X_j) = 1/n \sum_{i=1}^n X_{ij}$. Moreover, $\sigma^{(n)}_{\hat{f}(X_k)}$ denotes the sample standard deviation of the Winsorized estimator, i.e.

$$\sigma^{(n)}_{\hat{f}(X_k)} \equiv \sqrt{\frac{1}{n}\sum_{i=1}^n \left(\hat{f}(X_{ik}) - \mu_n(\hat{f})\right)^2}.$$

Recall that we treat the threshold estimates as given. In particular, we have here $\phi(\bar{\Gamma}^r_j)$, therefore note further that $\phi(\cdot)$, namely the density function of the standard normal is Lipschitz with Lipschitz constant $L_0 = (2\pi)^{-1/2}\exp(-1/2)$, s.t.

$$\left|\phi(\bar{\Gamma}^r_j) - \phi(\Gamma^r_j)\right| \leq L_0 \left|\bar{\Gamma}^r_j - \Gamma^r_j\right| \leq \left|\bar{\Gamma}^r_j - \Gamma^r_j\right|,$$

as $L_0 < 1$. Consequently, the statements regarding the accuracy of the threshold estimates in Section 3.4 still hold here.

The outline of the proof is as follows: We start by forming concentration bounds for the sample covariance and the sample standard deviation separately. Then, we argue that the quotient of the two will be accurate in terms of a Lipschitz condition on the corresponding compactum. Let us start with the sample covariance. To study the Winsorized estimator, we consider the interval $[g(-\sqrt{M\log n}), g(\sqrt{M\log n})]$ for a choice of $M > 2$. As the behavior of the estimator is different for the endpoints, we further split this interval into a middle and an end part, respectively, i.e.,

$$\begin{aligned}
\mathbb{M}_n &\equiv (g(-\sqrt{\beta\log n}), g(\sqrt{\beta\log n})) \\
\mathbb{E}_n &\equiv [g(-\sqrt{M\log n}), g(-\sqrt{\beta\log n})) \cup (g(\sqrt{\beta\log n}), g(\sqrt{M\log n})].
\end{aligned}$$

This is only necessary for $\hat{f}(X_k)$ since $X_j \in 1, \ldots, l_j$ is discrete and can therefore only take finitely many values. Now consider the sample covariance, where we have for any $t > 0$ that

$$\begin{aligned}
&P\left(\max_{j,k}\left|S_{\hat{f}(X_k)X_j} - S_{f(X_k)X_j}\right| > 2t\right) \\
&\quad = P\Bigg(\max_{j,k}\Bigg|\frac{1}{n}\sum_{i=1}^n \Big[\hat{f}(X_{ik})X_{ij} - f(X_{ik})X_{ij} \\
&\qquad - \mu_n(\hat{f})\mu_n(X_j) + \mu_n(f)\mu_n(X_j)\Big]\Bigg| > 2t\Bigg) \\
&\quad \leq P\left(\max_{j,k}\left|\frac{1}{n}\sum_{i=1}^n \left[(\hat{f}(X_{ik}) - f(X_{ik}))X_{ij})\right]\right| > t\right) \\
&\qquad + P\left(\max_{j,k}\left|(\mu_n(\hat{f}) - \mu_n(f))\mu_n(X_j)\right| > t\right).
\end{aligned}$$

Let us take a closer look at the second term

$$
\begin{aligned}
&P\left(\max_{j,k}\left|(\mu_n(\hat{f})-\mu_n(f))\mu_n(X_j)\right|>t\right)\\
&\qquad=P\left(\max_{j,k}\left|\frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}(X_{ik})-f(X_{ik})\right)\frac{1}{n}\sum_{i=1}^{n}X_{ij}\right|>t\right)\\
&\qquad=P\left(\max_{k}\left|\frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}(X_{ik})-f(X_{ik})\right)\right|\max_{j}\left|\frac{1}{n}\sum_{i=1}^{n}X_{ij}\right|>t\right)\\
&\qquad\le P\left(\max_{k}\left|\frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}(X_{ik})-f(X_{ik})\right)\right|>\frac{t}{l_{\max}}\right),
\end{aligned}
$$

where X_j is a discrete random variable with finite level set and $l_{\max}\equiv\max_j l_j>0$.

Now, define

$$\triangle_i(j,k)\equiv(\hat{f}(X_{ik})-f(X_{ik}))X_{ij}$$

and

$$\tilde{\triangle}_{r,s}\equiv(\hat{f}(s)-f(s))r,$$

for $r=1,\ldots,l_j$. Furthermore, consider the event $\mathcal{A}_n$, where

$$\mathcal{A}_n\equiv\{g(-\sqrt{M\log n})\le X_{1k},\ldots,X_{nk}\le g(\sqrt{M\log n}),k=d_1+1,\ldots,d\}.$$

The bound for the complement arises from the Gaussian maximal inequality Liu, Lafferty and Wasserman [24, Lemma 13], i.e.,

$$P(\mathcal{A}_n^{\mathrm{c}})\le P\left(\max_{i,k\in\{1,\ldots,n\}\times\{d_1+1,\ldots,d\}}\left|f(X_{ik})\right|>\sqrt{2\log(nd_2)}\right)\le\frac{1}{2\sqrt{\pi\log(nd_2)}}.$$

The following lemma gives insight into the behavior of the Winsorized estimator along the end region.

Lemma B.9. *On the event $\mathcal{A}_n$, consider $\beta=\frac{1}{2}$, $t\ge C_M\sqrt{\frac{\log d_2 log^2 n}{n^{1/2}}}$ and $A=\sqrt{\frac{2}{\pi}}(\sqrt{M}-\sqrt{\beta})$, then*

$$P\left(\max_{j,k}\frac{1}{n}\sum_{X_k\in\mathbb{E}_n}\left|(\hat{f}(X_{ik})-f(X_{ik}))X_{ij}\right|>\frac{t}{2}\right)\le\exp\Big(-\frac{k_1n^{3/4}\sqrt{\log n}}{k_2+k_3}\Big),$$

and

$$P\left(\max_{k}\frac{1}{n}\sum_{X_k\in\mathbb{E}_n}\left|\hat{f}(X_{ik})-f(X_{ik})\right|>\frac{t}{2}\right)\le\exp\Big(-\frac{k_1n^{3/4}\sqrt{\log n}}{k_2+k_3}\Big),$$

where $k_i,i\in\{1,2,3\}$ are generic constants independent of sample size and dimension.

Proof. Let $\theta_1 \equiv \frac{n^{\beta/2}t}{4A\sqrt{\log n}}$ and let us first consider the bound for the first inequality.

$$
\begin{aligned}
&P\bigg(\max_{j,k}\frac{1}{n}\sum_{X_k\in\mathbb{E}_n}\Big|(\hat{f}(X_{ik})-f(X_{ik}))X_{ij}\Big|>\frac{t}{2}\bigg)\\
&\qquad= P\Big(\max_{j,k}\frac{1}{n}\sum_{i:X_{ik}\in\mathbb{E}_n}\Big|(\hat{f}(X_{ik})-f(X_{ik}))X_{ij}\Big|>\frac{t}{2}\\
&\qquad\qquad\cap\max_{jk}\sup_{r\in\{1,\dots,l_j\},s\in\mathbb{E}_n}\Big|\hat{f}(t)-f(t)\Big||r|>\theta_1\Big)\\
&\qquad+P\Big(\max_{j,k}\frac{1}{n}\sum_{i:X_{ik}\in\mathbb{E}_n}\Big|(\hat{f}(X_{ik})-f(X_{ik}))X_{ij}\Big|>\frac{t}{2}\\
&\qquad\qquad\cap\max_{jk}\sup_{r\in\{1,\dots,l_j\},s\in\mathbb{E}_n}\Big|\hat{f}(s)-f(s)\Big||r|\le\theta_1\Big)\\
&\le P\Big(\max_{jk}\sup_{r\in\{1,\dots,l_j\},s\in\mathbb{E}_n}\Big|\hat{f}(s)-f(s)\Big||r|>\theta_1\Big)+P\Big(\frac{1}{n}\sum_{i=1}^{n}1_{\{X_{ik}\in\mathbb{E}_n\}}>\frac{t}{2\theta_1}\Big).
\end{aligned}
$$

Similarly, for the bound of the second inequality, we have

$$
\begin{aligned}
&P\Big(\max_{k}\frac{1}{n}\sum_{i:X_{ik}\in\mathbb{E}_n}\Big|\hat{f}(X_{ik})-f(X_{ik})\Big|>\frac{t}{2}\Big)\\
&=P\Big(\max_{k}\frac{1}{n}\sum_{i:X_{ik}\in\mathbb{E}_n}\Big|\hat{f}(X_{ik})-f(X_{ik})\Big|>\frac{t}{2}\cap\max_{k}\sup_{s\in\mathbb{E}_n}\Big|\hat{f}(s)-f(s)\Big|>\theta_1\Big)\\
&+P\Big(\max_{k}\frac{1}{n}\sum_{i:X_{ik}\in\mathbb{E}_n}\Big|\hat{f}(X_{ik})-f(X_{ik})\Big|>\frac{t}{2}\cap\max_{k}\sup_{s\in\mathbb{E}_n}\Big|\hat{f}(s)-f(s)\Big|\le\theta_1\Big)\\
&\qquad\le P\Big(\max_{k}\sup_{s\in\mathbb{E}_n}\Big|\hat{f}(s)-f(s)\Big|>\theta_1\Big)+P\Big(\frac{1}{n}\sum_{i=1}^{n}1_{\{X_{ik}\in\mathbb{E}_n\}}>\frac{t}{2\theta_1}\Big).
\end{aligned}
$$

