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 yet 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 flexible and scalable methodology for data with variables of entirely general mixed type. We study the key properties of the approaches theoretically and empirically, via extensive simulations as well an illustrative application to data from the UK Biobank concerning COVID-19 risk factors.

Keywords Generalized correlation · high-dimensional statistics · latent Gaussian copula · mixed data · polychoric/polyserial correlation · undirected graphical models

1 Introduction

Graphical models are widely used in the analysis of multivariate data, providing a convenient and interpretable way to study relationships among potentially large numbers of variables. They are key tools in modern statistics and
machine learning and play an important role in diverse applications. Undirected graphical models are used in a wide range of settings, including, among others, systems biology, omics, deep phenotyping [see, e.g. 1, 2, 3] and as a component within other analyses including two-sample testing, unsupervised learning, hidden Markov modelling and more [examples include 4, 5, 6, 7, 8].

A large part of the graphical models literature has focused on the case in which either only continuous variables or only discrete variables are present. Pertaining to the former case, Gaussian graphical models have been extensively studied, including in the high-dimensional regime [see among others 9, 10, 11, 12, 13, 14, 15]. 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 [16, 17, 2]. In the latter case, discrete graphical models – related to Ising-type models in statistical physics – have also been extensively studied [see e.g. 18, 19].

However, in many applications it is common to encounter data that is mixed with respect to variable type, i.e. where the data vector includes components that are 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 as well as discrete information about disease status or pain levels for example. In analysing such data, it is often of interest to estimate a joint multivariate graphical model spanning the various variable types. In practice, this is sometimes done using ad hoc pipelines and data transformations. However, in graphical modelling, 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 modelling for mixed data. The earliest approach is conditional Gaussian modelling of a mix of categorical and continuous data [20] as treated by Cheng et al. [21], Lee and Hastie [22]. 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. 23, 24, 25]. A third approach uses latent Gaussian models with a key recent reference being the paper of Fan et al. [26] 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, Fan et al. [26] consider only a mix of binary and continuous data, and do not allow for more general combinations (including counts or ordinal variables) as found in many real-world applications.

This third approach will be in the focus of this paper, which aims to provide a simple framework for working with latent Gaussian copula models in order 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 [27, 28] worked on the foundations of these ideas in 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 et al. [17] as a nonparametric extension to the Gaussian family. A random vector \( \mathbf{X} \in \mathbb{R}^d \) is a member of the nonparanormal family when \( f(\mathbf{X}) = (f_1(X_1), \ldots, f_d(X_d))^T \) is Gaussian, where \( \{f_k\}_{k=1}^d \) is a set of univariate monotone transformation functions. Moreover, if the \( f_j \)’s are differentiable on top of being monotone, the nonparanormal family is equivalent to the Gaussian copula family. As the polychoric and polyserial correlation assume that observed discrete data are generated from latent continuous variables, they in fact adhere to a latent copula approach.

We propose two estimators of the latent correlation matrix which can subsequently be plugged in to existing routines to estimate the precision matrix such as the graphical LASSO (glasso) [10], CLIME [15], or the graphical Dantzig selector [13]. The first one is appropriate under a latent Gaussian model and simply unifies the aforementioned MLEs. The second one is more general and is applicable under the latent Gaussian copula model. Both approaches can deal with discrete variables with general numbers of levels. We show that both estimators exhibit favorable theoretical properties and include simulation as well as real data results. Thus, the main contributions of the paper are as follows:
• We make the observation that incorporating polychoric and polyserial correlations into the latent Gaussian copula framework provides an elegant, simple and effective method for graphical modelling for fully general mixed data.
• We provide theoretical results concerning the behaviour of the proposed estimators, including in the high-dimensional setting. These concentration results demonstrate that the procedures proposed are statistically valid.
• We study the estimators empirically, via a range of simulations as well as an example using real phenotyping data of mixed type (from the UK Biobank). These results illustrate how the proposed methods can be used in practice and demonstrate that performance is often close to an oracle that is given access to true latent data.

Taken together, our results provide users a way to carry out graphical modelling of mixed data that is statistically sound and practically easy-to-use, involving no more overhead than standard high-dimensional Gaussian graphical modelling approaches and in particular no need to manually specify any model components (such as bridge functions) that are specific to the variable types.

The remainder of this paper is organized as follows. In Sections 2 and 3 we present the estimators based on polychoric and polyserial correlations. In Section 4 we describe the experimental setup used to test the proposed approaches on simulated data with the results themselves appearing in Section 5. Section 6 showcases an illustrative empirical application using real data from the UK Biobank.

We close with conclusions and directions to our R package hume which allows users to readily implement the herein developed methods.

2 Background and model set-up

The goal of this paper is to learn undirected graphical model structure for general mixed and high-dimensional data. To this end, we extend the Gaussian copula model [17, 29, 30] so as to allow inclusion of any type of discrete and continuous data.

**Definition 2.1** (latent Gaussian copula model for general mixed data). Assume we have a mix of ordinal and continuous variables, i.e. \( \mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2) \) where \( \mathbf{X}_1 \) denotes \( d_1 \)-dimensional possibly ordered discrete variables and \( \mathbf{X}_2 \) represents \( d_2 \)-dimensional continuous variables. Then, \( \mathbf{X} \) satisfies the latent Gaussian copula model, if there exists a corresponding \( d_1 \)-dimensional random vector of latent continuous variables \( \mathbf{Z}_1 = (Z_1, \ldots, Z_{d_1})' \) s.t. \( \mathbf{Z} := (\mathbf{Z}_1, \mathbf{X}_2) \sim NPN(\mu, \Sigma^*, f) \) where \( \mu = (\mu_j)_{j=1,...,d} \) is the mean vector and \( \Sigma^* = (\Sigma^*_{jk})_{1 \leq j,k \leq d} \) the correlation matrix and \( f = \{f_1, \ldots, f_d\} \) a set of monotone univariate functions.

Let further
\[
X_j = x^d_r \quad \text{if} \quad \gamma^d_{r-1} \leq Z_j < \gamma^d_r \quad \text{for all} \quad j = 1, \ldots, d_1 \quad \text{and} \quad r = 1, \ldots, l_{X_j},
\]
where \( \gamma^d_r \) represent some thresholds. For simplicity, we denote \( \gamma^d_0 = -\infty \) and \( \gamma^d_{l_{X_j}} = +\infty \), \( x^d_r \in \mathbb{N}_0 \) and \( l_{X_j} \) the number of levels of \( X_j, j = 1, \ldots, d_1 \).

Then we say that the \( d_1 \cup d_2 = d \)-dimensional vector \( \mathbf{X} \) satisfies the latent Gaussian copula model, i.e. \( \mathbf{X} \sim LNPN(\mu, \Sigma^*, f, \Gamma) \) where \( \Gamma = (\gamma^1, \ldots, \gamma^d) \) is a collection of thresholds. If \( \mathbf{Z} \sim N(\mu, \Sigma^*) \), then \( \mathbf{X} \) satisfies the latent Gaussian model, i.e. \( LN(\mu, \Sigma^*, \Gamma) \).

Let \( \Omega^* = \Sigma^*^{-1} \) denote the latent precision matrix. Then, as shown in Liu et al. [17], the zero-pattern of \( \Omega^* \) under the latent Gaussian copula model still encodes the conditional independencies of the latent continuous variables. Thus, the underlying undirected graph is represented by \( \Omega^* \) just as for the parametric normal. Note that the latent Gaussian copula model for general mixed data in Definition 2.1 agrees with that of Quan et al. [31] and of Feng and Ning [32]. The problem phrased by Fan et al. [26] is a special case of Definition 2.1. A more detailed comparison between both approaches can be found in Section 3. Nominal discrete variables need to be transformed to a dummy system.

For the remainder of the paper we assume we have an independent \( n \)-sample of the \( d \)-dimensional vector \( \mathbf{X} \). We estimate \( \Sigma^* \) by considering the corresponding entries separately i.e. the couples \((X_j, X_k)\). 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. \( X_j, X_k \subset X_2 \).
- **Case II**: \( X_j \) is discrete and \( X_k \) is continuous, i.e. \( X_j \subset X_1 \) and \( X_k \subset X_2 \). By symmetry the case where \( X_j \) is continuous and \( X_k \) is ordinal is identical to case II.
- **Case III**: Both \( X_j \) and \( X_k \) are discrete, i.e. \( X_j, X_k \subset X_1 \).
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. Let us start with 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 Ravikumar et al. [14]. Hence, the estimator for $\Sigma^*$ when both $X_j$ and $X_k$ are continuous is:

**Definition 2.2** (Estimator $\hat{\Sigma}^{(n)}$ of $\Sigma^*$; Case I). Let $\bar{x}_j$ denote the sample mean of $X_j$. The estimator $\hat{\Sigma}^{(n)} = (\hat{\Sigma}^{(n)}_{jk})_{1 \leq j, k \leq d}$ of the correlation matrix $\Sigma^*$ is defined by:

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

(2)

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

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

Turning to case II, let $X_j$ be ordinal and $X_k$ be continuous. We are interested in the product-moment correlation $\Sigma_{jk}$ between two jointly Gaussian variables, where $X_j$ is not directly observed but only the ordered categories (see Eq. (1)). This is called the *polyserial* correlation [33]. The likelihood and log-likelihood of the $n$-sample are defined by:

$$
L^{(n)}(\Sigma_{jk}, x_r^j, x_k) = \prod_{i=1}^{n} p(x_{ir}^j, x_{ik}, \Sigma_{jk}) = \prod_{i=1}^{n} p(x_{ik})p(x_{ir}^j | x_{ik}, \Sigma_{jk}) ;
$$

(3)

$$
\ell^{(n)}(\Sigma_{jk}, x_r^j, x_k) = \sum_{i=1}^{n} \left[ \log(p(x_{ik})) + \log(p(x_{ir}^j | x_{ik}, \Sigma_{jk})) \right].
$$

where $p(x_{ir}^j, x_{ik}, \Sigma_{jk})$ denotes the joint probability of $X_j$ and $X_k$ and $p(x_{ik})$ the marginal density of the Gaussian variable $X_k$. For notational simplicity the subscripts in $L_{jk}$ and $\ell_{jk}$ will be omitted. MLEs are obtained by differentiating the log-likelihood in Eq. (3) with respect to the unknown parameters and setting the partial derivatives to zero and solving the system of equations for $\Sigma_{jk}, \mu, \sigma^2$, and $\gamma_r, r = 1, \ldots, l_{X_j} - 1$.

**Definition 2.3** (Estimator $\hat{\Sigma}^{(n)}$ of $\Sigma^*$; Case II). Recall the log-likelihood in Eq. (3). The estimator $\hat{\Sigma}^{(n)} = (\hat{\Sigma}^{(n)}_{jk})_{1 \leq j, k \leq d}$ of the correlation matrix $\Sigma^*$ is defined by:

$$
\hat{\Sigma}^{(n)}_{jk} = \arg\max_{|\Sigma_{jk}| \leq 1} \ell^{(n)}(\Sigma_{jk}, x_r^j, x_k)
$$

(4)

$$
= \arg\max_{|\Sigma_{jk}| \leq 1} \frac{1}{n} \ell^{(n)}(\Sigma_{jk}, x_r^j, x_k)
$$

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

Regularity conditions ensuring consistency and asymptotic normality can be verified to hold here [34].