Recall that $\sup\{1,\dots,l_j\}=l_j>0$. Furthermore, Lemma 16 in [24] states that for all n

$$
\sup_{s\in\mathbb{E}_n}\Big|\hat{f}(s)-f(s)\Big|<\sqrt{2(M+2)\log n} \tag{44}
$$

With this in mind, we have

$$
P\Big(\max_{k}\sup_{s\in\mathbb{E}_n}\Big|\hat{f}(s)-f(s)\Big|>\theta_1\Big)\le d_2P\Big(\sup_{s\in\mathbb{E}_n}\Big|\hat{f}(s)-f(s)\Big|>\theta_1\Big).
$$

Recall that $C_M=8/\sqrt{\pi}(\sqrt{2M}-1)(M+2)$ and since $t\ge C_M\sqrt{\frac{\log d_2 log^2 n}{n^{1/2}}}$, we have

$$
\theta_1=\frac{n^{\beta/2}t}{4A\sqrt{\log n}}\ge\frac{C_M\sqrt{\log d_2 log^2 n}}{4A\sqrt{\log n}}=2(M+2)\log n.
$$

Consequently, we have

$$\theta_1 \geq 2(M+2)\log n \geq \sqrt{2(M+2)\log n},$$

as well as

$$\frac{\theta_1}{l_j} \geq \sqrt{2(M+2)\log n},$$

such that

$$P\Big(\sup_{t\in\mathbb{E}_n}\Big|\hat{f}(t)-f(t)\Big|>\theta_1\Big) = P\Big(\sup_{r\in\{1,\dots,l_j\},s\in\mathbb{E}_n}\Big|\hat{f}(s)-f(s)\Big||r|>\theta_1\Big) = 0.$$

In both cases, the second term is equivalent. We have

$$\begin{aligned}
P\Big(\frac{1}{n}\sum_{i=1}^{n}1_{\{X_{ik}\in\mathbb{E}_n\}} > \frac{t}{2\theta_1}\Big) &= P\Big(\sum_{i=1}^{n}1_{\{X_{ik}\in\mathbb{E}_n\}} > \frac{nt}{2\theta_1}\Big)\\
&= P\Big(\sum_{i=1}^{n}\big(1_{\{X_{ik}\in\mathbb{E}_n\}} - P(X_{1k}\in\mathbb{E}_n)\big) > \frac{nt}{2\theta_1} - P(X_{1k}\in n\mathbb{E}_n)\Big)\\
&\leq P\Big(\sum_{i=1}^{n}\big(1_{\{X_{ik}\in\mathbb{E}_n\}} - P(X_{1k}\in\mathbb{E}_n)\big) > \frac{nt}{2\theta_1} - nA\sqrt{\frac{\log n}{n^\beta}}\Big).
\end{aligned}$$

Choosing θ_1 this way guarantees that

$$\frac{nt}{2\theta_1} - nA\sqrt{\frac{\log n}{n^\beta}} = nA\sqrt{\frac{\log n}{n^\beta}} > 0.$$

Then, using Bernstein's inequality, we get

$$\begin{aligned}
P\Big(\frac{1}{n}\sum_{i=1}^{n}1_{\{X_{ik}\in\mathbb{E}_n\}} > \frac{t}{2\theta_1}\Big)&\\
&\leq P\Big(\sum_{i=1}^{n}\big(1_{\{X_{ik}\in\mathbb{E}_n\}} - P(X_{1k}\in\mathbb{E}_n)\big) > nA\sqrt{\frac{\log n}{n^\beta}}\Big)\\
&\leq \exp\Big(-\frac{k_1 n^{2-\beta}\log n}{k_2 n^{1-\beta/2}\sqrt{\log n} + k_3 n^{1-\beta/2}\sqrt{\log n}}\Big),
\end{aligned}$$

where $k_1, k_2, k_3 > 0$ are generic constants independent of n and d_2. Collecting terms finishes the proof. □

Turning back to the first decomposition of the sample covariance, we have

$$
\begin{aligned}
&P\left(\max_{j,k}\left|\frac{1}{n}\sum_{i=1}^{n}\Big[(\hat{f}(X_{ik})-f(X_{ik}))X_{ij})\Big]\right|>t\right)\\
&\qquad\qquad\leq P\left(\max_{j,k}\left|\frac{1}{n}\sum_{i=1}^{n}\triangle_i(j,k)\right|>t,\mathcal{A}_n\right)+P(\mathcal{A}_n^c)\\
&\qquad\qquad\leq P\left(\max_{j,k}\left|\frac{1}{n}\sum_{i=1}^{n}\triangle_i(j,k)\right|>t\cap\mathcal{A}_n\right)+\frac{1}{2\sqrt{\pi\log(nd_2)}}.
\end{aligned}
$$

Further, we have

$$
\begin{aligned}
&P\left(\max_{j,k}\left|\frac{1}{n}\sum_{i=1}^{n}\triangle_i(j,k)\right|>t\cap\mathcal{A}_n\right)\\
&\leq P\left(\max_{j,k}\frac{1}{n}\sum_{X_k\in\mathbb{M}_n}\big|\triangle_i(j,k)\big|>\frac{t}{2}\right)+P\left(\max_{j,k}\frac{1}{n}\sum_{X_k\in\mathbb{E}_n}\big|\triangle_i(j,k)\big|>\frac{t}{2}\right)\\
&+\frac{1}{2\sqrt{\pi\log(nd_2)}}\\
&\leq P\left(\max_{j,k}\frac{1}{n}\sum_{X_k\in\mathbb{M}_n}\big|\triangle_i(j,k)\big|>\frac{t}{2}\right)+\exp\Big(-\frac{k_1n^{3/4}\sqrt{\log n}}{k_2+k_3}\Big)\\
&+\frac{1}{2\sqrt{\pi\log(nd_2)}},
\end{aligned}
$$

where $X_k\in\mathbb{M}_n$ is shorthand notation for $i:X_{ik}\in\mathbb{M}_n$. We derive the bound of the second term in Lemma B.9. Thus, let us continue with the first term, where

$$
\begin{aligned}
&P\left(\max_{j,k}\frac{1}{n}\sum_{X_k\in\mathbb{M}_n}\big|\triangle_i(j,k)\big|>\frac{t}{2}\right)\\
&\qquad\qquad\leq d^2P\left(\sup_{r\in\{1,\ldots,l_j\},s\in\mathbb{M}_n}\left|\tilde{\triangle}_{r,s}\right|>\frac{t}{2}\right)\\
&\qquad\qquad=d^2P\left(\sup_{r\in\{1,\ldots,l_j\},s\in\mathbb{M}_n}\Big|(\hat{f}(s)-f(s))\Big||r|>\frac{t}{2}\right)\\
&\qquad\qquad=d^2P\left(\sup_{s\in\mathbb{M}_n}\Big|(\hat{f}(s)-f(s))\Big|>\frac{t}{2l_j}\right),
\end{aligned}
$$

where clearly $\sup\{1,\ldots,l_j\}=l_j>0$. Define the event

$$
\mathbb{B}_n\equiv\{\delta_n\leq\hat{F}_{X_k}(g_j(s))\leq1-\delta_n,\quad k\in[d_2]\}.
$$

Now, from the definition of the Winsorized estimator, we observe that

$$
\begin{aligned}
&d^2P\left(\sup_{s\in\mathbb{M}_n}\left|(\hat{f}(s)-f(s))\right|>\frac{t}{2l_j}\right)\\
&\quad\le d^2P\left(\sup_{s\in\mathbb{M}_n}\left|\Phi^{-1}(W_{\delta_n}[\hat{F}_{X_k}(s)])-\Phi^{-1}(F_{X_k}(s))\right|>\frac{t}{2l_j}\cap\mathbb{B}_n\right)+P(\mathbb{B}_n^c)\\
&\quad\le d^2P\left(\sup_{s\in\mathbb{M}_n}\left|\Phi^{-1}(\hat{F}_{X_k}(s))-\Phi^{-1}(F_{X_k}(s))\right|>\frac{t}{2l_j}\right)\\
&\quad+2\exp\left(2\log d-\frac{\sqrt{n}}{8\pi\log n}\right),
\end{aligned}
$$

where the expression for $P(\mathbb{B}_n^c))$ follows directly from Lemma 19 in [24]. Now, define

$$
\begin{aligned}
T_{1n}&\equiv\max\left\{F_{X_k}(g(\sqrt{\beta\log n})),1-\delta_n\right\}\\
T_{2n}&\equiv1-\min\left\{F_{X_k}(g(-\sqrt{\beta\log n})),\delta_n\right\},
\end{aligned}
$$

where it follows directly that $T_{1n}=T_{1n}=1-\delta_n$. Consequently, we apply the mean value theorem and get

$$
\begin{aligned}
&P\left(\sup_{s\in\mathbb{M}_n}\left|\Phi^{-1}(\hat{F}_{X_k}(s))-\Phi^{-1}(F_{X_k}(s))\right|>\frac{t}{2l_j}\right)\\
&\qquad\le P\left((\Phi^{-1})'(\max(T_{1n},T_{2n}))\sup_{s\in\mathbb{M}_n}\left|\hat{F}_{X_k}(s)-F_{X_k}(s)\right|>\frac{t}{2l_j}\right)\\
&\qquad=P\left((\Phi^{-1})'(1-\delta_n)\sup_{s\in\mathbb{M}_n}\left|\hat{F}_{X_k}(s)-F_{X_k}(s)\right|>\frac{t}{2l_j}\right)\\
&\qquad\le P\left(\sup_{s\in\mathbb{M}_n}\left|\hat{F}_{X_k}(s)-F_{X_k}(s)\right|>\frac{t}{(\Phi^{-1})'(1-\delta_n)2l_j}\right)\\
&\qquad\le 2\exp\left(-2\frac{nt^2}{4l_j^2[(\Phi^{-1})'(1-\delta_n)]^2}\right),
\end{aligned}
$$

where the last inequality arises from applying the Dvoretzky-Kiefer-Wolfowitz inequality. Now, we have that

$$
\begin{aligned}
(\Phi^{-1})'(1-\delta_n)&=\frac{1}{\phi\big(\Phi^{-1}(1-\delta_n)\big)}\\
&\le\frac{1}{\phi\Big(\sqrt{2\log\frac{1}{\delta_n}}\Big)}=\sqrt{2\pi}\Big(\frac{1}{\delta_n}\Big)=8\pi n^{\beta/2}\sqrt{\beta\log n}.
\end{aligned}
$$

Therefore,

$$d^2 P\left(\sup_{s\in\mathbb{M}_n} \left|\Phi^{-1}(\hat{F}_{X_k}(s)) - \Phi^{-1}(F_{X_k}(s))\right| > \frac{t}{2l_j} \right) \\ \leq 2\exp\left(2\log d - \frac{\sqrt{n}t^2}{64l_j^2\pi^2\log n}\right).$$

Collecting the remaining terms we have

$$P\left(\max_{j,k} \frac{1}{n} \sum_{X_k\in\mathbb{M}_n} \left|\triangle_i(j,k)\right| > \frac{t}{2} \right) \\ \leq 2\exp\left(2\log d - \frac{\sqrt{n}t^2}{64l_j^2\pi^2\log n}\right) + 2\exp\left(2\log d - \frac{\sqrt{n}}{8\pi\log n}\right).$$

Thus, we have for the first term in the covariance matrix decomposition

$$\begin{aligned} P\left(\max_{j,k} \left| \frac{1}{n}\sum_{i=1}^{n} \left[(\hat{f}(X_{ik}) - f(X_{ik}))X_{ij} \right] \right| > t \right) \\ \leq P\left(\max_{j,k} \frac{1}{n} \sum_{X_k\in\mathbb{M}_n} \left|\triangle_i(j,k)\right| > \frac{t}{2} \right) \\ + \exp\Big(- \frac{k_1 n^{3/4}\sqrt{\log n}}{k_2+k_3}\Big) + \frac{1}{2\sqrt{\pi\log(nd_2)}} \\ \leq 2\exp\left(2\log d - \frac{\sqrt{n}t^2}{64l_j^2\pi^2\log n}\right) + 2\exp\left(2\log d - \frac{\sqrt{n}}{8\pi\log n}\right) \\ + \exp\Big(- \frac{k_1 n^{3/4}\sqrt{\log n}}{k_2+k_3}\Big) + \frac{1}{2\sqrt{\pi\log(nd_2)}}. \end{aligned}$$

Let us turn back to the second term of the first sample covariance decomposition, i.e.

$$P\left(\max_{j,k} \left|(\mu_n(\hat{f}) - \mu_n(f))\mu_n(X_j)\right| > t \right) \\ \leq P\left(\max_{k} \left| \frac{1}{n}\sum_{i=1}^{n} \Big(\hat{f}(X_{ik}) - f(X_{ik})\Big) \right| > \frac{t}{l_{\max}} \cap \mathcal{A}_n \right) + \frac{1}{2\sqrt{\pi\log(nd_2)}}.$$

Now, analogous to before, we find

$$\begin{aligned}
P\Big(\max_k \frac{1}{n}\sum_{i=1}^{n}\Big|\hat{f}(X_{ik}) - f(X_{ik})\Big| > \frac{t}{l_{\max}} \cap \mathcal{A}_n\Big) \\
&\le P\Big(\max_k \frac{1}{n}\sum_{X_k\in\mathbb{M}_n}\Big|\hat{f}(X_{ik}) - f(X_{ik})\Big| > \frac{t}{2l_{\max}}\Big) \\
&+ P\Big(\max_k \frac{1}{n}\sum_{X_k\in\mathbb{E}_n}\Big|\hat{f}(X_{ik}) - f(X_{ik})\Big| > \frac{t}{2l_{\max}}\Big) \\
&\le P\Big(\max_k \frac{1}{n}\sum_{X_k\in\mathbb{M}_n}\Big|\hat{f}(X_{ik}) - f(X_{ik})\Big| > \frac{t}{2l_{\max}}\Big) \\
&+ \exp\Big(-\frac{k_1 n^{3/4}\sqrt{\log n}}{k_2+k_3}\Big) \\
&\le d_2 P\Big(\sup_{t\in\mathbb{M}_n}\Big|\hat{f}(t) - f(t)\Big| > \frac{t}{2l_{\max}}\Big) \\
&+ \exp\Big(-\frac{k_1 n^{3/4}\sqrt{\log n}}{k_2+k_3}\Big).
\end{aligned} \tag{45}$$

Let us take a closer look at

$$\begin{aligned}
d_2 P\Big(\sup_{t\in\mathbb{M}_n}\Big|\hat{f}(t) - f(t)\Big| > \frac{t}{2l_{\max}}\Big) \\
&\le d_2 P\left(\sup_{t\in\mathbb{M}_n}\Big|\Phi^{-1}(W_{\delta_n}[\hat{F}_{X_k}(t)]) - \Phi^{-1}(F_{X_k}(t))\Big| > \frac{t}{2l_{\max}} \cap \mathbb{B}_n\right) \\
&+ d_2 P(\mathbb{B}_n^c) \\
&\le d_2 P\left(\sup_{t\in\mathbb{M}_n}\Big|\Phi^{-1}(\hat{F}_{X_k}(t)) - \Phi^{-1}(F_{X_k}(t))\Big| > \frac{t}{2l_{\max}}\right) \\
&+ 2\exp\Big(\log d_2 - \frac{\sqrt{n}}{8\pi\log n}\Big).
\end{aligned}$$

The definition of the event $\mathbb{B}_n$ is the same as above. Then applying once more the Dvoretzky–Kiefer–Wolfowitz inequality we end up with the following upper bound:

$$\begin{aligned}
d_2 P\left(\sup_{t\in\mathbb{M}_n}\Big|\Phi^{-1}(\hat{F}_{X_k}(t)) - \Phi^{-1}(F_{X_k}(t))\Big| > \frac{t}{2l_{\max}}\right) \\
&\le 2\exp\Big(\log d_2 - \frac{\sqrt{n}t^2}{64\ l_{\max}^2\ \pi^2\log n}\Big).
\end{aligned}$$

Collecting terms and simplifying yields

$$
\begin{aligned}
&P\Bigg(\max_{j,k} \Bigg| \frac{1}{n} \sum_{i=1}^{n} \Big[(\hat{f}(X_{ik}) - f(X_{ik}))X_{ij}) \Big] \Bigg| > t \Bigg) \\
&\qquad + P\Bigg(\max_{j,k} \Big| (\mu_n(\hat{f}) - \mu_n(f))\mu_n(X_j) \Big| > t \Bigg) \\
&\qquad \leq 2\exp\left(2\log d - \frac{\sqrt{n}t^2}{64\ l_j^2\ \pi^2 \log n} \right) + 2\exp\Big(2\log d - \frac{\sqrt{n}}{8\pi \log n} \Big) \\
&\qquad + \exp\Big(- \frac{k_1 n^{3/4}\sqrt{\log n}}{k_2 + k_3} \Big) + \frac{1}{2\sqrt{\pi \log(nd_2)}} \\
&\qquad + 2\exp\left(\log d_2 - \frac{\sqrt{n}t^2}{64\ l_{\max}^2\ \pi^2 \log n} \right) + 2\exp\Big(\log d_2 - \frac{\sqrt{n}}{8\pi \log n} \Big) \\
&\qquad + \exp\Big(- \frac{k_1 n^{3/4}\sqrt{\log n}}{k_2 + k_3} \Big) + \frac{1}{2\sqrt{\pi \log(nd_2)}} \\
&\qquad \leq 4\exp\left(2\log d - \frac{\sqrt{n}t^2}{64\ l_{\max}^2\ \pi^2 \log n} \right) + 4\exp\Big(2\log d - \frac{\sqrt{n}}{8\pi \log n} \Big) \\
&\qquad + 2\exp\Big(- \frac{k_1 n^{3/4}\sqrt{\log n}}{k_2 + k_3} \Big) + \frac{1}{\sqrt{\pi \log(nd_2)}}.
\end{aligned}
$$

Then

$$
\begin{aligned}
&P\Bigg(\max_{j,k} \Big| S_{\hat{f}(X_k)X_j} - S_{f(X_k)X_j} \Big| > 2t \Bigg) \\
&\qquad \leq 4\exp\left(2\log d - \frac{\sqrt{n}t^2}{64\ l_{\max}^2\ \pi^2 \log n} \right) + 4\exp\Big(2\log d - \frac{\sqrt{n}}{8\pi \log n} \Big) \\
&\qquad + 2\exp\Big(- \frac{k_1 n^{3/4}\sqrt{\log n}}{k_2 + k_3} \Big) + \frac{1}{\sqrt{\pi \log(nd_2)}},
\end{aligned}
\tag{46}
$$

which completes the considerations regarding the sample covariance.