Lastly, consider case III, where both $X_j$ and $X_k$ are ordinal. Consider the probability of an observation with $X_j = x_r^j$ and $X_k = x_r^k$:

$$
\pi_{rs} := P(X_j = x_r^j, X_k = x_r^k) = P(g_{j-1}^r \leq Z_j < g_j^r, g_{k-1}^s \leq Z_k < g_k^s) = \int_{g_{j-1}^r}^{g_j^r} \int_{g_{k-1}^s}^{g_k^s} \phi(z_j, z_k, \Sigma_{jk})dz_j dz_k,
$$

(5)

where $r = 1, \ldots, l_{X_j} - 1$ and $s = 1, \ldots, l_{X_k} - 1$ and $\phi(x, y, \rho)$ denotes the standard bivariate density with correlation $\rho$. Then, as outlined by Olsson [35] the likelihood and log-likelihood of the $n$-sample are defined as:

$$
L^{(n)}(\Sigma_{jk}, x_r^j, x_k) = C \prod_{r=1}^{l_{X_j}} \prod_{s=1}^{l_{X_k}} \pi_{rs}^{n_{rs}},
$$

(6)

$$
\ell^{(n)}(\Sigma_{jk}, x_r^j, x_k) = \log(C) + \sum_{r=1}^{l_{X_j}} \sum_{s=1}^{l_{X_k}} n_{rs} \log(\pi_{rs}).
$$
where $C$ is a constant and $n_{rs}$ denotes the observed frequency of $X_j = x_j^l$ and $X_k = x_k^k$ in a sample of size $n = \sum_{r=1}^{l_x} \sum_{s=1}^{l_x} n_{rs}$. Differentiating the log-likelihood, setting it to zero, and solving for the unknown parameters yields the estimator for $\Sigma^*$ for case III:

**Definition 2.4** (Estimator $\hat{\Sigma}^{(n)}$ of $\Sigma^*$; Case III). Recall the log-likelihood in Eq. (6). The estimator $\hat{\Sigma}^{(n)} = (\hat{\Sigma}^{(n)}_{jk})_{1 \leq j, k \leq d}$ of the correlation matrix $\Sigma^*$ is defined by:

$$\hat{\Sigma}^{(n)}_{jk} = \arg \max_{|\Sigma_{jk}| \leq 1} \ell^{(n)}(\Sigma_{jk}, x^l, x^k)$$

(7)

$$= \arg \max_{|\Sigma_{jk}| \leq 1} \frac{1}{n} \ell^{(n)}(\Sigma_{jk}, x^l, x^k)$$

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 [36].

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

### 3 Latent Gaussian Copula Models

Fan et al. [26] propose a special case of the latent Gaussian copula model in Definition 2.1 where they consider a mix of binary and continuous variables. In the spirit of the nonparanormal SKEPTIC [29] they avoid estimating the monotone transformation functions $\{f_j\}_{j=1}^d$ directly by making use of rank correlation measures such as Kendall’s tau or Spearman’s rho. These measures are invariant under monotone transformations and for case I there exists a well-known mapping between Kendall’s tau and Spearman’s rho and the underlying Pearson correlation coefficient $\Sigma^*$. Consequently, the main contribution of Fan et al. [26] is the derivation of corresponding bridge functions for cases II and III. When pondering the general mixed case, they recommend binarizing all ordinal variables.

This thought has been taken up by Feng and Ning [32] who propose to first binarize all ordinal variables to form preliminary estimators and subsequently combine them meaningfully by some weighted aggregate.

In an attempt to work on generalizations regarding the bridge functions, Quan et al. [31] extended the binary latent Gaussian copula model to the setting where a mix of continuous, binary, and ternary variables is present. However, a considerable drawback of this procedure becomes apparent. While for the binary-continuous mix three bridge functions were needed – one for each case – the number of mappings increases for each discrete variable with distinct state space. Indeed, a mix of continuous variables and discrete variables with say 5 different state spaces already requires $\binom{5+2}{2} = 21$ bridge functions.

For this reason, we take a different avenue to the latent Gaussian copula model for general mixed data where the discrete variables are allowed to have any number of states. In this approach, the number of cases we need to consider remains exactly three as already introduced in the previous section.

#### 3.1 Nonparanormal Case I

For case I, the mapping between $\Sigma^*_{jk}$ and the population versions of Spearman’s rho and Kendall’s tau is well known [17]. Here we make use of Spearman’s rho $\rho^{Sp}_{jk} = \text{corr}(F_j(X_j), F_k(X_k))$ with $F_j$ and $F_k$ denoting the cumulative distribution functions (CDFs) of $X_j$ and $X_k$, respectively. Then $\Sigma^*_{jk} = 2 \sin \frac{\pi}{6} \rho^{Sp}_{jk}$ for $d_1 < j < k \leq d_2$. In practice, we use the sample estimate

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

with $R_{ji}$ corresponding to the rank of $X_{ji}$ among $X_{j1}, \ldots, X_{jn}$ and $\bar{R}_j = 1/n \sum_{i=1}^n R_{ji} = (n + 1)/2$ [similar setup as 29]. From this we obtain the following estimator:

**Definition 3.1** (Estimator $\hat{\Sigma}^{(n)}$ of $\Sigma^*$; Case I nonparanormal). The estimator $\hat{\Sigma}^{(n)} = (\hat{\Sigma}^{(n)}_{jk})_{1 \leq j, k \leq d}$ of the correlation matrix $\Sigma^*$ is defined by:

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

(8)

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

5
3.2 Nonparanormal Case II

For case II, matters become more involved. In order to make use of the rank-based approach regarding the nonparanormal model the ML procedure can no longer be applied as we do not observe the continuous variable in its Gaussian form. Instead, we will proceed by suitably modifying other approaches that address the Gaussian case through more direct examination of the relationship between $\Sigma_{jk}$ and the point polyserial correlation [37, 38].

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 $Z$. According to Definition 2.2, we have the following Gaussian conditional expectation

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

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 $f(Z)$, i.e.

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

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

$$\Sigma^*_jk = \frac{E[f(X_k)X_j]}{\sigma_{f(X_k)} E[f(Z)] X_j} = \frac{r_{f(X_k)X_j} \sigma_{X_j}}{E[f(Z)] X_j},$$

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. (10). Let us start with the expectation in the denominator $E[f(Z)] X_j$. By assumption $f(Z) \sim N(0, \Sigma^*)$ and therefore w.l.o.g. $f(Z) \sim N(0, 1)$ for all $j = 1, \ldots, d$. Consequently, we have:

$$E[f(Z)_d X_j] = \sum_{r=1}^{l_x_d} x_r^d \int_{\gamma_{r-1}^j}^{\gamma_r^j} f(z_j) dF(f(z_j)) = \sum_{r=1}^{l_x_d} x_r^d \int_{\gamma_{r-1}^j}^{\gamma_r^j} f(z_j) \phi(f(z_j)) dz_j$$

$$= \sum_{r=1}^{l_x_d} (x_r^d - x_{r+1}^d) \phi(\gamma_r^j)$$

where $\phi(t)$ denotes the standard normal density. Whenever the ordinal states are consecutive integers we have $\sum_{r=1}^{l_x_d} \phi(\gamma_r^j)$. Based on this derivation it is straightforward to give an estimate of $E[f(Z)_d X_j]$, once estimates of the thresholds $\gamma_r^j$ have been formed (see Section 3.4 for more details). Let us turn to the numerator of Eq. (10). The standard deviation of $X_j$ does not require any special treatment, and we simply use $\sigma_{X_j}^{(n)} = \sqrt{1/n \sum_{i=1}^n (X_{ji} - X_j)^2}$ in order 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) back-transformed version of the continuous variables. Therefore, we proceed to estimate the back-transformation.

To this end, consider the marginal distribution function of $X_k$, namely $F_{X_k}(x) = P(X_k \leq x) = P(f(X_k) < f(x)) = \Phi(f(x))$ such that $f(x) = \Phi^{-1}(F_{X_k}(x))$. In this setting, Liu et al. [17] 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 diverges quickly when evaluated at the boundaries of the $[0, 1]$ interval. More precisely, consider $\hat{f}(u) = \Phi^{-1}(\hat{W}_{\delta_n}(\hat{F}_{X_k}(u)))$, where $\Phi^{-1}(\cdot)$ is again the quantile function of the standard normal and $\hat{W}_{\delta_n}$ is a Winsorization operator, i.e. $\hat{W}_{\delta_n}(u) \equiv \delta_n I(u < \delta_n) + u I(\delta_n \leq u \leq (1 - \delta_n)) + (1 - \delta_n) I(u > (1 - \delta_n))$. The truncation constant $\delta_n$ can be chosen in several ways. Liu et al. [17] 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.

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

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

6
Definition 3.2 (Estimator \( \hat{\Sigma}^{(n)} \) of \( \Sigma^* \); Case II nonparanormal). The estimator \( \hat{\Sigma}^{(n)} = \hat{\Sigma}^{(n)}_{1 \leq j, k \leq d} \) of the correlation matrix \( \Sigma^* \) is defined by:

\[
\hat{\Sigma}^{(n)}_{jk} = \frac{\sum_{r=1}^n f^{(n)}(X_k, X_j, \sigma^{(n)}_j) x_r^j}{\sum_{r=1}^n \phi(\gamma_{r}^j)(x_{r+1}^j - x_r^j)} \tag{12}
\]

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 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 2.2. 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 (Estimator \( \hat{\Sigma}^{(n)} \) of \( \Sigma^* \); Case III nonparanormal). The estimator \( \hat{\Sigma}^{(n)} = \hat{\Sigma}^{(n)}_{1 \leq j, k \leq d} \) of the correlation matrix \( \Sigma^* \) is defined by:

\[
\hat{\Sigma}^{(n)}_{jk} = \arg \max_{|\Sigma_{jk}| \leq 1} \frac{1}{n} \ell^{(n)}(\Sigma_{jk}, x^j, x^k) \tag{13}
\]

for all \( 1 < j < k \leq d_1 \).

In summary, the estimator \( \hat{\Sigma}^{(n)} \) 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 \( \Gamma \) play a key role in linking the observed discrete to the latent continuous variables. Therefore, being able to form an accurate estimate of the \( (\gamma^1_j, \ldots, \gamma^{d_1}_j) \) is crucial for both the likelihood-based procedures as well as the nonparanormal estimator outlined above.

We start by highlighting that we set the model up such that for each \( \gamma^j_r, \ r = 1, \ldots, l_{x_j} - 1 \) there exists a constant \( G \) such that \( |\gamma^j_r| \leq G \) for all \( r = 1, \ldots, l_{x_j} - 1 \), i.e. the estimable thresholds are bounded away from infinity. Let us define the cumulative probability vector \( \pi^j = (\pi^j_1, \ldots, \pi^j_{l_{x_j} - 1}) \). Then, by Eq. (1), it is easy to see that

\[
\pi^j_r = \sum_{i=1}^r P(X_j = x^j_i) = P(X_j \leq x^j_r) = P(Z_j \leq \gamma^j_r) = \Phi(\gamma^j_r),
\]

where \( \Phi(\cdot) \) denotes the cumulative distribution function of a standard normal random variable. 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 by its sample equivalent, namely

\[
\hat{\pi}^j_r = \sum_{k=1}^r \left[ \frac{1}{n} \sum_{i=1}^n \mathbb{1}(X_{ji} = x^j_k) \right] = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(X_{ji} \leq x^j_r),
\]

and plug it into the identity, i.e. \( \hat{\gamma}^j_r = \Phi^{-1}(\hat{\pi}^j_r) \) for \( j = 1, \ldots, d_1 \). The following lemma assures that these threshold estimates can be formed with high accuracy.

**Lemma 3.1.** Consider the event \( A^j_r = \{ |\hat{\gamma}^j_r| \leq 2G \} \). The following bound holds for all \( j = 1, \ldots, d_1 \) and \( r = 1, \ldots, l_{x_j} - 1 \) for some Lipschitz constant \( L_1 \)

\[
P(\bar{A}^j_r) \leq 2(l_{x_j} - 1) \exp \left( \frac{-2G^2 n}{L_1^2} \right),
\]

where \( A^j_r = \cap_{r=1}^{l_{x_j} - 1} A^j_r \).