As a next step, we need to bound the error of the sample standard deviation of the Winsorized estimator

$$
\sigma_{\hat{f}(X_k)}^{(n)} \equiv \sqrt{\frac{1}{n} \sum_{i=1}^{n} \Big(\hat{f}(X_{ik}) - \mu_n(\hat{f}) \Big)^2}.
$$

Consider the following decomposition of the standard deviation of the Win-

sorized estimator,

$$
\begin{aligned}
&\left|\sigma^{(n)}_{\hat{f}(X_k)} - \sigma^{(n)}_{f(X_k)}\right| \\
&\quad= \left|\sqrt{1/n\sum_{i=1}^{n}\Big(\hat{f}(X_{ik}) - \mu_n(\hat{f})\Big)^2} - \sqrt{1/n\sum_{i=1}^{n}\Big(f(X_{ik}) - \mu_n(f)\Big)^2}\right| \\
&\quad= \frac{1}{\sqrt{n}}\left|\Big(\|\hat{f}(X_k) - \mu_n(\hat{f})\|_2 - \|f(X_k) - \mu_n(f)\|_2\Big)\right| \\
&\quad\leq \frac{1}{\sqrt{n}}\|\hat{f}(X_k) - \mu_n(\hat{f}) - f(X_k) + \mu_n(f)\|_2 \\
&\quad\leq \frac{1}{\sqrt{n}}\sqrt{n}\|\hat{f}(X_k) - \mu_n(\hat{f}) - f(X_k) + \mu_n(f)\|_\infty \\
&\quad= \sup_{i:X_{ik}\in\{1,\ldots,n\}}\left|\hat{f}(X_{ik}) - f(X_{ik}) + \mu_n(f) - \mu_n(\hat{f})\right| \\
&\quad= \sup_{i:X_{ik}\in\{1,\ldots,n\}}\left|\hat{f}(X_{ik}) - f(X_{ik})\right| + \left|\mu_n(\hat{f}) - \mu_n(f)\right| \\
&\quad\leq \sup_{i:X_{ik}\in\{1,\ldots,n\}}\left|\hat{f}(X_{ik}) - f(X_{ik})\right| + \frac{1}{n}\sum_{i=1}^{n}\left|\hat{f}(X_{ik}) - f(X_{ik})\right|.
\end{aligned}
$$

The first inequality is due to the reverse triangle inequality. The ensuing inequalities arise from applying standard norm equivalences. As before, we analyze both terms separately since we have to take care of the behavior of the Winsorized estimator, taking values at the end and the middle interval. We have for any $t > 0$,

$$
\begin{aligned}
&P\left(\max_k\left|\sigma^{(n)}_{\hat{f}(X_k)} - \sigma^{(n)}_{f(X_k)}\right| > 2t\right) \\
&\qquad\leq P\Big(\max_k \sup_{i:X_{ik}\in\{1,\ldots,n\}}\left|\hat{f}(X_{ik}) - f(X_{ik})\right| > t, \mathcal{A}_n\Big) \\
&\qquad\quad + P\Big(\max_k \frac{1}{n}\sum_{i=1}^{n}\left|\hat{f}(X_{ik}) - f(X_{ik})\right| > t, \mathcal{A}_n\Big) + P(\mathcal{A}_n^c).
\end{aligned}
$$

Note that the second term is, in effect, equivalent to Eq. (45) above such that

we can immediately conclude that

$$
\begin{aligned}
&P\Big(\max_k \frac{1}{n}\sum_{i=1}^{n}\Big|\hat{f}(X_{ik}) - f(X_{ik})\Big| > t \cap \mathcal{A}_n\Big) \\
&\qquad \leq d_2 P\Big(\sup_{t\in\mathbb{M}_n}\Big|\hat{f}(t) - f(t)\Big| > \frac{t}{2}\Big) + \exp\Big(-\frac{k_1 n^{3/4}\sqrt{\log n}}{k_2 + k_3}\Big) \\
&\qquad \leq 2\exp\Big(\log d_2 - \frac{\sqrt{n}t^2}{64\pi^2 \log n}\Big) + 2\exp\Big(\log d_2 - \frac{\sqrt{n}}{8\pi \log n}\Big) \\
&\qquad + \exp\Big(-\frac{k_1 n^{3/4}\sqrt{\log n}}{k_2 + k_3}\Big).
\end{aligned}
$$

The bound for the end region follows again from Lemma B.9.

Similarly, we find that

$$
\begin{aligned}
&P\Big(\max_k \sup_{i:X_{ik}\in\{1,\ldots,n\}}\Big|\hat{f}(X_{ik}) - f(X_{ik})\Big| > t \cap \mathcal{A}_n\Big) \\
&\qquad \leq d_2 P\Big(\sup_{t\in\mathbb{M}_n}\Big|\hat{f}(t) - f(t)\Big| > \frac{t}{2}\Big) + d_2 P\Big(\sup_{t\in\mathbb{E}_n}\Big|\hat{f}(t) - f(t)\Big| > \frac{t}{2}\Big) \\
&\qquad = d_2 P\Big(\sup_{t\in\mathbb{M}_n}\Big|\hat{f}(t) - f(t)\Big| > \frac{t}{2}\Big),
\end{aligned}
$$

where the bound over the end region has been shown in Lemma B.9. Thus, we only have to take care of

$$
\begin{aligned}
&P\Big(\sup_{t\in\mathbb{M}_n}\Big|\hat{f}(t) - f(t)\Big| > \frac{t}{2}\Big) \\
&\leq P\left(\sup_{t\in\mathbb{M}_n}\Big|\Phi^{-1}(W_{\delta_n}[\hat{F}_{X_k}(t)]) - \Phi^{-1}(F_{X_k}(t))\Big| > \frac{t}{2} \cap \mathbb{B}_n\right) + P(\mathbb{B}_n^c) \\
&\leq P\left(\sup_{t\in\mathbb{M}_n}\Big|\Phi^{-1}(\hat{F}_{X_k}(t)) - \Phi^{-1}(F_{X_k}(t))\Big| > \frac{t}{2}\right) + 2\exp\Big(\log d_2 - \frac{\sqrt{n}}{8\pi \log n}\Big).
\end{aligned}
$$

The definition of the event $\mathbb{B}_n$ is the same as above. Then again, by the Dvoretzky-Kiefer-Wolfowitz inequality, we end up with the following upper bound:

$$
P\left(\sup_{t\in\mathbb{M}_n}\Big|\Phi^{-1}(\hat{F}_{X_k}(t)) - \Phi^{-1}(F_{X_k}(t))\Big| > \frac{t}{2}\right) \leq 2\exp\Big(-\frac{\sqrt{n}t^2}{64\pi^2 \log n}\Big).
$$

Collecting terms, we arrive at the concentration bound for the sample standard

deviation:

$$P\left(\max_k\left|\sigma^{(n)}_{\hat{f}(X_k)}-\sigma^{(n)}_{f(X_k)}\right|>2t\right)$$
$$\leq 4\exp\Big(\log d_2-\frac{\sqrt{n}t^2}{64\pi^2\log n}\Big)+4\exp\Big(\log d_2-\frac{\sqrt{n}}{8\pi\log n}\Big)$$
$$+\exp\Big(-\frac{k_1 n^{3/4}\sqrt{\log n}}{k_2+k_3}\Big)+\frac{1}{2\sqrt{\pi\log(nd_2)}}.$$

With these intermediate results, we have shown that the sample covariance (numerator) and the sample standard deviation (denominator) can be estimated accurately.

The following lemma provides us with the means to form a probability bound for

$$\max_{jk}\left|\hat{\Sigma}^{(n)}_{jk}-\Sigma^{*}_{jk}\right|.$$

Lemma B.10. *Consider the polyserial ad hoc estimator $\hat{\Sigma}^{(n)}_{jk}$ for $1\leq j\leq d_1<k\leq d_2$ and let $\epsilon\in\Big[C_M\sqrt{\frac{\log d_2 log^2 n}{n^{1/2}}},8(1+4c^2)\Big]$, where c is the corresponding sub-Gaussian parameter of the discrete variable. Both the numerator and the denominator are bounded, i.e.*

$$S_{\hat{f}(X_k)X_j}\in[-(1+\epsilon),1+\epsilon],$$

and

$$\sigma^{(n)}_{\hat{f}(X_k)}\in[1-\epsilon,1+\epsilon].$$

Consequently, $\hat{\Sigma}^{(n)}_{jk}$ is Lipschitz with constant L. The following decomposition holds

$$\max_{jk}\left|\hat{\Sigma}^{(n)}_{jk}-\Sigma^{*}_{jk}\right|=\max_{jk}\left|\hat{\Sigma}^{(n)}_{jk}-\Sigma^{(n)}_{jk}+\Sigma^{(n)}_{jk}-\Sigma^{*}_{jk}\right|$$
$$\leq\max_{jk}\left|\hat{\Sigma}^{(n)}_{jk}-\Sigma^{(n)}_{jk}\right|+\max_{jk}\left|\Sigma^{(n)}_{jk}-\Sigma^{*}_{jk}\right|$$
$$\leq L\Big(\max_{jk}\left|S_{\hat{f}(X_k)X_j}-S_{f(X_k)X_j}\right|+C_\Gamma\max_k\left|\sigma^{(n)}_{\hat{f}(X_k)}-\sigma^{(n)}_{f(X_k)}\right|$$
$$+\max_{jk}\left|S_{f(X_k)X_j}-S^{*}_{f(X_k)X_j}\right|+C_\Gamma\max_k\left|\sigma^{(n)}_{f(X_k)}-1\right|\Big),$$

where $C_\Gamma\equiv\sum_{r=1}^{l_j}\phi(\bar{\Gamma}^r_j)(x^{r+1}_j-x^r_j)$.

Proof. Let us assume w.l.o.g. that X_j - the discrete variable - has mean zero and variance one. By the Cauchy-Schwarz inequality, the true covariance of the pair is bounded from above by 1, i.e.

$$\left|S^{*}_{f(X_k)X_j}\right|\leq\sigma^2_{f(X_k)}\sigma^2_{X_j}=1.$$

Earlier we have shown that for some $t \geq C_M \sqrt{\frac{\log d_2 log^2 n}{n^{1/2}}}$ we have

$$\begin{aligned} P\Bigg(\max_{j,k} \Big| S_{\hat{f}(X_k)X_j} - S_{f(X_k)X_j} \Big| > 2t \Bigg) \\ \leq 4 \exp\left(2 \log d - \frac{\sqrt{n} t^2}{64\ l_{\max}^2\ \pi^2 \log n} \right) + 4 \exp\left(2 \log d - \frac{\sqrt{n}}{8\pi \log n} \right) \\ + 2 \exp\Big(- \frac{k_1 n^{3/4} \sqrt{\log n}}{k_2 + k_3} \Big) + \frac{1}{\sqrt{\pi \log(n d_2)}}. \end{aligned} \tag{47}$$

According to Lemma 1 in [37] with sub-Gaussian parameter $c > 0$, $f(X_k)$ is standard Gaussian and thus sub-Gaussian and X_j is discrete and bounded and therefore also sub-Gaussian we have the following tail bound

$$P\left(\max_{jk} \Big| S_{f(X_k)X_j} - S^*_{f(X_k)X_j} \Big| \geq t \right) \leq 4d^2 \exp\left\{ - \frac{nt^2}{128(1+4c^2)^2} \right\},$$

for all $t \in (0, 8(1+4c^2))$. Therefore with high probability for $j \in [d_1], k \in [d_2]$, we have

$$S_{\hat{f}(X_k)X_j} \in [S_{f(X_k)X_j} - 2t, S_{f(X_k)X_j} + 2t],$$

and since

$$S_{f(X_k)X_j} \in [-1-t, 1+t],$$

with high probability, we have

$$S_{\hat{f}(X_k)X_j} \in [-(1+3t), 1+3t].$$

Similar considerations hold for the sample standard deviation. We have already shown that

$$\begin{aligned} P\left(\max_{k} \Big| \sigma^{(n)}_{\hat{f}(X_k)} - \sigma^{(n)}_{f(X_k)} \Big| > 2t \right) \\ \leq 4 \exp\Big(\log d_2 - \frac{\sqrt{n} t^2}{64\pi^2 \log n} \Big) + 4 \exp\Big(\log d_2 - \frac{\sqrt{n}}{8\pi \log n} \Big) \\ + \exp\Big(- \frac{k_1 n^{3/4} \sqrt{\log n}}{k_2 + k_3} \Big) + \frac{1}{2\sqrt{\pi \log(n d_2)}}. \end{aligned}$$

Furthermore, we use Lemma 1 again in [37] to form a bound for the variance. Since $f(X_k)$ is standard Gaussian and hence sub-Gaussian with parameter $c = 1$ we immediately get

$$P\left(\max_{k} \Big| (\sigma^{(n)}_{f(X_k)})^2 - 1 \Big| \geq t \right) \leq 4 d_2 \exp\left\{ - \frac{nt^2}{128(1+4)^2} \right\}.$$

Put differently, with high probability

$$(\sigma^{(n)}_{f(X_k)})^2 \in [1-t, 1+t],$$

and consequently, we also have with high probability

$$\sigma^{(n)}_{f(X_k)} \in [\sqrt{1-t}, \sqrt{1+t}].$$

Since the interval $[1-t, 1+t]$ is always as least as wide as $[\sqrt{1-t}, \sqrt{1+t}]$, for all $t > 0$ with high probability we then also have

$$\sigma^{(n)}_{f(X_k)} \in [1-t, 1+t].$$

Putting these things together, we obtain that with high probability

$$\sigma^{(n)}_{\hat{f}(X_k)} \in [1-3t, 1+3t].$$

In order to finish the proof consider the following function $h : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ defined by

$$h(u, v) = \frac{u}{v},$$

with $\nabla h = (1/v, -u/v^2)^T$.

As we have just shown for the polyserial ad hoc estimator, $\nabla h = (1/v, -u/v^2)^T$ is bounded with

$$\sup \|\nabla h\|_2 = \sqrt{\left(\frac{1}{1-3t}\right)^2 + \left(-\frac{(1+3t)}{(1-3t)^2}\right)^2} \coloneqq L.$$

Consequently, h is Lipschitz, and we have the following decomposition

$$\begin{aligned} \big|h(u,v) - h(u',v')\big| &= \big|h(u,v) - h(\tilde{u},\tilde{v}) + h(\tilde{u},\tilde{v}) - h(u',v')\big| \\ &\leq \big|h(u,v) - h(\tilde{u},\tilde{v})\big| + \big|h(\tilde{u},\tilde{v}) - h(u',v')\big| \\ &\leq L\big(|u-\tilde{u}| + |v-\tilde{v}|\big) + L\big(\big|\tilde{u}-u'\big| + \big|\tilde{v}-v'\big|\big). \end{aligned}$$

Finally, taking $\epsilon = 3t$ finishes the proof. □

At last, collecting terms, we find that for $j \in [d_1]$ and $k[d_2]$ and any

$$\epsilon \in \left[C_M \sqrt{\frac{\log d \log^2 n}{\sqrt{n}}}, 8(1+4c^2)\right]$$

the following bound holds

$$
\begin{aligned}
&P\left(\max_{jk}\left|\hat{\Sigma}^{(n)}_{jk}-\Sigma^{*}_{jk}\right|\geq\epsilon\right)\\
&\quad\leq P\left(\max_{j,k}\left|S_{\hat{f}(X_k)X_j}-S_{f(X_k)X_j}\right|>\frac{\epsilon}{4L}\right)\\
&\quad\quad+P\left(\max_{k}\left|\sigma^{(n)}_{\hat{f}(X_k)}-\sigma^{(n)}_{f(X_k)}\right|>\frac{\epsilon}{4LC_\Gamma}\right)\\
&\quad\quad+P\left(\max_{jk}\left|S_{f(X_k)X_j}-S^{*}_{f(X_k)X_j}\right|\geq\frac{\epsilon}{4L}\right)\\
&\quad\quad+P\left(\max_{k}\left|(\sigma^{(n)}_{f(X_k)})^2-1\right|\geq\frac{\epsilon}{4lC_\Gamma}\right)
\end{aligned}
$$

The conclusion of Theorem 3.3 follows by plugging in the corresponding concentration bounds and simplifying.

Appendix C: Additional simulation setup and results

C.1. FPR and TPR results for binary and general mixed data

This section provides additional simulation results for the binary-continuous and general mixed data setting. In particular, we report TPR and FPR for the latent Gaussian model and the LGCM with $f_j(x) = x^{1/3}$ for all j. The results are depicted in Figures 5 and 6 for the binary-continuous and general mixed data settings, respectively.

Figure 5 illustrates that our `poly` estimator performs almost identical to the gold standard `bridge` estimator proposed by Fan et al. [12]. In the right column, the `mle` estimator is misspecified, and TPR is noticeably lower than for the remaining estimators. Interestingly, it appears that the misspecified `mle` estimator does not infer additional edges, which is reflected in the FPR holding level with the other estimators.

The general mixed data setting results reported in Figure 6 are similar. The `poly` estimator performs almost identically to the `mle` estimator under the latent Gaussian model in terms of FPR and TPR. The `bridge` estimator has higher FPR and TPR in the $d = 50$ case and lower FPR and TPR in the $d = 750$ case. Under the LGCM, similar to the binary-continuous data setting, the `mle` estimator is misspecified and has a lower TPR than the other estimators while the FPR holds level. The `bridge` ensemble estimator has an even higher FPR and TPR in the $d = 50$ case and a slightly lower FPR and TPR in the $d = 750$ case.

C.2. Ternary mixed data results

We additionally compare our `poly` and `mle` estimators with a generalization of the bridge function approach proposed by Quan, Booth and Wells [35] given a

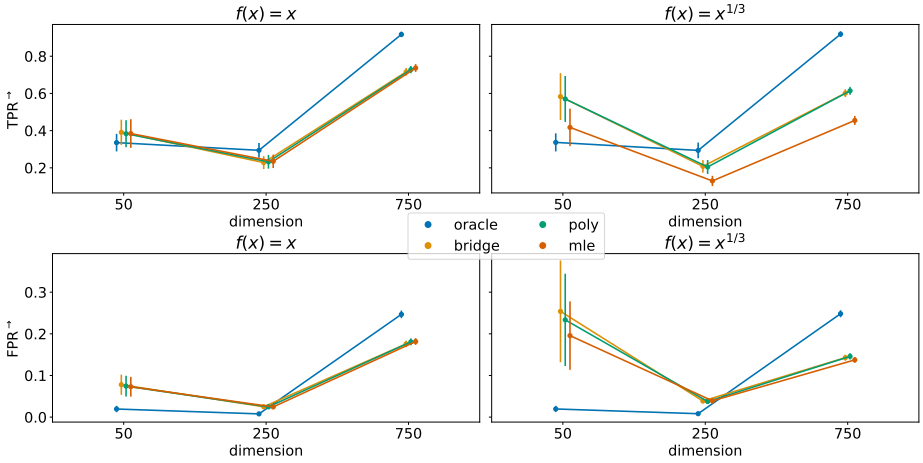

FIG 5. *Simulation results for the binary-continuous data setting based on* 100 *simulation runs. The left column corresponds to the latent Gaussian model, where the transformation function is the identity. The right colum depicts results for the LGCM with* $f_j(x) = x^{1/3}$ *for all* j*. The top and bottom rows report mean and standard deviation of the TPR and FPR along simulation runs, respectively. The y-axis labels have superscript arrows attached to indicate the direction of improvement:* $\rightarrow$ *implies that larger values are better.*

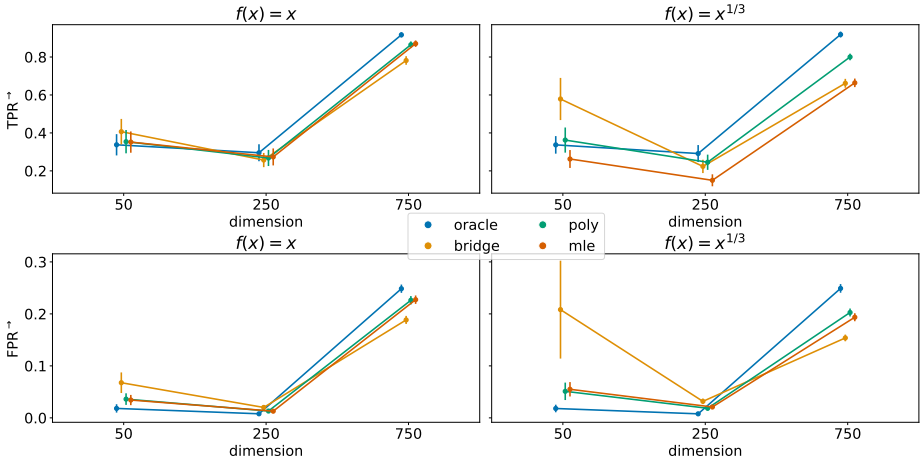

FIG 6. *Simulation results for the general mixed data setting based on* 100 *simulation runs. The left column corresponds to the latent Gaussian model, where the transformation function is the identity. The right colum depicts results for the LGCM with* $f_j(x) = x^{1/3}$ *for all* j*. The top and bottom rows report mean and standard deviation of the TPR and FPR along simulation runs, respectively. The y-axis labels have superscript arrows attached to indicate the direction of improvement:* $\rightarrow$ *implies that larger values are better.*

mix of ternary, binary, and continuous data. In this case, $\binom{2+2}{2} = 6$ individual bridge functions are needed.