The proof of Lemma 3.1 is given in Section 4 of the Supplementary Materials. All herein developed methods are applied in a two-step fashion. We stress this by denoting the estimated thresholds as \( \hat{\gamma}^j_r \) in the ensuing theoretical results.
3.5 Concentration results

We start by stating the following assumptions:

**Assumption 3.1.** For all $1 \leq j < k \leq d_1$, $\Sigma^*_{jk} \neq 1$. In other words, there exists a constant $\delta > 0$ such that $|\Sigma^*_{jk}| \leq 1 - \delta$.

**Assumption 3.2.** For $\gamma^r_j$ there exists a constant $G$ such that $|\gamma^r_j| \leq G$ for any $j = 1, \ldots, d_1$ and for all $r = 1, \ldots, \bar{l}$.

**Assumption 3.3.** Let $j < k$ and consider the log-likelihood functions in Definition 2.3 and in Definition 2.4. 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 $\Sigma^*_{jk}$ is non-degenerate.
- All joint and conditional states of the discrete variables have positive probability.

Assumptions 3.1 and 3.2 guarantee that $f(X_j)$ and $f(X_k)$ are not perfectly linearly dependent and that the thresholds are bounded away from infinity, respectively (these impose few restrictions in practice).

Assumption 3.3 assures that the likelihood functions under the latent Gaussian model behave in a “nice” way. This is again a requirement that resembles a mild technical assumption. The following theorem, relies on Mei et al. [39] and requires four conditions that are verified to hold in Section 2 of the Supplementary Materials. We note that a similar approach has been employed by Anne et al. [40] in the context of zero-inflated Gaussian data under double truncation.

**Theorem 3.2.** Assume that Assumptions 3.1–3.3 hold and let $j, k \in 1, \ldots, d_1$ and $k \in d_1 + 1, \ldots, d$ (case II) or $j, k \in 1, \ldots, d_1$ (case III). Further, let $0 < \alpha < 1$. There exist some known constants $B, C,$ and $D$ independent of $(n, d)$ but depending on cases II and III. Now, let $n \geq 4C \log(n) \log \left( \frac{B}{\alpha} \right)$, such that $\hat{\Sigma}^{(n)}_{jk}$ satisfies the following inequality:

$$P \left( \max_{j,k} \left| \hat{\Sigma}^{(n)}_{jk} - \Sigma^*_{jk} \right| > D \sqrt{\frac{C \log(n)}{n}} \right) \leq \frac{d(d-1)}{2} \alpha. \quad (14)$$

Case I of the latent Gaussian model deals only with observed Gaussian variables and concentration results can be retrieved for example from Ravikumar et al. [14]. Having treated the latent Gaussian model, we now turn to the nonparanormal extension.

In principle the three cases will have to be considered again.

- **Case I:** When both random variables are continuous concentration results follow immediately from Liu et al. [29] 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 the important observation that Theorem 3.2 above still applies and needs not be altered. We do not observe the continuous variables directly but only their discretized versions. As a consequence, 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.

**Theorem 3.3.** Suppose that Assumptions 3.1 and 3.2 hold and $j, k \in 1, \ldots, d_1$ and $k \in d_1 + 1, \ldots, d$. Then for any $\epsilon \in \left[ C_M \sqrt{\frac{\log \log^2 n}{\sqrt{n}}}, 8(1 + 4e^2) \right]$ with sub-Gaussian parameter $c$, generic constants $k_i$, $i = 1, 2, 3$ and constant $C_M = \frac{48}{\sqrt{\epsilon}}(\sqrt{2M} - 1)(M + 2)$ for some $M \geq 2 \left( \frac{\log d_1}{\log n} + 1 \right)$ with $C_{\gamma} = \sum_{r=1}^{\bar{l}} \phi(\bar{\gamma}_r)(x_{r+1}^j - x_r^j)$ and Lipschitz
constant \( L \) the following probability bound holds

\[
P(\max_{j,k} | \hat{\Sigma}^{(n)}_{jk} - \Sigma^*_{jk} | \geq \epsilon) \\
\leq 8 \exp\left(2 \log d - \frac{\sqrt{n} \epsilon^2}{(64 L C_\gamma l_{\max} \pi)^2 \log n}\right) \\
+ 8 \exp\left(2 \log d - \frac{ne^2}{(4L C_\gamma)^2 (1+4\epsilon^2)^2}\right) \\
+ 8 \exp\left(2 \log d - \frac{\sqrt{n}}{8\pi \log n}\right) + 4 \exp\left(- \frac{k_1 n^{3/4} \sqrt{\log n}}{k_2 + k_3}\right) + \frac{2}{\sqrt{\pi \log(n \epsilon^2)}}.
\]

The proof of the theorem is given in Section 5 of the Supplementary Materials. With regards to the scaling of the dimension in terms of sample size the ensuing corollary follows immediately.

**Corollary 3.4.** For some known constant \( K_\Sigma \) independent of \( d \) and \( n \) we have

\[
P(\max_{j,k} | \hat{\Sigma}^{(n)}_{jk} - \Sigma^*_{jk} | > K_\Sigma \frac{\log d \log n}{\sqrt{n}}) = o(1).
\]  

### 3.6 Estimating the precision matrix

Similar to Fan et al. [26], we plug our estimate of the sample correlation matrix into existing routines for estimating \( \Omega^* \). In particular, we employ the graphical lasso (glasso) estimator [10], i.e.

\[
\hat{\Omega} = \arg \min_{\Omega \succeq 0} \text{tr}(\hat{\Sigma}^{(n)}\Omega) - \log|\Omega| + \lambda \sum_{j \neq k} |\Omega_{jk}|,
\]

where \( \lambda > 0 \) is a regularization parameter. As \( \hat{\Sigma}^{(n)} \) exhibits at worst the same theoretical properties as established in Liu et al. [17], convergence rate and graph selection results follow immediately.

We do not penalize diagonal entries of \( \Omega \) and therefore have to make sure that \( \hat{\Sigma}^{(n)} \) is at least positive semidefinite to establish convergence in Eq. (16). Hence, we need to project \( \hat{\Sigma}^{(n)} \) into the cone of positive semidefinite matrices [compare also 29, 26]. In practice, we use an efficient implementation of the alternating projections method proposed by Higham [41].

In order to select the tuning parameter in Eq. (16) Foygel and Drton [42] introduce an extended BIC (eBIC) in particular for Gaussian graphical models establishing consistency in higher dimensions under mild asymptotic assumptions. We consider

\[
eBIC_\theta = -2\ell^{(n)}(\hat{\Omega}(E)) + |E| \log(n) + 4 |E| \theta \log(d),
\]

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 \( \ell^{(n)}(\hat{\Omega}(E)) \) denotes the corresponding maximized log-likelihood [see 42, for more details] which in turn depends on \( \lambda \) from Eq. (16).

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 of the models in the path and select the one with the minimal value.

### 4 Experimental Setup

In order to numerically assess the accuracy of our mixed graph estimation approach we begin with a simulation study in which the estimators can be assessed in a gold-standard fashion and compared against oracles.

**Simulation strategy.** To facilitate comparison, we follow a similar data-generating strategy to Fan et al. [26]. We split the experiments into three parts. First, we consider a binary mixed case and benchmark the performance of our approach against Fan et al. [26, Section 6.1 scenarios c) and d)]. Second, we generate a mix of binary-ternary-continuous data. Although Quan et al. [31] do not report any numerical results in their simulation study, we compare our approach to their extension of Fan et al. [26]. Third, from \( Z \) we generate discrete data with arbitrary numbers of levels and compare...
performance with the latent continuous oracle. The results for the binary-ternary-continuous data simulations as well as a detailed description of the simulation setup can be found in Section 6 of the Supplementary Materials. We set the dimension to \( d = (50, 250, 750) \) for sample size \( n = (200, 200, 300) \) and choose \( c \) such that the number of edges is roughly equal to the dimension – except for \( d = 50 \), where we allow for 200 edges in accordance with Fan et al. [26].

**Performance metrics.** To assess performance, we report the mean estimation error \( \| \hat{\Omega} - \Omega^* \|_F \) as evaluated by the Frobenius norm. Furthermore, we consider graph recovery metrics. To this end, we define the number of true positives \( TP(\lambda) \) and false positives \( FP(\lambda) \) depending on the glasso path as the number of nonzero lower off-diagonal elements that agree both in \( \Omega^* \) and \( \hat{\Omega} \) and the number of nonzero lower off-diagonal elements in \( \Omega \) that are actually zero in \( \hat{\Omega}^* \), respectively. The true positive rate \( TPR(\lambda) \) and the false positive rate \( FPR(\lambda) \) are defined as \( TPR = \frac{TP(\lambda)}{TP(\lambda) + FP(\lambda)} \) and \( FPR = \frac{FP(\lambda)}{d(d - 1)/2 - |E|} \), respectively. Lastly, we consider the area under the curve (AUC) where a value of 0.5 corresponds to random guessing of the presence of an edge and a value of 1 corresponds to perfect error-free recovery of the underlying latent graph (in the rank sense of ROC analysis).

5 Simulation Results

5.1 Simulation results: binary-continuous mix

We start by considering a mix of binary and continuous variables generated as outlined in Section 4 in order to be able to compare results with those of Fan et al. [26]. For this purpose, Table 1 reports mean estimation errors \( \| \hat{\Omega} - \Omega^* \|_F \) under the different \((d, n)\) regimes. When for all \( j \), \( f_j(x) = x \) we recover the latent Gaussian and when \( f_j(x) = x^3 \) the latent Gaussian copula model.

The oracle estimator in the third column of Table 1 corresponds to estimating \( \hat{\Sigma}^{(n)} \) from Definition 2.2 based on realization of the latent data \((Z_1, Z_2)\). Next in column four, the binary \( \hat{\Omega}_r \) indicates the nonparanormal estimator proposed by Fan et al. [26]. The next two columns, namely \( \hat{\Omega}_{MLE} \) and \( \hat{\Omega}_r \) indicate the ML approach and the general mixed estimator developed in Sections 2.1 and 3, respectively.

As expected, the estimation error for \( d = 50 \) and \( n = 200 \) in the latent Gaussian setting is almost identical for \( \hat{\Omega}_r \) and our proposed nonparanormal estimator \( \hat{\Omega}_r \). Additionally, \( \hat{\Omega}_{MLE} \) performs best in this case. Compared to the oracle estimator, all three approaches exhibit very little loss in accuracy that arises due to the binarization. Looking at graph recovery in terms of FPR, TPR, and AUC the picture is similar.

Turning to the nonparanormal setting with \( f_j(x) = x^3 \) under \( d = 50 \) and \( n = 200 \), as expected \( \hat{\Omega}_r \) and \( \hat{\Omega}_r \) remain largely unchanged but for small numerical differences. However, for \( \hat{\Omega}_{MLE} \) accuracy both in terms of estimation error and graph recovery drop noticeably. However, the FPR remains unchanged, indicating that whilst detecting fewer correct edges in the graph the number of incorrect edges is not affected by this particular transformation. When increasing the number of variables to \( d = 250 \) and \( d = 750 \) the picture is similar to before. To sum up, when compared to the estimator \( \hat{\Omega}_r \) proposed by Fan et al. [26] both \( \hat{\Omega}_{MLE} \) and \( \hat{\Omega}_r \) perform similarly, under the latent Gaussian assumption even somewhat better. \( \hat{\Omega}_r \) performs slightly better than \( \hat{\Omega}_r \) in all scenarios considered. Loss of accuracy that arises from the discretization is not too severe under all \((d, n)\) regimes considered. Note, that the binary-continuous mix is merely a special case of the general mixed data scheme we consider in our paper. Therefore, being able to show similar or even improved performance to the current gold standard is important.