The (d, n)-setup is analogous to the binary-continuous data setting from the main text. Starting with `mle`, the pattern from the binary-continuous setting continues to show in the ternary-binary-continuous mix. Whenever $f(x) = x$, the `mle` estimator generally performs best, in particular concerning graph recovery.

Table 1: Ternary mixed data structure learning; Simulated data with 100 simulation runs. Standard errors in brackets

$d, n, f(x)$		Oracle $\hat{\Omega}$	ternary $\hat{\Omega}_\tau$	$\hat{\Omega}_{\text{MLE}}$	$\hat{\Omega}_r$
$50, 200, x$	$\|\hat{\Omega} - \Omega\|_F$	2.860 (0.098)	2.936 (0.105)	2.935 (0.106)	2.930 (0.109)
	FPR	0.016 (0.005)	0.067 (0.017)	0.071 (0.021)	0.075 (0.023)
	TPR	0.340 (0.046)	0.370 (0.061)	0.381 (0.068)	0.389 (0.070)
	AUC	0.880 (0.013)	0.758 (0.019)	0.769 (0.019)	0.764 (0.020)
$50, 200, x^3$	$\|\hat{\Omega} - \Omega\|_F$	2.856 (0.116)	2.942 (0.102)	3.053 (0.098)	2.935 (0.108)
	FPR	0.016 (0.007)	0.068 (0.019)	0.076 (0.020)	0.075 (0.022)
	TPR	0.342 (0.051)	0.372 (0.059)	0.280 (0.051)	0.391 (0.066)
	AUC	0.882 (0.015)	0.759 (0.019)	0.691 (0.020)	0.768 (0.019)
$250, 200, x$	$\|\hat{\Omega} - \Omega\|_F$	3.185 (0.097)	3.742 (0.090)	3.709 (0.089)	3.711 (0.091)
	FPR	0.006 (0.001)	0.025 (0.003)	0.024 (0.003)	0.025 (0.003)
	TPR	0.308 (0.034)	0.238 (0.033)	0.237 (0.031)	0.235 (0.030)
	AUC	0.884 (0.014)	0.759 (0.018)	0.773 (0.018)	0.768 (0.018)
$250, 200, x^3$	$\|\hat{\Omega} - \Omega\|_F$	3.199 (0.096)	3.757 (0.096)	3.894 (0.096)	3.724 (0.087)
	FPR	0.006 (0.001)	0.025 (0.003)	0.026 (0.003)	0.025 (0.003)
	TPR	0.302 (0.034)	0.239 (0.032)	0.143 (0.027)	0.237 (0.032)
	AUC	0.882 (0.012)	0.759 (0.016)	0.691 (0.016)	0.767 (0.015)

$750, 300, x$	$\|\hat{\Omega} - \Omega\|_F$	11.181 (0.134)	10.830 (0.129)	10.640 (0.122)	10.659 (0.118)
	FPR	0.256 (0.006)	0.179 (0.006)	0.180 (0.006)	0.179 (0.005)
	TPR	0.937 (0.009)	0.723 (0.016)	0.744 (0.017)	0.736 (0.016)
	AUC	0.939 (0.006)	0.820 (0.009)	0.831 (0.009)	0.828 (0.009)
$750, 300, x^3$	$\|\hat{\Omega} - \Omega\|_F$	11.196 (0.130)	10.838 (0.129)	11.250 (0.130)	10.646 (0.137)
	FPR	0.256 (0.006)	0.180 (0.006)	0.173 (0.006)	0.179 (0.006)
	TPR	0.937 (0.009)	0.724 (0.016)	0.590 (0.020)	0.737 (0.016)
	AUC	0.939 (0.006)	0.820 (0.008)	0.743 (0.011)	0.828 (0.009)

However, when $f_j(x) = x^3$, performance drops notably, which is driven not by an increased FPR but by a decreased TPR. Again, results for `bridge` and `poly` behave similarly across metrics. No performance reduction, neither in estimation error nor in graph recovery, can be detected when comparing `poly` to the ternary `bridge` estimator.

Appendix D: Application to COVID-19 data

This section presents the results of an analysis of real-world health data (from the UK Biobank). We are interested in investigating associations between the severity of a COVID-19 infection and various potential risk factors. This analysis is intended to illustrate the use of the proposed methods in a real-world, mixed variable type example.

Table 2. Estimated partial correlations between COVID-19 severity and the listed variables for data sets A, B and C.

Covid-19 severity assoc. Variables	Data set A	Data set B	Data set C
age	0.162	0.134	0.140
waist circ.	0.031	0.009	0.011
deprev. idx	0.016	-	-
sex	0.007	-	-
hypertension	0.075	0.035	0.037
heart attack	-	0.073	0.065
diabetes	-	0.062	0.055
chr. bronch.	-	-	0.012
wisd. teeth surg.	-	-0.003	-

D.1. Data set and variables

We first describe the data set used here, which is a part of the UK Biobank COVID-19 resource in which UK Biobank data were linked to clinical COVID data. To construct an indicator of COVID-19 severity, we consider subjects who tested positive for COVID-19 at some point in 2020. Based on that, we created an indicator variable (COVID severity) to capture whether each subject had a severe outcome within six weeks of infection (meaning either hospitalized, hospitalized, receiving critical care, or died). Around 14% experienced such a severe outcome. The analysis includes $n = 8672$ observations on $d = 712$ variables (risk factors

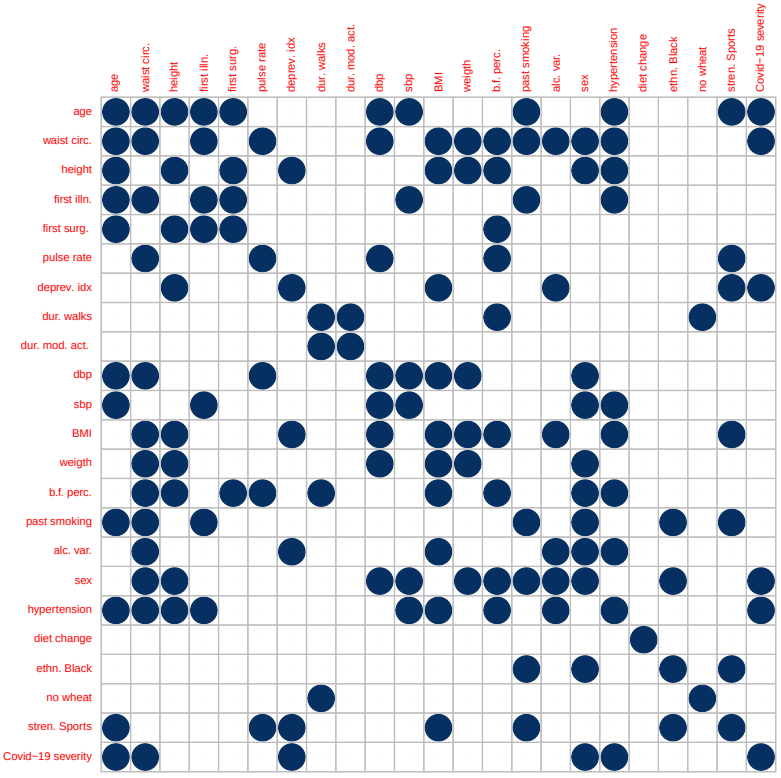

Fig 7. Plot of the estimated adjacency matrix of data set A.

and covariates with less than 40% missingness). Missing values were imputed using `missForest` R-package using default settings. Variables expressing more than 20 states were treated as continuous. The remaining data include 665 binary variables, 25 count variables, and 8 categorical variables. Many of the binary variables represent the status for relatively rare conditions. This means that the share of the minority class of these indicators (i.e., the fraction of samples with the least frequent value of the variable) can often be minimal. To understand

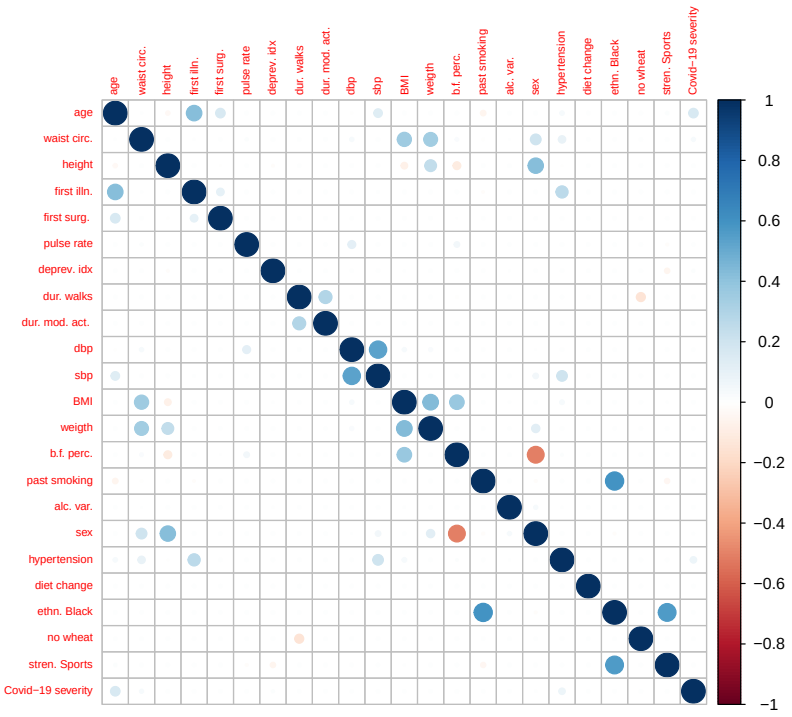

Fig 8. Plot of the estimated precision matrix of data set A.

the effects of such rare events on the analysis, we define three data sets (named A, B, and C) with inclusion rules requiring respectively at least a $25\%, 2\%, 1\%$ share of observations falling into the minority class.