5.2 Arbitrary mixed data results

We turn to the set of experiments where we generate a mix of continuous and discrete data with arbitrary numbers of levels. As existing approaches, and in particular the bridge function approach, do not extend to this fully general case, we can no longer compare the proposed method to an existing one. Instead in Table 2, we report a second oracle estimator, namely oracle \( \hat{\Omega}_r \), i.e. applying Definition 3.1 to realization from \( Z = (Z_1, X_2) \) in order to get more insight into the cases where \( f_j(x) = x^3 \).

When comparing \( \| \hat{\Omega} - \Omega^* \|_F \) across the different settings, similar to the previous results, efficiency loss is almost negligible.

This is true both for \( \hat{\Omega}_r \) and for \( \hat{\Omega}_{MLE} \) in the latent Gaussian settings. Furthermore, graph recovery in terms of AUC improves overall owing to the fact that more information regarding the latent variable is available. All remaining
| $d, n, f(x)$ | Oracle $\hat{\Omega}$ | binary $\hat{\Omega}_r$ | $\hat{\Omega}_{\text{MLE}}$ | $\hat{\Omega}_r$ |
|-------------|----------------|----------------|----------------|----------------|
| $\|\hat{\Omega} - \Omega\|_F$ | 2.858 | 3.126 | 3.095 | 3.111 |
| FPR | 0.016 | 0.250 | 0.222 | 0.231 |
| TPR | 0.340 | 0.587 | 0.566 | 0.567 |
| AUC | 0.880 | 0.713 | 0.720 | 0.715 |
| $d, n, f(x)$ | 50, 200, $x$ | 50, 200, $x^3$ | 250, 200, $x$ | 250, 200, $x^3$ |
| $\|\hat{\Omega} - \Omega\|_F$ | 2.851 | 3.153 | 3.241 | 3.115 |
| FPR | 0.016 | 0.263 | 0.187 | 0.228 |
| TPR | 0.345 | 0.604 | 0.410 | 0.576 |
| AUC | 0.881 | 0.718 | 0.650 | 0.721 |
| $\|\hat{\Omega} - \Omega\|_F$ | 3.186 | 4.090 | 4.099 | 4.098 |
| FPR | 0.006 | 0.040 | 0.039 | 0.039 |
| TPR | 0.304 | 0.222 | 0.228 | 0.224 |
| AUC | 0.884 | 0.717 | 0.726 | 0.720 |
| $\|\hat{\Omega} - \Omega\|_F$ | 3.194 | 4.086 | 4.287 | 4.105 |
| FPR | 0.006 | 0.040 | 0.042 | 0.040 |
| TPR | 0.304 | 0.223 | 0.144 | 0.226 |
| AUC | 0.883 | 0.716 | 0.649 | 0.719 |
| $\|\hat{\Omega} - \Omega\|_F$ | 11.187 | 10.264 | 10.265 | 10.270 |
| FPR | 0.256 | 0.142 | 0.144 | 0.143 |
| TPR | 0.938 | 0.602 | 0.619 | 0.611 |
| AUC | 0.939 | 0.770 | 0.777 | 0.774 |
| $\|\hat{\Omega} - \Omega\|_F$ | 11.197 | 10.273 | 10.845 | 10.287 |
| FPR | 0.256 | 0.142 | 0.136 | 0.144 |
| TPR | 0.937 | 0.601 | 0.455 | 0.611 |
| AUC | 0.938 | 0.769 | 0.695 | 0.773 |

Table 1: Binary mixed data structure learning; Simulated data with 100 simulation runs. Standard errors in brackets.
| $d, n, f(x)$ | $\|\hat{\Omega} - \Omega\|_F$ | Oracle $\hat{\Omega}$ | Oracle $\hat{\Omega}_\rho$ | $\hat{\Omega}_{MLE}$ | $\hat{\Omega}_r$ |
|------|-------|----------------|----------------|----------------|----------------|
| $50, 200, x$ | $2.860$ | $2.845$ | $2.892$ | $2.890$ |
| | $0.015$ | $0.019$ | $0.042$ | $0.044$ |
| | $0.335$ | $0.343$ | $0.356$ | $0.356$ |
| | $0.881$ | $0.864$ | $0.816$ | $0.811$ |
| $50, 200, x^3$ | $2.869$ | $2.866$ | $3.014$ | $2.912$ |
| | $0.016$ | $0.019$ | $0.044$ | $0.042$ |
| | $0.336$ | $0.338$ | $0.244$ | $0.352$ |
| | $0.879$ | $0.862$ | $0.730$ | $0.807$ |
| $250, 200, x$ | $3.201$ | $3.212$ | $3.503$ | $3.507$ |
| | $0.006$ | $0.008$ | $0.016$ | $0.016$ |
| | $0.297$ | $0.300$ | $0.257$ | $0.255$ |
| | $0.879$ | $0.861$ | $0.816$ | $0.811$ |
| $250, 200, x^3$ | $3.192$ | $3.209$ | $3.708$ | $3.503$ |
| | $0.006$ | $0.008$ | $0.018$ | $0.017$ |
| | $0.300$ | $0.289$ | $0.153$ | $0.255$ |
| | $0.881$ | $0.863$ | $0.734$ | $0.812$ |
| $750, 300, x$ | $11.207$ | $11.154$ | $10.913$ | $10.928$ |
| | $0.257$ | $0.247$ | $0.214$ | $0.213$ |
| | $0.936$ | $0.917$ | $0.835$ | $0.830$ |
| | $0.938$ | $0.925$ | $0.878$ | $0.874$ |
| $750, 300, x^3$ | $11.199$ | $11.165$ | $11.398$ | $10.939$ |
| | $0.256$ | $0.248$ | $0.202$ | $0.214$ |
| | $0.937$ | $0.919$ | $0.687$ | $0.831$ |
| | $0.938$ | $0.925$ | $0.789$ | $0.874$ |

Table 2: General mixed data structure learning; Simulated data with 100 simulation runs. Standard errors in brackets.
characteristics established in the previous scenario translate to the general case. In the nonparanormal settings \( \hat{\Omega}_{\text{MLE}} \) fails to establishing true positives rather than producing more FPs.

In conclusion, the results of the simulation study reveal that both estimators developed in this paper perform favorably when compared to the state of the art. Particularly, the nonparanormal estimator \( \hat{\Omega}_r \) performs well, but is simple, removing the need to derive potentially large numbers of bridge functions to generalize the setting in Fan et al. [26] further. Instead, the polychoric and polyserial correlations agree naturally with the latent Gaussian copula model.

6 Application to COVID-19 data

In this section, we present 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 a variety of potential risk factors. This analysis is intended to illustrate the use of the proposed methods in a real-world, mixed variable type example.

| 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     | -          |

Table 3: Estimated partial correlations between COVID-19 severity and the listed variables for data sets A, B and C.

6.1 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. In order to construct an indicator of COVID-19 severity we consider subjects who were 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 6 weeks of infection (meaning either hospitalised, hospitalised receiving critical care or died). Around 14% experienced such a severe outcome. Overall the analysis includes \( n = 8672 \) observations on \( d = 712 \) variables (risk factors and covariates with less than 40% missingness). Missing values were imputed using \texttt{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 minority class of these indicators (i.e. the fraction of samples with the least frequent value of the variable) can often be very small. To understand the effects of such rare events on the analysis, we defined 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.

6.2 Results

We present results of a joint analysis of the variables considered, using the real data. However, we emphasize that the analysis is aimed at illustrating behaviour of the proposed estimators and not at fully understanding risk factors for severe COVID-19. There has been much work done on factors influencing risk of severe COVID-19 and on its treatment [see, among others, 43, 44] and we direct the interested reader to the references for further information.

Table 3 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. However, 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 43].
Finally, we present more detailed results of the analysis of data set A. Figure 1 shows the estimated adjacency matrix and Figure 2 depicts the estimated precision matrix $\hat{\Omega}_A$. These results highlight the type of output, spanning different kinds of variables, that is readily available from the proposed method.

7 Conclusion

Estimating high-dimensional undirected graphs from general mixed data is a challenging task. We propose an approach for this problem that combines classical, generalized correlation measures, and in particular polychoric and polyserial correlations, with recent ideas from high-dimensional graphical modelling and copulas.

In particular, we make the simple but we think relevant observation that polychoric and polyserial correlations can be usefully considered via a latent Gaussian copula model. While it requires some care to tailor the polyserial correlation to the nonparanormal case, the polychoric correlation does not require any adjustments. The resulting estimators enjoy favorable theoretical properties (also in high dimensions) and show very good empirical performance in our simulation study.

The framework we advocate for builds on a line of work that extends the graphical lasso for Gaussian observations to nonparanormal models and then mixed data as in the work of Fan et al. [26] and later [31] and Feng and Ning [32]. A key distinction is that in our approach there is no need to specify bridge functions, and we can directly cope with general types of mixed data with no additional effort on the user’s part, as we illustrated in an analysis of phenotyping data from the UK Biobank.

8 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 in order to run the simulation study conducted in the paper including a small sample simulation is available under https://github.com/konstantingoe/mixed_hidim_graphs

Figure 1: Plot of the estimated adjacency matrix of data set A.
Supplementary Materials

The reader is referred to the Supplementary Materials for technical appendices, as well as proofs of theorems and lemmas in the main manuscript. Additionally, we present further simulation results and details regarding the real-world data application.

Funding

This work was supported by the Helmholtz AI project “Scalable and Interpretable Models for Complex And Structured Data” (SIMCARD), the Medical Research Council [programme number MC UU 00002/17] and the National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre.

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

Acknowledgments

Conflict of Interest: None declared.

References

[1] Adrian Dobra, Chris Hans, Beatrix Jones, Joseph R. Nevins, Guang Yao, and Mike West. Sparse graphical models for exploring gene expression data. *J. Multivariate Anal.*, 90(1):196–212, 2004.

[2] Michael Finegold and Mathias Drton. Robust graphical modeling of gene networks using classical and alternative t-distributions. *Ann. Appl. Stat.*, 5(2A):1057–1080, 2011.
[3] Ricardo Pio Monti, Peter Hellyer, David Sharp, Robert Leech, Christoforos Anagnostopoulos, and Giovanni Montana. Estimating time-varying brain connectivity networks from functional MRI time series. *NeuroImage*, 103:427–443, 2014.

[4] Z. Wei and H. Li. A markov random field model for network-based analysis of genomic data. *Bioinformatics*, 23(12):1537–1544, 2007.

[5] N. Verzelen and P. Villers. Tests for Gaussian graphical models. *Comput. Statist. Data Anal.*, 53(5):1894–1905, 2009.

[6] Nicolas Stäudler and Sach Mukherjee. Penalized estimation in high-dimensional hidden Markov models with state-specific graphical models. *Ann. Appl. Stat.*, 7(4):2157–2179, 2013.

[7] Nicolas Stäudler and Sach Mukherjee. Multivariate gene-set testing based on graphical models. *Biostatistics*, 16(1):47–59, 2015.

[8] Konstantinos Perrakis, Thomas Lartigue, Frank Dondelinger, and Sach Mukherjee. Regularized joint mixture models, 2019. arXiv:1908.07869.

[9] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 2006.

[10] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2007.

[11] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d’Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.*, 9:485–516, 2008.

[12] Clifford Lam and Jianqing Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.*, 37(6B):4254–4278, 2009.

[13] Ming Yuan. High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.*, 11:2261–2286, 2010.

[14] Pradeep Ravikumar, Martin J. Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing $\ell_1$-penalized log-determinant divergence. *Electron. J. Stat.*, 5:935–980, 2011.

[15] Tony Cai, Weidong Liu, and Xi Luo. A constrained $\ell_1$ minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.*, 106(494):594–607, 2011.