D.2. Results

We present the results of a joint analysis of the variables using real UK Biobank data. We emphasize that the analysis aims to illustrate the proposed estimators' behavior and not fully understand the risk factors for severe COVID-19. There has been much work done on factors influencing the risk of severe COVID-19 and its treatment [see, among others, 45, 5] and we direct the interested reader to the references for further information.

Table 2 gives a summary of the estimated links (indicated as a visualization of the partial correlations) between the variables (including COVID-19 severity). Considering in particular links to COVID-19 severity, we see that `age`, `waist circ.`, `hypertension`, `heart attack` and `diabetes` are quite stable links throughout the different data sets. The effect sizes in terms of partial

correlations are penalized and should be interpreted in relative terms. In particular, `age` retains a relatively large signal, which is in line with the known strong influence of age on COVID-19 severity [see, e.g. 45].

Finally, we present more detailed results of the analysis of data set A. Figure 7 shows the estimated adjacency matrix and Figure 8 depicts the estimated precision matrix $\hat{\boldsymbol{\Omega}}_A$. These results highlight the type of output, spanning different kinds of variables, that is readily available from the proposed method.

D.3. Variable Description for real-world data application

Table 3 gives an overview of the variables in the UK Biobank data set.

Table 3: Variable description of the real world application

Variable Name	Description
age	age in. years in 2020
waist circ.	waist circumference in cm
height	standing in height in cm
first illn.	age at which illness first occurred
first surg.	age at which operation was done first
pulse rate	pulse rate measured in bpm
deprev. idx	Townsend deprivation index at recruitment
dur. walks	duration of walks in minutes per day
dur. mod. act.	duration of moderate activity in minutes per day
dbp	diastolic blood pressure in mmHg
sbp	systolic blood pressure in mmHg
BMI	in kg/m2
weigth	in kg
b.f. perc.	body fat percentage in %
walking	number of days per week walked 10+ minutes
mod. phys. act.	number of days per week of moderate physical activity 10+ minutes
vig. phys. act.	number of days per week of vigorous physical activity 10+ minutes
cheese	answer to “How often do you eat cheese per week?"
stair climb.	answer to "At home, during the last 4 weeks, about how many times a DAY do you climb a flight of stairs? (approx 10 steps)"
curr. smoking	categorial, "Do you smoke tobacco now?" (yes, no, occasionally)
past smoking	categorial, "How often did you smoke tobacco?" (never, once/twice, occasionally, on most days)
diet var.	categorial, "Does your diet change?" (never, sometimes, often)

alc. freq.	categorial, "How often do you drink alcohol?" (never, special occasions only, 1-3 per month, 1-2 per week, 3-4 per week, almost daily)
alc. var.	categorial, "Compared to 10 years ago, do you drink?" (more, about the same, less)
sex	binary indicator with 0=female, 1=male
hypertension	hypertension, binary indicator with 0=no, 1=yes
angina	angina, binary indicator with 0=no, 1=yes
heart attack	heart attack, binary indicator with 0=no, 1=yes
stroke	stroke, binary indicator with 0=no, 1=yes
dvt	deep venous thrombosis, binary indicator with 0=no, 1=yes
asthma	asthma, binary indicator with 0=no, 1=yes
chr. bronch.	emphysema/chronic bronchitis, binary indicator with 0=no, 1=yes
gord	gastro-oesophageal reflux/gastric reflux, binary indicator with 0=no, 1=yes
ibs	irritable bowel syndrome, binary indicator with 0=no, 1=yes
gall stones	cholelithiasis/gall stones, binary indicator with 0=no, 1=yes
kidn./bladder stone	kidney stone/ureter stone/bladder stone, binary indicator with 0=no, 1=yes
diabetes	diabetes, binary indicator with 0=no, 1=yes
diabtes 2	type 2 diabetes, binary indicator with 0=no, 1=yes
myxoedema	hypothyroidism/myxoedema, binary indicator with 0=no, 1=yes
migraine	migraine, binary indicator with 0=no, 1=yes
glaucoma	glaucoma, binary indicator with 0=no, 1=yes
cataract	cataract, binary indicator with 0=no, 1=yes
depression	depression, binary indicator with 0=no, 1=yes
panic attacks	anxiety/panic attacks, binary indicator with 0=no, 1=yes
back probl.	back problems, binary indicator with 0=no, 1=yes
osteoporosis	osteoporosis, binary indicator with 0=no, 1=yes
spine arthr.	spine arthritis/spondylitis, binary indicator with 0=no, 1=yes
slipped disc	prolapsed disc/slipped disc, binary indicator with 0=no, 1=yes
anaemia	iron deficiency anaemia, binary indicator with 0=no, 1=yes
ut. fibroids	uterine fibroids, binary indicator with 0=no, 1=yes
allerg. rhinitis	heyfever/allergic rhinitis, binary indicator with 0=no, 1=yes
enlarged prost.	enlarged prostate, binary indicator with 0=no, 1=yes
pneumonia	pneumonia, binary indicator with 0=no, 1=yes

endometr.	endometriosis, binary indicator with 0=no, 1=yes
ear disor.	ear/vestibular disorder, binary indicator with 0=no, 1=yes
headaches	headaches (not migraine), binary indicator with 0=no, 1=yes
ecz./dermat.	eczema/dermatitis, binary indicator with 0=no, 1=yes
psoriasis	psoriasis, binary indicator with 0=no, 1=yes
div. disease	diverticular disease/diverticulitis, binary indicator with 0=no, 1=yes
osteoarthr.	osteoarthritis, binary indicator with 0=no, 1=yes
gout	gout, binary indicator with 0=no, 1=yes
high chol.	high cholesterol, binary indicator with 0=no, 1=yes
hiat. hern.	hiatus hernia, binary indicator with 0=no, 1=yes
sciatica	sciatica, binary indicator with 0=no, 1=yes
appendic.	appendicitis, binary indicator with 0=no, 1=yes
back pain	back pain, binary indicator with 0=no, 1=yes
arthritis	arthritis (nos), binary indicator with 0=no, 1=yes
measles	measles/morbillivirus, binary indicator with 0=no, 1=yes
chickpox	chickenpox, binary indicator with 0=no, 1=yes
tonsillitis	tonsillitis, binary indicator with 0=no, 1=yes
ptca	coronary angioplasty (ptca)+/-stent, binary indicator with 0=no, 1=yes
ear surg.	ear surgery, binary indicator with 0=no, 1=yes
sinus surg.	nasal/sinus,nose surgery, binary indicator with 0=no, 1=yes
vasectomy	vasectomy, binary indicator with 0=no, 1=yes
soft tiss. surg.	mucsle/soft tissue surgery, binary indicator with 0=no, 1=yes
hip repl.	hip replacement/revision, binary indicator with 0=no, 1=yes
knee repl.	knee replacement/revision, binary indicator with 0=no, 1=yes
spine surg.	spine or back surgery, binary indicator with 0=no, 1=yes
bil. ooph.	bilateral oophorectomy, binary indicator with 0=no, 1=yes
hysterect.	hysterectomy, binary indicator with 0=no, 1=yes
steril.	sterilisation, binary indicator with 0=no, 1=yes
lumpect.	lumpectomy, binary indicator with 0=no, 1=yes
ing. hernia rep.	inguinal/femoral hernia repair, binary indicator with 0=no, 1=yes
umb. hernia rep.	umbilical hernia repair, binary indicator with 0=no, 1=yes