[16] Masashi Miyamura and Yutaka Kano. Robust Gaussian graphical modeling. *J. Multivariate Anal.*, 97(7):1525–1550, 2006.

[17] Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.*, 10:2295–2328, 2009.

[18] M.J. Wainwright and M.J. Jordan. Log-determinant relaxation for approximate inference in discrete markov random fields. *IEEE Transactions on Signal Processing*, 54(6):2099–2109, 2006.

[19] Pradeep Ravikumar, Martin J. Wainwright, and John D. Lafferty. High-dimensional Ising model selection using $\ell_1$-regularized logistic regression. *Ann. Statist.*, 38(3):1287–1319, 2010.

[20] Steffen L. Lauritzen. *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press, Oxford University Press, New York, 1996. ISBN 0-19-852219-3. Oxford Science Publications.

[21] Jie Cheng, Tianxi Li, Elizaveta Levina, and Ji Zhu. High-dimensional mixed graphical models. *J. Comput. Graph. Statist.*, 26(2):367–378, 2017.

[22] Jason D. Lee and Trevor J. Hastie. Learning the structure of mixed graphical models. *J. Comput. Graph. Statist.*, 24(1):230–253, 2015.

[23] Shizhe Chen, Daniela M. Witten, and Ali Shojaie. Selection and estimation for mixed graphical models. *Biometrika*, 102(1):47–64, 2015.

[24] Eunho Yang, Yulia Baker, Pradeep Ravikumar, Genevera Allen, and Zhandong Liu. Mixed Graphical Models via Exponential Families. In Samuel Kaski and Jukka Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 1042–1050, Reykjavik, Iceland, 2014. PMLR.

[25] Zhuoran Yang, Yang Ning, and Han Liu. On semiparametric exponential family graphical models. *J. Mach. Learn. Res.*, 19:Paper No. 57, 59, 2018.

[26] Jianqing Fan, Han Liu, Yang Ning, and Hui Zou. High dimensional semiparametric latent graphical model for mixed data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79(2):405–421, 2017.
[27] Karl Pearson. 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(262-273):1–47, 1900.

[28] Karl Pearson. On the measurement of the influence of "broad categories" on correlation. *Biometrika*, 9(1/2):116–139, 1913.

[29] Han Liu, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman. High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.*, 40(4):2293–2326, 2012.

[30] Lingzhou Xue and Hui Zou. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann. Statist.*, 40(5):2541–2571, 2012.

[31] Xiaoyun Quan, James G. Booth, and Martin T. Wells. Rank-based approach for estimating correlations in mixed ordinal data. 2018. arXiv: 1809.06255.

[32] Huijie Feng and Yang Ning. High-dimensional mixed graphical model with ordinal data: Parameter estimation and statistical inference. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 654–663. PMLR, 2019.

[33] Ulf Olsson, Fritz Drasgow, and Neil J. Dorans. The polychoric correlation coefficient. *Psychometrika*, 47(3):337–347, 1982.

[34] N. R. Cox. Estimation of the correlation between a continuous and a discrete variable. *Biometrics*, 30:171–178, 1974.

[35] Ulf Olsson. Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44(4):443–460, 1979.

[36] Shaobo Jin and Fan Yang-Wallentin. Asymptotic robustness study of the polychoric correlation estimation. *Psychometrika*, 82(1):67–85, 2017.

[37] Edward J. Bedrick. A comparison of generalized and modified sample biserial correlation estimators. *Psychometrika*, 57(2):183–201, 1992.

[38] Edward J. Bedrick and Frederick C. Breslin. Estimating the polyserial correlation coefficient. *Psychometrika*, 61(3):427–443, 1996.

[39] Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *Ann. Statist.*, 46(6A):2747–2774, 2018.

[40] Gégout-Petit Anne, Gueudin-Muller Aurélie, and Karmann Clémence. Graph estimation for Gaussian data zero-inflated by double truncation, 2019. arXiv:1911.07694.

[41] Nicholas J. Higham. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra Appl.*, 103:103–118, 1988.

[42] Rina Foygel and Mathias Drton. Extended bayesian information criteria for Gaussian graphical models. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23, pages 604–612. Curran Associates, Inc., 2010.

[43] Elizabeth J. Williamson, Alex J. Walker, Krishnan Bhaskaran, Seb Bacon, Chris Bates, Caroline E. Morton, Helen J. Curtis, Amir Mehrkar, David Evans, Peter Inglesby, Jonathan Cockburn, Helen I. McDonald, Brian MacKenna, Laurie Tomlinson, Ian J. Douglas, Christopher T. Rentsch, Rohini Mathur, Angel Y. S. Wong, Richard Grieve, David Harrison, Harriet Forbes, Anna Schultzje, Richard Croker, John Parry, Frank Hester, Sam Harper, Rafael Perera, Stephen J. W. Evans, Liam Smeeth, and Ben Goldacre. Factors associated with COVID-19-related death using OpenSAFELY. *Nature*, 584(7821):430–436, 2020.

[44] David A. Berlin, Roy M. Gulick, and Fernando J. Martinez. Severe covid-19. *New England Journal of Medicine*, 383(25):2451–2460, 2020.
Supplementary Materials

1 Methodology

1.1 Case II MLE derivation

Recall case II, where we assume that $X_j$ is ordinal and $X_k$ is continuous and we are interested in the product-moment correlation $\Sigma_{jk}$ between two jointly Gaussian variables, where $X_j$ is not directly observed but only the ordered categories (Eq. (1) in the Manuscript) are given. The likelihood of the $n$-sample is defined by:

$$L^{(n)}(\Sigma_{jk}, x^j, x_k) = \prod_{i=1}^{n} p(x^j_{ir}, x_{ik}, \Sigma_{jk})$$

$$= \prod_{i=1}^{n} p(x_{ik}) p(x^j_{ir} \mid x_{ik}, \Sigma_{jk}),$$

where $p(x^j_{ir}, x_{ik}, \Sigma_{jk})$ 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.

$$p(x_{ik}) = (2\pi\sigma)^{-\frac{1}{2}} \exp \left[ -\frac{1}{2} \left( \frac{x_{ik} - \mu}{\sigma_{jk}} \right)^2 \right].$$
Furthermore, the conditional probability of $X_j$ in Eq. (3) in the Manuscript can be written as:

\[ p(X_j = x_j^r \mid X_k, \Sigma_{jk}) = p(\gamma_{j,r-1} \leq Z_j < \gamma_{j,r} \mid X_k, \Sigma_{jk}) \]
\[ = p(Z_j \leq \gamma_{j,r} \mid X_k, \Sigma_{jk}) - p(Z_j \leq \gamma_{j,r-1} \mid X_k, \Sigma_{jk}) \]
\[ \Phi(\tilde{z}_j) - \Phi(\bar{z}_j), \quad r = 1, \ldots, l_{X_j} - 1, \] (2)

where

\[ \tilde{z}_j = \frac{\gamma_j - \Sigma_{jk} \bar{X}_k}{\sqrt{1 - (\Sigma_{jk})^2}} \]

with $\bar{X}_k = X_k - \bar{X}$ 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 $\Sigma_{jk} \bar{X}_k$ and variance $(1 - (\Sigma_{jk})^2)$. The log-likelihood is then $\ell^{(n)}(\Sigma_{jk}, x_j^r, x_k)$ with:

\[ \ell^{(n)}(\Sigma_{jk}, x_j^r, x_k) = \sum_{i=1}^{n} \left[ \log(p(x_{ik})) + \log(p(x_{ir}^j \mid x_{ik}, \Sigma_{jk})) \right]. \] (3)

Due to the heavy computational burden involved when estimating all parameters simultaneously, a two-step estimator has been proposed Olsson et al. [1]. That is, in a first step the thresholds $\gamma_j^r, r = 1, \ldots, l_{X_j} - 1$ 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 $\Sigma_{jk}$ now with the readily computed estimates from the first step:

\[ \frac{\partial \ell^{(n)}(\Sigma_{jk}, x_j^r, x_k)}{\partial \Sigma_{jk}} = \sum_{i=1}^{n} \frac{1}{p(x_{ir}^j \mid x_{ik}, \Sigma_{jk})} \frac{\partial p(x_{ir}^j \mid x_{ik}, \Sigma_{jk})}{\partial \Sigma_{jk}}. \] (4)

Let us take a closer look at the partial derivative of the conditional probability in Eq. (4):

\[ \frac{\partial p(x_{ir}^j \mid x_{ik}, \Sigma_{jk})}{\partial \Sigma_{jk}} = \frac{\partial \Phi(\tilde{z}_j)}{\partial \Sigma_{jk}} - \frac{\partial \Phi(\bar{z}_j)}{\partial \Sigma_{jk}} \]
\[ = \phi(\tilde{z}_j) \frac{\partial \tilde{z}_j}{\partial \Sigma_{jk}} - \phi(\bar{z}_j) \frac{\partial \bar{z}_j}{\partial \Sigma_{jk}} \]
\[ = (1 - (\Sigma_{jk})^2)^{-\frac{3}{2}} \left[ \phi(\tilde{z}_j)(\gamma_j \Sigma_{jk} - \bar{X}_k) - \phi(\bar{z}_j)(\gamma_j \Sigma_{jk} - \bar{X}_k) \right], \] (5)

where $\bar{X}_k = X_k - \bar{X}$ and $\tilde{z}_j = \frac{\gamma_j - \Sigma_{jk} \bar{X}_k}{\sqrt{1 - (\Sigma_{jk})^2}}$. The last equality follows from taking the derivative and applying the chain-rule.

Hence putting all the pieces together we obtain:

\[ \frac{\partial \ell^{(n)}(\Sigma_{jk}, x_j^r, x_k)}{\partial \Sigma_{jk}} = \sum_{i=1}^{n} \frac{1}{p(x_{ir}^j \mid x_{ik}, \Sigma_{jk})} (1 - (\Sigma_{jk})^2)^{-\frac{3}{2}} \]
\[ \phi(\tilde{z}_j)(\gamma_j \Sigma_{jk} - \bar{X}_k) - \phi(\bar{z}_j)(\gamma_j \Sigma_{jk} - \bar{X}_k) \]
\[ = \frac{1}{p(x_{ir}^j \mid x_{ik}, \Sigma_{jk})} \phi(z_j, z_k, \Sigma_{jk}) dz_j dz_k, \] (6)

1.2 Case III MLE derivation

Consider case III, where both $X_j$ and $X_k$ are ordinal variables. The probability of an observation with $X_j = x_j^r$ and $X_k = x_k^s$ is:

\[ \pi_{rs} := p(X_j = x_j^r, X_k = x_k^s) \]
\[ = p(\gamma_{j,r-1} \leq Z_j < \gamma_{j,r} \mid X_k = x_k^s, \Sigma_{jk}) \]
\[ = \int_{\gamma_{j,r-1}}^{\gamma_{j,r}} \int_{\gamma_{s,r-1}}^{\gamma_{s,r}} \phi(z_j, z_k, \Sigma_{jk}) dz_j dz_k, \] (7)
where \( r = 1, \ldots, l \), \( s = 1, \ldots, l \), and \( \phi(x, y, \rho) \) denotes the standard bivariate density with correlation \( \rho \). Then, as in the main text the likelihood and log-likelihood of the \( n \)-sample are defined as:

\[
L^{(n)}(\Sigma_{jk}, x^j_i, x^k_s) = C \prod_{r=1}^{l_x} \prod_{s=1}^{l_x} \pi_{rs}^{n_{rs}},
\]

\[
\ell^{(n)}(\Sigma_{jk}, x^j_i, x^k_s) = \log(C) + \sum_{r=1}^{l_x} \sum_{s=1}^{l_x} n_{rs} \log(\pi_{rs}).
\]

where \( C \) is a constant and \( n_{rs} \) denotes the observed frequency of \( X_j = x^j_i \) and \( X_k = x^k_s \) in a sample of size \( n = \sum_{r=1}^{l_x} \sum_{s=1}^{l_x} n_{rs} \). 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 \( \ell^{(n)}(\Sigma_{jk}, x^j_i, x^k_s) \) with respect to \( \Sigma_{jk} \) explicitly. First, recall that from Eq. (5)

\[
\pi_{rs} = \int_{\gamma_{r-1}}^{\gamma_r} \int_{\gamma_{s-1}}^{\gamma_s} \phi(z_j, z_k, \Sigma_{jk}) dz_j dz_k
\]

\[
= \Phi_2(\gamma_r^j, \gamma_s^k, \Sigma_{jk}) - \Phi_2(\gamma_{r-1}^j, \gamma_s^k, \Sigma_{jk}) - \Phi_2(\gamma_r^j, \gamma_{s-1}^k, \Sigma_{jk}) + \Phi_2(\gamma_{r-1}^j, \gamma_{s-1}^k, \Sigma_{jk}),
\]

where \( \Phi_2(u, v, \rho) \) is the standard bivariate normal distribution function with correlation parameter \( \rho \). Note also that we have \( \frac{\partial \Phi_2(u, v, \rho)}{\partial \rho} = \phi_2(u, v, \rho) \), where \( \phi_2 \) is the bivariate normal density function [2].

Thus, taking the derivative of \( \ell^{(n)}(\Sigma_{jk}, x^j_i, x^k_s) \) with respect to \( \Sigma_{jk} \) yields

\[
2 \frac{\partial \ell^{(n)}(\Sigma_{jk}, x^j_i, x^k_s)}{\partial \Sigma_{jk}} = \sum_{r=1}^{l_x} \sum_{s=1}^{l_x} n_{rs} \frac{\partial \pi_{rs}}{\partial \Sigma_{jk}}
\]

\[
= \sum_{r=1}^{l_x} \sum_{s=1}^{l_x} n_{rs} \left[ \phi_2(\gamma_r^j, \gamma_s^k, \Sigma_{jk}) - \phi_2(\gamma_{r-1}^j, \gamma_s^k, \Sigma_{jk}) - \phi_2(\gamma_r^j, \gamma_{s-1}^k, \Sigma_{jk}) + \phi_2(\gamma_{r-1}^j, \gamma_{s-1}^k, \Sigma_{jk}) \right].
\]

2 Proof of Theorem 3.2

Condition 2.1 (Gradient statistical noise). The gradient of the log-likelihood function is \( \tau^2 \)-sub-Gaussian. That is, for any \( \lambda \in \mathbb{R} \) and \( \forall \Sigma_{jk} \in [-1 + \delta, 1 - \delta] \) for \( 1 \leq j < k \leq d \)

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

where \( \ell_{jk} \) corresponds to the respective log-likelihood functions in Definitions 2.3 and 2.4 of the main Manuscript.

Let us consider case II, where we recall that

\[
\frac{\partial \ell(\Sigma_{jk}, x^j_i, x^k_s)}{\partial \Sigma_{jk}} = \frac{1}{p(x^j_i | x_{ik}, \Sigma_{jk})} \frac{\partial p(x^j_i | x_{ik}, \Sigma_{jk})}{\partial \Sigma_{jk}}.
\]

Thus:

\[
\frac{\partial \ell(\Sigma_{jk}, x^j_i, x^k_s)}{\partial \Sigma_{jk}} = \frac{(1 - (\Sigma_{jk})^2)^{-\frac{3}{2}}}{\Phi(\gamma_r^j) - \Phi(\gamma_{r-1}^j)} \left[ \phi(\tilde{x}_r^j)(\gamma_r^j \Sigma_{jk} - \tilde{x}_k) - \phi(\gamma_{r-1}^j)(\gamma_{r-1}^j \Sigma_{jk} - \tilde{x}_k) \right].
\]

It is easy to see that \( p(x^j_i | x_{ik}, \Sigma_{jk}) \in (0, 1) \) almost surely. Assumption 3.3 makes sure that we exclude impossible events where \( p(x^j_i | x_{ik}, \Sigma_{jk}) = 0 \). Moreover, we require that \( \gamma_r^j > \gamma_{r-1}^j, \forall j = 1, \ldots, d \) this implies that \( \Phi(\gamma_r^j) > \Phi(\gamma_{r-1}^j) \). In other words, there exists a \( \kappa > 0 \) such that \( p(x^j_i | x_{ik}, \Sigma_{jk}) \leq \frac{1}{\kappa} \).
Let us now turn to $\partial p(x^j_i \mid x_{ik}, \Sigma_{jk}) / \partial \Sigma_{jk}$. 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. $|\phi(t)| \leq (2\pi)^{-\frac{1}{2}}$ for all $x \in \mathbb{R}$. Similarly

$$|\phi(\tilde{\gamma}^j_i)(\gamma^j_i \Sigma_{jk} - \tilde{x}_k) - \phi(\tilde{\gamma}^{j-1}_i)(\gamma^{j-1}_i \Sigma_{jk} - \tilde{x}_k)| \leq |\phi(\tilde{\gamma}^j_i)(\gamma^j_i \Sigma_{jk} - \tilde{x}_k)| \leq L_1,$$

due to Assumption 3.2. Therefore,

$$\left| \frac{\partial \ell(\Sigma_{jk}, x^j_i, x_k)}{\partial \Sigma_{jk}} \right| \leq \kappa L_1,$$

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 (1963) lemma the gradient of the log-likelihood function is $\tau^2$-sub-Gaussian with $\tau = 2\kappa L_1$

Turning to case III, recall that we have:

$$\pi_{rs} = \Phi_2(\gamma^j_i, \gamma^k_i, \Sigma_{jk}) - \Phi_2(\gamma^j_{r-1}, \gamma^k_s, \Sigma_{jk})$$

$$- \Phi_2(\gamma^j_i, \gamma^k_{s-1}, \Sigma_{jk}) + \Phi_2(\gamma^j_{r-1}, \gamma^k_{s-1}, \Sigma_{jk}),$$

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

Now let us show that

$$\frac{\partial \pi_{rs}}{\partial \Sigma_{jk}} = \left[ \phi_2(\gamma^j_i, \gamma^k_i, \Sigma_{jk}) - \phi_2(\gamma^j_{r-1}, \gamma^k_s, \Sigma_{jk})$$

$$- \phi_2(\gamma^j_i, \gamma^k_{s-1}, \Sigma_{jk}) + \phi_2(\gamma^j_{r-1}, \gamma^k_{s-1}, \Sigma_{jk}) \right]$$

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 $|\phi_2(x, y)| \leq c$. 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 $\tau^2$-sub-Gaussian with $\tau = 2\kappa K_1$.

From these arguments follows that the gradient statistical noise condition is satisfied.

**Condition 2.2** (Hessian statistical noise). The hessian of the log-likelihood function is $\tau^2$-sub-exponential, i.e. for all $\Sigma_{jk} \in [-1+\delta, 1-\delta]$ and for $1 \leq j < k \leq d$:

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

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

$$\|X\|_{\psi_1} := \sup_{p \geq 1} \frac{1}{p} \mathbb{E} \left( \left| X - \mathbb{E}(X) \right|^p \right)^{\frac{1}{p}}.$$

Again, $\ell_{jk}$ corresponds to the respective log-likelihood functions in Definitions 2.3 and 2.4.

Let us start with case II. We have

$$\frac{\partial^2 \ell_{jk}}{\partial \Sigma_{jk}^2} = \frac{\partial^2 p(x^j_i \mid x_k, \Sigma_{jk}) / \partial \Sigma_{jk}^2}{p(x^j_i \mid x_k, \Sigma_{jk})} - \left( \frac{\partial p(x^j_i \mid x_k, \Sigma_{jk}) / \partial \Sigma_{jk}}{p(x^j_i \mid x_k, \Sigma_{jk})} \right)^2$$

Clearly,

$$\left| \frac{\partial^2 p(x^j_i \mid x_k, \Sigma_{jk}) / \partial \Sigma_{jk}^2}{p(x^j_i \mid x_k, \Sigma_{jk})} \right| \leq \left| \frac{\partial p(x^j_i \mid x_k, \Sigma_{jk}) / \partial \Sigma_{jk}}{p(x^j_i \mid x_k, \Sigma_{jk})} \right| + \left( \frac{\partial p(x^j_i \mid x_k, \Sigma_{jk}) / \partial \Sigma_{jk}}{p(x^j_i \mid x_k, \Sigma_{jk})} \right)^2$$

$$\leq \kappa L_2 + \kappa^2 L_1^2,$$
where it remains to show that $|\frac{\partial^2 p(x^j_{i_r} \mid x_{i_k}, \Sigma_{jk})}{\partial \Sigma_{jk}^2}| \leq L_2$. Indeed, we can rewrite our objective as:

$$
\frac{\partial}{\partial \Sigma_{jk}} \left( \frac{\partial \ell(x_{i_r}, x^j_{i_r}, x_k)}{\partial \Sigma_{jk}} \right)
= \frac{\partial}{\partial \Sigma_{jk}} \left( 1 - (\Sigma_{jk})^2 \right)^{-\frac{3}{2}} \left[ \phi(\gamma^j_{r_2}) (\gamma^j_{r_2} \Sigma_{jk} - \bar{x}_k) - \phi(\gamma^j_{r_1}) (\gamma^j_{r_1} \Sigma_{jk} - \bar{x}_k) \right]
= \frac{3\Sigma_{jk}}{1 - \Sigma_{jk}^2} \left( 1 - (\Sigma_{jk})^2 \right)^{-\frac{3}{2}} \phi(\gamma^j_{r_2}) (\gamma^j_{r_2} \Sigma_{jk} - \bar{x}_k) + \frac{\phi'(\gamma^j_{r_2}) (\gamma^j_{r_2} \Sigma_{jk} - \bar{x}_k)}{(1 - \Sigma_{jk})^3}
+ \frac{\phi(\gamma^j_{r_1}) (\gamma^j_{r_1} \Sigma_{jk} - \bar{x}_k)}{(1 - \Sigma_{jk})^{\frac{3}{2}}}
$$

(13)

Thus:

$$
\left| \frac{\partial}{\partial \Sigma_{jk}} \left( \frac{\partial \ell(x_{i_r}, x^j_{i_r}, x_k)}{\partial \Sigma_{jk}} \right) \right| \leq \left| \frac{3\Sigma_{jk}}{1 - \Sigma_{jk}^2} \left( 1 - (\Sigma_{jk})^2 \right)^{-\frac{3}{2}} \phi(\gamma^j_{r_2}) (\gamma^j_{r_2} \Sigma_{jk} - \bar{x}_k) \right|
+ \frac{\phi'(\gamma^j_{r_2}) (\gamma^j_{r_2} \Sigma_{jk} - \bar{x}_k)}{(1 - \Sigma_{jk})^3}
+ \frac{\phi(\gamma^j_{r_1}) (\gamma^j_{r_1} \Sigma_{jk} - \bar{x}_k)}{(1 - \Sigma_{jk})^{\frac{3}{2}}} \leq L_2,
$$
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. (12) is in fact valid 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 $2(\kappa L_2 + \kappa^2 L_1^2)$. This implies that for all $p \geq 1$

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

Finally, for $\tau = 2\kappa L_1$ we can choose $L_1$ and $\kappa$ such that $2(\kappa L_2 + \kappa^2 L_1^2) \leq \tau^2 = 4\kappa^2 L_1^2$ and the Hessian statistical noise-condition for case II is satisfied.