cataract extr.	catarct extraction/lens implant, binary indicator with 0=no, 1=yes
red./fix. bone frac.	reduction or fixationof bone fracture, binary indicator with 0=no, 1=yes
cholecystect.	cholecystectomy/gall bladder removal, binary indicator with 0=no, 1=yes
appendicect.	appendicectomy, binary indicator with 0=no, 1=yes
c-sec.	caesarian section, binary indicator with 0=no, 1=yes
tonsillest.	tonsillectomy, binary indicator with 0=no, 1=yes
var. vein surg.	varicose vein surgery, binary indicator with 0=no, 1=yes
wisd. teeth surg.	wisdom teeth surgery, binary indicator with 0=no, 1=yes
piles surg.	haemorroidectomy/piles surgery/banding of piles, binary indicator with 0=no, 1=yes
male circ.	male circumcision, binary indicator with 0=no, 1=yes
squint corr.	squint correction, binary indicator with 0=no, 1=yes
arthrosc.	arthroscopy (nos), binary indicator with 0=no, 1=yes
foot surg.	foot surgery, binary indicator with 0=no, 1=yes
knee surg.	knee surgery (not replacement), binary indicator with 0=no, 1=yes
shoulder surg.	shoulder surgery, binary indicator with 0=no, 1=yes
car. tunn. surg.	carpal tunnel surgery, binary indicator with 0=no, 1=yes
valg. surg.	bunion/hallus valgus surgery, binary indicator with 0=no, 1=yes
rem. mole	removal of mole/skin lesion, binary indicator with 0=no, 1=yes
ov. cyst. rem.	ovarian cyst removal/surgery, binary indicator with 0=no, 1=yes
d+c	dilatation and curettage, binary indicator with 0=no, 1=yes
cone biops.	cone biopsy, binary indicator with 0=no, 1=yes
endosc.	endoscopy/gastroscopy, binary indicator with 0=no, 1=yes
colonosc.	colonoscopy/sigmoidoscopy, binary indicator with 0=no, 1=yes
laparosc.	laparoscopy, binary indicator with 0=no, 1=yes
rhinoplast.	rhinoplasty/nose surgery, binary indicator with 0=no, 1=yes
tonsil surg.	tonsillectomy/tonsil surgery, binary indicator with 0=no, 1=yes
ing. hern. rep.	inguinal hernia repair, binary indicator with 0=no, 1=yes
illn. ind. diet	Major dietary changes in the last 5 years because of illness, binary indicator with 0=no, 1=yes

diet change	Major dietary changes in the last 5 years because of other reason, binary indicator with 0=no, 1=yes
ethn. Mixed	Ethnicity - mixed, binary indicator with 0=no, 1=yes
ethn. Asian	Ethnicity - Asian, binary indicator with 0=no, 1=yes
ethn. Black	Ethnicity - Black, binary indicator with 0=no, 1=yes
no eggs	Never eat eggs or foods containing eggs, binary indicator with 0=no, 1=yes
no dairy	Never dairy products, binary indicator with 0=no, 1=yes
no wheat	Never eat wheat, binary indicator with 0=no, 1=yes
no sugar	Never eat sugar or foods/drinks containing sugar, binary indicator with 0=no, 1=yes
walk. f. pleas.	Types of physical activity in last 4 weeks - walking for pleasure, binary indicator with 0=no, 1=yes
exercises	Types of physical activity in last 4 weeks - other exercises (swimming, bowling etc.), binary indicator with 0=no, 1=yes
stren. Sports	Types of physical activity in last 4 weeks - strenuous sports, binary indicator with 0=no, 1=yes
Covid-19 severity	Covid-19 severity, binary indicator with 0=mild outcome and 1=severe outcome

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.

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].

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. https://doi.org/10.1007/BF02294504 MR1173589
[4] Bedrick, E. J. and Breslin, F. C. (1996). Estimating the polyserial correlation coefficient. *Psychometrika* **61** 427–443. https://doi.org/10.1007/BF02294548 MR1424910
[5] Berlin, D. A., Gulick, R. M. and Martinez, F. J. (2020). Severe Covid-19. *New England Journal of Medicine* **383** 2451–2460. https://doi.org/10.1056/nejmcp2009575
[6] Broyden, C. G. (1965). A class of methods for solving nonlinear simultaneous equations. *Math. Comp.* **19** 577–593. https://doi.org/10.2307/2003941 MR198670
[7] 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. https://doi.org/10.1198/jasa.2011.tm10155 MR2847973
[8] Chen, S., Witten, D. M. and Shojaie, A. (2015). Selection and estimation for mixed graphical models. *Biometrika* **102** 47–64. https://doi.org/10.1093/biomet/asu051 MR3335095
[9] Cheng, J., Li, T., Levina, E. and Zhu, J. (2017). High-dimensional mixed graphical models. *J. Comput. Graph. Statist.* **26** 367–378. https://doi.org/10.1080/10618600.2016.1237362 MR3640193
[10] Cox, N. R. (1974). Estimation of the correlation between a continuous and a discrete variable. *Biometrics* **30** 171–178. https://doi.org/10.2307/2529626 MR334376
[11] 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. https://doi.org/10.1016/j.jmva.2004.02.009 MR2064941
[12] 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. https://doi.org/10.1111/rssb.12168 MR3611752
[13] 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.
[14] Finegold, M. and Drton, M. (2011). Robust graphical modeling of gene networks using classical and alternative t-distributions. *Ann. Appl. Stat.* **5** 1057–1080. https://doi.org/10.1214/10-AOAS410 MR2840186
[15] Fox, J. (2022). polycor: Polychoric and Polyserial Correlations R package version 0.8-1.
[16] Foygel, R. and Drton, M. (2010). Extended Bayesian Information Cri-

teria 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.

[17] Friedman, J., Hastie, T. and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441. https://doi.org/10.1093/biostatistics/kxm045

[18] Higham, N. J. (1988). Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra Appl.* **103** 103–118. https://doi.org/10.1016/0024-3795(88)90223-6 MR943997

[19] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30. MR144363

[20] Jin, S. and Yang-Wallentin, F. (2017). Asymptotic robustness study of the polychoric correlation estimation. *Psychometrika* **82** 67–85. https://doi.org/10.1007/s11336-016-9512-2 MR3614808

[21] Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. https://doi.org/10.1214/09-AOS720 MR2572459

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

[23] Lee, J. D. and Hastie, T. J. (2015). Learning the structure of mixed graphical models. *J. Comput. Graph. Statist.* **24** 230–253. https://doi.org/10.1080/10618600.2014.900500 MR3328255

[24] 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

[25] Liu, H., Han, F., Yuan, M., Lafferty, J. and Wasserman, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.* **40** 2293–2326. https://doi.org/10.1214/12-AOS1037 MR3059084

[26] Mei, S., Bai, Y. and Montanari, A. (2018). The landscape of empirical risk for nonconvex losses. *Ann. Statist.* **46** 2747–2774. https://doi.org/10.1214/17-AOS1637 MR3851754

[27] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. https://doi.org/10.1214/009053606000000281 MR2278363

[28] Miyamura, M. and Kano, Y. (2006). Robust Gaussian graphical modeling. *J. Multivariate Anal.* **97** 1525–1550. https://doi.org/10.1016/j.jmva.2006.02.006 MR2275418

[29] 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. https://doi.org/10.1016/j.neuroimage.2014.07.033

[30] Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika* **44** 443–460. https://doi.org/10.1007/BF02296207 MR554892

[31] Olsson, U., Drasgow, F. and Dorans, N. J. (1982). The polyserial correlation coefficient. *Psychometrika* **47** 337–347. https://doi.org/10.1007/BF02294164 MR678066

[32] 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.

[33] Pearson, K. (1913). On the measurement of the influence of "broad categories" on correlation. *Biometrika* **9** 116–139.

[34] Perrakis, K., Lartigue, T., Dondelinger, F. and Mukherjee, S. (2019). Regularized joint mixture models. arXiv:1908.07869.

[35] Quan, X., Booth, J. G. and Wells, M. T. (2018). Rank-based approach for estimating correlations in mixed ordinal data. arXiv: 1809.06255.

[36] 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. https://doi.org/10.1214/09-AOS691 MR2662343

[37] 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. https://doi.org/10.1214/11-EJS631 MR2836766

[38] 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. https://doi.org/10.1214/13-AOAS662 MR3161717

[39] Städler, N. and Mukherjee, S. (2015). Multivariate gene-set testing based on graphical models. *Biostatistics* **16** 47–59. https://doi.org/10.1093/biostatistics/kxu027 MR3365410

[40] Tallis, G. M. (1962). The maximum likelihood estimation of correlation from contingency tables. *Biometrics* **18** 342–353. https://doi.org/10.2307/2527476 MR145613

[41] Yang, Z., Ning, Y. and Liu, H. (2018). On semiparametric exponential family graphical models. *J. Mach. Learn. Res.* **19** Paper No. 57, 59. MR3899759

[42] Verzelen, N. and Villers, F. (2009). Tests for Gaussian graphical models. *Comput. Statist. Data Anal.* **53** 1894–1905. https://doi.org/10.1016/j.csda.2008.09.022 MR2649554

[43] 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. https://doi.org/10.1109/tsp.2006.874409

[44] Wei, Z. and Li, H. (2007). A Markov random field model for network-based analysis of genomic data. *Bioinformatics* **23** 1537–1544. https://doi.org/10.1093/bioinformatics/btm129

[45] Williamson, E. J., Walker, A. J., Bhaskaran, K., Bacon, S., Bates, C., Morton, C. E., Curtis, H. J., Mehrkar, A., Evans, D., In-

GLESBY, P., COCKBURN, J., MCDONALD, H. I., MACKENNA, B., TOMLINSON, L., DOUGLAS, I. J., RENTSCH, C. T., MATHUR, R., WONG, A. Y. S., GRIEVE, R., HARRISON, D., FORBES, H., SCHULTZE, A., CROKER, R., PARRY, J., HESTER, F., HARPER, S., PERERA, R., EVANS, S. J. W., SMEETH, L. and GOLDACRE, B. (2020). Factors associated with COVID-19-related death using OpenSAFELY. *Nature* **584** 430–436. https://doi.org/10.1038/s41586-020-2521-4

[46] XUE, L. and ZOU, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann. Statist.* **40** 2541–2571. https://doi.org/10.1214/12-AOS1041 MR3097612

[47] 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.

[48] YOON, G., MÜLLER, C. L. and GAYNANOVA, I. (2021). Fast Computation of Latent Correlations. *Journal of Computational and Graphical Statistics* **30** 1249-1256. https://doi.org/10.1080/10618600.2021.1882468

[49] YUAN, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **11** 2261–2286. MR2719856