Let us consider case III:

$$
\frac{\partial^2 \ell(x_{i_r}, x^j_{i_r}, x_k)}{\partial \Sigma_{jk}^2} = \frac{\partial^2 \pi_{rs}}{\partial \Sigma_{jk}^2} \left( \frac{\partial \pi_{rs}}{\partial \Sigma_{jk}} \right) = \frac{\partial^2 \pi_{rs}}{\partial \Sigma_{jk}^2} \left( \frac{\partial \pi_{rs}}{\partial \Sigma_{jk}} \right)^2
$$

(15)

Thus:

$$
\left| \frac{\partial^2 \ell(x_{i_r}, x^j_{i_r}, x_k)}{\partial \Sigma_{jk}^2} \right| \leq \left| \frac{\partial^2 \pi_{rs}}{\partial \Sigma_{jk}^2} \right| \left( \left| \frac{\partial \pi_{rs}}{\partial \Sigma_{jk}} \right| \right)^2
\leq \xi K_2 + \xi^2 K_1^2.
$$

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

$$
\left| \frac{\partial}{\partial \Sigma_{jk}} \left( \frac{\partial \pi_{rs}}{\partial \Sigma_{jk}} \right) \right|
= \left| \frac{\partial}{\partial \Sigma_{jk}} \phi_2(\gamma^j_{r_2}, \gamma^k_{s_2}, \Sigma_{jk}) - \phi_2(\gamma^j_{r_1}, \gamma^k_{s_1}, \Sigma_{jk}) \right|
- \phi_2(\gamma^j_{r_1}, \gamma^k_{s_1}, \Sigma_{jk}) + \phi_2(\gamma^j_{r_1}, \gamma^k_{s_1}, \Sigma_{jk})
\leq \left| \frac{\partial}{\partial \Sigma_{jk}} \phi_2(\gamma^j_{r_2}, \gamma^k_{s_2}, \Sigma_{jk}) \right| + \left| \frac{\partial}{\partial \Sigma_{jk}} \phi_2(\gamma^j_{r_1}, \gamma^k_{s_1}, \Sigma_{jk}) \right|
+ \left| \frac{\partial}{\partial \Sigma_{jk}} \phi_2(\gamma^j_{r_1}, \gamma^k_{s_1}, \Sigma_{jk}) \right| + \left| \frac{\partial}{\partial \Sigma_{jk}} \phi_2(\gamma^j_{r_1}, \gamma^k_{s_1}, \Sigma_{jk}) \right|
\leq K_2,
$$

5
where each of the derivatives of the bivariate density are clearly bounded since they are continuous functions tending to zero at infinity.

Parallel to case II, for \( \tau = 2\xi K_1 \) we can choose \( K_1 \) and \( \xi \) such that \( 2(\xi k_2 + \xi^2 k_1^2) \leq \tau^2 = 4\xi^2 K_1^2 \) and the Hessian statistical noise-condition for case III is satisfied. This then validates the Hessian statistical noise-condition.

With regard to third condition we introduce some additional notation. Let the sample risk be denoted by \( \hat{R}_n(\Sigma_{jk}) \). In order to avoid too many subscripts, \( \hat{R}_n(\Sigma_{jk}) \) represents the sample risk both in case II and case III, i.e.

\[
\hat{R}_n(\Sigma_{jk}) = \frac{1}{n} \sum_{i=1}^{n} \left[ \log(p(x_{jk})) + \log(p(x_{jr}^{1} | x_{ik}, \Sigma_{jk})) \right],
\]

for case II and

\[
\hat{R}_n(\Sigma_{jk}) = \frac{1}{n} \sum_{r=1}^{l_{x_j}} \sum_{s=1}^{l_{x_k}} n_{rs} \log(\pi_{rs})
\]

for case III. Lastly, we define \( R(\Sigma_{jk}) = \mathbb{E}_{\Sigma_{jk}} \hat{R}_n(\Sigma_{jk}) \) to be the population risk for each of the respective cases. Now we are ready to confirm the third condition.

**Condition 2.3 (Hessian regularity).** The Hessian regularity condition consists of three parts:

1. The second derivative of the population risk \( R(\Sigma_{jk}) \) is bounded at one point. That is, there exists one \( |\Sigma_{jk}| \leq 1 - \delta \) and \( H > 0 \) such that \( |R''(\Sigma_{jk})| \leq H \).

2. The second derivative of the log-likelihood with respect to \( \Sigma_{jk} \) is Lipschitz continuous with integrable Lipschitz constant, i.e. there exists a \( M^* > 0 \) such that \( \mathbb{E}[M] \leq M^* \), where

\[
M = \sup_{|\Sigma^{(1)}_{jk}|, |\Sigma^{(2)}_{jk}| \leq 1 - \delta, \Sigma^{(1)}_{jk} \neq \Sigma^{(2)}_{jk}} \frac{|\ell''(\Sigma^{(1)}_{jk}) - \ell''(\Sigma^{(2)}_{jk})|}{|\Sigma^{(1)}_{jk} - \Sigma^{(2)}_{jk}|}.
\]

3. The constants \( H \) and \( M^* \) are such that \( H \leq \tau^2 \) and \( M^* \leq \tau^3 \).

We need some intermediate results that make it easier to deal with \( R(\Sigma_{jk}) \). First, note that \( \mathbb{E}_{\Sigma_{jk}} \hat{R}_n(\Sigma_{jk}) = \mathbb{E}_{\Sigma_{jk}} \ell(\Sigma_{jk}) \). Second, for all \( \Sigma_{jk} \in [-1 + \delta, 1 - \delta] \), for \( 1 \leq j < k \leq d \), and for \( m \in 1, 2 \)

\[
R^m(\Sigma_{jk}) = \frac{\partial^m}{\partial \Sigma_{jk}^m} \mathbb{E}_{\Sigma_{jk}} \ell(\Sigma_{jk}) = \mathbb{E}_{\Sigma_{jk}} \frac{\partial^m}{\partial \Sigma_{jk}^m} \ell(\Sigma_{jk}),
\]

by Lemma 3.3 and Corollary 3.4.

Starting with the first part of the Hessian regularity condition, recall that by Eq. (12) and Eq. (15) for all \( \Sigma_{jk} \in [-1 + \delta, 1 - \delta] \) we have \( \left| \frac{\partial^2}{\partial \Sigma_{jk}^2} \ell(\Sigma_{jk}) \right| \leq \kappa L_2 + \kappa^2 L_1^2 \) and \( \left| \frac{\partial^2}{\partial \Sigma_{jk}^2} \ell(\Sigma_{jk}) \right| \leq \xi K_2 + \xi^2 K_1^2 \) for cases II and III, respectively. Clearly, then any \( |\Sigma_{jk}| \leq 1 - \delta \) and \( H = \kappa L_2 + \kappa^2 L_1^2 \) and \( H = \xi K_2 + \xi^2 K_1^2 \) for cases II and III, respectively satisfy the requirement of the first part. What’s more we also have \( H \leq \tau^2 = 4\kappa L_1^2 \) and \( H \leq \tau^3 = 4\xi K_1^2 \) for both our cases.

The second part requires that the second derivative of the log-likelihood with respect to \( \Sigma_{jk} \) 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.
Let us start with case II, where we have:
\[
\frac{\partial^3}{\partial \Sigma_{jk}^3} \ell(\Sigma_{jk}) = \frac{\partial}{\partial \Sigma_{jk}} \left[ \frac{\partial^2 \ell_{jk}}{\partial \Sigma_{jk}^2} \right]
\]
\[
= \frac{\partial}{\partial \Sigma_{jk}} \left[ \frac{\partial^2 \rho(x_r \mid x_k, \Sigma_{jk})/\partial \Sigma_{jk}^2}{\rho(x_r \mid x_k, \Sigma_{jk})} \right] - \left( \frac{\partial \rho(x_r \mid x_k, \Sigma_{jk})/\partial \Sigma_{jk}}{\rho(x_r \mid x_k, \Sigma_{jk})} \right)^2
\]
\[
= \frac{\partial^3 \rho(x_r \mid x_k, \Sigma_{jk})/\partial \Sigma_{jk}^3}{\rho(x_r \mid x_k, \Sigma_{jk})} - 3 \left( \frac{\partial \rho(x_r \mid x_k, \Sigma_{jk})/\partial \Sigma_{jk}}{\rho(x_r \mid x_k, \Sigma_{jk})} \right) \left( \frac{\partial^2 \rho(x_r \mid x_k, \Sigma_{jk})/\partial \Sigma_{jk}^2}{\rho(x_r \mid x_k, \Sigma_{jk})} \right)
\]
\[
+ 2 \left( \frac{\partial \rho(x_r \mid x_k, \Sigma_{jk})/\partial \Sigma_{jk}}{\rho(x_r \mid x_k, \Sigma_{jk})} \right)^3
\]

Hence:
\[
\left| \frac{\partial^3}{\partial \Sigma_{jk}^3} \ell(\Sigma_{jk}) \right| \leq \kappa L_3 + 3\kappa^2 L_2 L_1 + 2\kappa^3 L_1^3.
\]

It remains to show therefore, that \( \left| \frac{\partial^3 \rho(x_r \mid x_k, \Sigma_{jk})/\partial \Sigma_{jk}^3} \right| \leq L_3 \). Due to the candid but tedious nature of the expression when taking the derivative of Eq. (13) we will merely argue that the resulting statement is clearly 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 \( M \leq \kappa L_3 + 3\kappa^2 L_2 L_1 + 2\kappa^3 L_1^3 \) and the natural choice for \( M^* = \kappa L_3 + 3\kappa^2 L_2 L_1 + 2\kappa^3 L_1^3 \) where we have \( M^* \leq \tau^3 = 8\kappa^3 L_1^3 \).

For case III we proceed similarly by considering:
\[
\frac{\partial^3}{\partial \Sigma_{jk}^3} \ell(\Sigma_{jk}) = \frac{\partial}{\partial \Sigma_{jk}} \left[ \frac{\partial^2 \ell_{jk}(\Sigma_{jk}, x^r, x^k)}{\partial \Sigma_{jk}^2} \right]
\]
\[
= \frac{\partial}{\partial \Sigma_{jk}} \left[ \frac{\partial^2 \pi_{rs} / \partial \Sigma_{jk}^2}{\pi_{rs}} - \left( \frac{\partial \pi_{rs} / \partial \Sigma_{jk}}{\pi_{rs}} \right)^2 \right]
\]
\[
= \frac{\partial^3 \pi_{rs} / \partial \Sigma_{jk}^3}{\pi_{rs}} - 3 \left( \frac{\partial \pi_{rs} / \partial \Sigma_{jk}}{\pi_{rs}} \right) \left( \frac{\partial^2 \pi_{rs} / \partial \Sigma_{jk}^2}{\pi_{rs}^2} \right)
\]
\[
+ 2 \left( \frac{\partial \pi_{rs} / \partial \Sigma_{jk}}{\pi_{rs}} \right)^3.
\]

Hence:
\[
\left| \frac{\partial^3}{\partial \Sigma_{jk}^3} \ell(\Sigma_{jk}) \right| \leq |\xi K_3 + 3\xi^2 K_2 K_1 + 2\xi^3 K_1^3|.
\]

Taking a closer look at \( \frac{\partial^3 \pi_{rs} / \partial \Sigma_{jk}^3} \), boundedness again follows from the fact that the quadratic function in the exponential of the bivariate normal density does not vanish. We thus have \( M \leq |\xi K_3 + 3\xi^2 K_2 K_1 + 2\xi^3 K_1^3| \) and the natural choice for \( M^* = |\xi K_3 + 3\xi^2 K_2 K_1 + 2\xi^3 K_1^3| \) where we have \( M^* \leq \tau^3 = 8\xi^3 K_1^3 \). This validates the Hessian regularity condition.

**Condition 2.4 (Population risk is strongly Morse).** There exist \( \epsilon > 0 \) and \( \eta > 0 \) such that \( R(\Sigma_{jk}) \) is \( (\epsilon, \eta) \)-strongly Morse, i.e.

1. For all \( \Sigma_{jk} \) such that \( |\Sigma_{jk}| = 1 - \delta \) we have that \( |R'(\Sigma_{jk})| > \epsilon \).
2. For all \( \Sigma_{jk} \) such that \( |\Sigma_{jk}| \leq 1 - \delta \): \( |R'(\Sigma_{jk})| \leq \epsilon \implies |R''(\Sigma_{jk})| \geq \eta \).

Put differently, \( R(\Sigma_{jk}) \) is \( (\epsilon, \eta) \)-strongly Morse if the boundaries \(-1 + \delta \) and \( 1 - \delta \) are not critical points of \( R(\Sigma_{jk}) \) and moreover if \( R(\Sigma_{jk}) \) 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 3.3 and Corollary 3.4 we can rewrite $R''(\Sigma_{jk})$ and obtain

$$R''(\Sigma_{jk}) = E_{\Sigma_{jk}} \left[ \frac{\partial^2 \ell(\Sigma_{jk})}{\partial \Sigma_{jk}^2} \right] = E_{\Sigma_{jk}} \left\{ \begin{array}{ll}
\frac{\partial^2 p(x_j^t | x_k, \Sigma_{jk}^*)}{\partial \Sigma_{jk}^2} p(x_j^t | x_k, \Sigma_{jk}^*) & \text{for case II,} \\
\frac{\partial^2 \pi(\Sigma_{jk})_{r,s} / \partial \Sigma_{jk}^2}{\pi(\Sigma_{jk})_{r,s}} - \left( \frac{\partial \pi(\Sigma_{jk})_{r,s} / \partial \Sigma_{jk}}{\pi(\Sigma_{jk})_{r,s}} \right)^2 & \text{for case III,}
\end{array} \right.$$ 

where in $\pi(\Sigma_{jk})_{r,s}$ we made the dependence on $\Sigma_{jk}$ explicit. Note, that for case II we have

$$E_{\Sigma_{jk}} \left[ \frac{\partial^2 p(x_j^t | x_k, \Sigma_{jk}^*)}{\partial \Sigma_{jk}^2} \right] = \int_{-\infty}^{\infty} \sum_{r=1}^{l_{x_j}} \frac{\partial^2 p(x_j^t | x_k, \Sigma_{jk}^*)}{\partial \Sigma_{jk}^2} p(x_j^t | x_k, \Sigma_{jk}^*) dx_k$$

$$= \sum_{r=1}^{l_{x_j}} \frac{\partial^2 p(x_j^t | x_k, \Sigma_{jk}^*)}{\partial \Sigma_{jk}^2}$$

with

$$\sum_{r=1}^{l_{x_j}} \frac{\partial^2 p(x_j^t | x_k, \Sigma_{jk}^*)}{\partial \Sigma_{jk}^2}$$

$$= \sum_{r=1}^{l_{x_j}} \left[ \frac{3 \Sigma_{jk}}{1 - \Sigma_{jk}^2} (1 - (\Sigma_{jk})^2)^{-\frac{3}{2}} \phi(\hat{\theta}_j^r \Sigma_{jk} - \hat{x}_k) - \frac{\phi'(\hat{\theta}_j^r \Sigma_{jk} - \hat{x}_k)}{(1 - \Sigma_{jk}^2)^{\frac{3}{2}}} + \frac{\phi(\hat{\theta}_j^r \gamma r)}{(1 - \Sigma_{jk}^2)^{-\frac{3}{2}}} \right.$$

$$\left. - \frac{3 \Sigma_{jk}}{1 - \Sigma_{jk}^2} (1 - (\Sigma_{jk})^2)^{-\frac{3}{2}} \phi(\hat{\theta}_j^{r-1} \Sigma_{jk} - \hat{x}_k) - \frac{\phi'(\hat{\theta}_j^{r-1} \Sigma_{jk} - \hat{x}_k)}{(1 - \Sigma_{jk}^2)^{\frac{3}{2}}} + \frac{\phi(\hat{\theta}_j^{r-1} \gamma_{r-1})}{(1 - \Sigma_{jk}^2)^{-\frac{3}{2}}} \right]$$

$$= 0,$$
since all terms except the ones involving \( \phi(\tilde{\gamma}_j^r) \) and \( \phi(\tilde{\gamma}_l^j) \) cancel and furthermore \( \lim_{t \to \pm \infty} \phi(t) = \lim_{t \to \pm \infty} \phi'(t) = 0 \). Similarly, we have for case III:

\[
\mathbb{E}_{\Sigma_{jk}^*} \left[ \frac{\partial^2 \pi(x_r^j, x_s^k; \Sigma_{jk})}{\partial \Sigma_{jk}^2} \right] = \sum_r \sum_s \left[ \frac{\partial^2 \pi(x_r^j, x_s^k; \Sigma_{jk})}{\partial \Sigma_{jk}^2} \right] P(X_j = x_r^j, X_k = x_s^k)
\]

\[
= \sum_r \sum_s \left[ \frac{\partial^2 \pi(x_r^j, x_s^k; \Sigma_{jk})}{\partial \Sigma_{jk}^2} \right] P(X_j = x_r^j, X_k = x_s^k)
\]

\[
= \sum_r \sum_s \left[ q(\gamma_r^j, \gamma_s^k, \Sigma_{jk}) \phi_2(\gamma_r^j, \gamma_s^k, \Sigma_{jk}) - q(\gamma_r^{j-1}, \gamma_s^k, \Sigma_{jk}) \phi_2(\gamma_r^{j-1}, \gamma_s^k, \Sigma_{jk}) \right]
\]

\[
= \sum_r \sum_s \left[ q(\gamma_r^j, \gamma_{s-1}^k, \Sigma_{jk}) \phi_2(\gamma_r^j, \gamma_{s-1}^k, \Sigma_{jk}) + q(\gamma_r^{j-1}, \gamma_{s-1}^k, \Sigma_{jk}) \phi_2(\gamma_r^{j-1}, \gamma_{s-1}^k, \Sigma_{jk}) \right]
\]

\[
= \sum_r \sum_s \left[ q(\gamma_r^j, \gamma_{s-1}^k, \Sigma_{jk}) \phi_2(\gamma_r^j, \gamma_{s-1}^k, \Sigma_{jk}) + q(\gamma_r^{j-1}, \gamma_{s-1}^k, \Sigma_{jk}) \phi_2(\gamma_r^{j-1}, \gamma_{s-1}^k, \Sigma_{jk}) \right] = 0,
\]

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

This means that \( R''(\Sigma_{jk}^*) \) can only be zero if \( \partial \pi(x_r^j \mid x_k, \Sigma_{jk})/\partial \Sigma_{jk} \) for case II and \( \partial \pi(\Sigma_{jk}^*)_{r,s}/\partial \Sigma_{jk} \) for case III were zero. But this is not possible due to Assumptions 3.2 and 3.3. To see this note that in Eq. (5) \( \partial \pi(x_r^j \mid x_k, \Sigma_{jk})/\partial \Sigma_{jk} \) can only be zero if either \( \gamma_j^r = \gamma_j^{r-1} \) which we ruled out in Eq. (1) or if \( \gamma_j^r = \infty \) which is ruled out by Assumption 3.2. If we had \( r = \{0, L\} \) then we would not observe any discrete states. Assumption 3.3 rules this case out. Consequently, there exist \( \epsilon > 0 \) and \( \eta > 0 \) such that \( R(\Sigma_{jk}) \) is \( (\epsilon, \eta) \)-strongly Morse

With these considerations, we have verified the required four conditions to hold such that Theorem 3.2 is applicable for each couple \((j, k)\) with \( j < k \). More precisely, let \( \alpha \in (0, 1) \). Now, letting \( n \geq 4C \log(n) \log(\frac{L}{\tau}) \) where \( C = C_0 \left( \frac{n}{\tau} \vee \frac{n}{\eta} \vee \frac{n^2}{\eta^2} \right) \) and \( B = \tau(1 - \delta) \) with \( \tau = 2[\kappa L_1 \vee \xi K_1] \) and \( C_0 \) denoting a universal constant. Letting further \( L = \sup_{\Sigma_{jk} \mid |\Sigma_{jk}| \leq 1 - \delta} |R''(\Sigma_{jk})| \) we obtain

\[
P \left( \left| \frac{\Sigma_{jk}^* - \Sigma_{jk}^*}{\Sigma_{jk}^*} \right| \leq \frac{2\tau}{\eta} \left[ C_0 \log(n) \log \left( \frac{\tau(1 - \delta)}{\alpha} \right) \vee 1 \right] \geq 1 - \alpha, \right)
\]

and consequently the result in Theorem 3.2 follows.

\section{Proof of Lemmas 3.1 to 3.4}

\textbf{Lemma 3.1.} For all \( |\Sigma_{jk}| \in 1 - \delta \) and all \( j \in \{1, \ldots, d_1\}, k \in d_1 + 1, \ldots, d_2 \) 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 \( \mu \) is the counting measure on \( S \), the corresponding discrete space.

\textbf{Proof.} Clearly, from Eq. (5) we have

\[
\frac{\partial}{\partial \Sigma_{jk}} p(x_r^j \mid x_k, \Sigma_{jk}) = (1 - (\Sigma_{jk})^2)^{-\frac{n}{2}} \left[ \phi(\tilde{\gamma}_r^j)(\gamma_r^j \Sigma_{jk} - \tilde{x}_k) - \phi(\tilde{\gamma}_r^{j-1})(\gamma_r^{j-1} \Sigma_{jk} - \tilde{x}_k) \right],
\]
and therefore
\[
\int_S (1 - (\Sigma_{jk})^2)^{-\frac{3}{2}} \left[ \phi(\tilde{\gamma}_j^r) (\gamma_j^r \Sigma_{jk} - \tilde{x}_k) - \phi(\tilde{\gamma}_{r-1}^j) (\gamma_{r-1}^j \Sigma_{jk} - \tilde{x}_k) \right] d\mu(x_j^r) = 0
\]

so that
\[
(1 - (\Sigma_{jk})^2)^{-\frac{3}{2}} \sum_{r=1}^{l_{x_j}} \left[ \phi(\tilde{\gamma}_j^r) (\gamma_j^r \Sigma_{jk} - \tilde{x}_k) - \phi(\tilde{\gamma}_{r-1}^j) (\gamma_{r-1}^j \Sigma_{jk} - \tilde{x}_k) \right] = 0
\]

since all terms except the ones involving \(\phi(\tilde{\gamma}_j^r)\) and \(\phi(\tilde{\gamma}_{l_{x_j}}^j)\) 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 3.2.** For all \(|\Sigma_{jk}| \leq 1 - \delta\) and all \(j = 1, \ldots, d_1, k = d_1 + 1, \ldots, d_2\) we have
\[
\int_S \frac{\partial^2}{\partial \Sigma_{jk}} p(x_j^r | x_k, \Sigma_{jk}) d\mu(x_j^r) = \frac{\partial^2}{\partial \Sigma_{jk}} \int_S p(x_j^r | x_k, \Sigma_{jk}) d\mu(x_j^r),
\]
where again \(\mu\) is the counting measure on \(S\), the corresponding discrete space.

**Proof.** From Eq. (13) we obtain
\[
\frac{\partial^2}{\partial \Sigma_{jk}} p(x_j^r | x_k, \Sigma_{jk}) = \frac{3 \Sigma_{jk}}{1 - \Sigma_{jk}^2} (1 - (\Sigma_{jk})^2)^{-\frac{3}{2}} \phi(\tilde{\gamma}_j^r) (\gamma_j^r \Sigma_{jk} - \tilde{x}_k) + \frac{\phi'(\tilde{\gamma}_j^r) (\gamma_j^r \Sigma_{jk} - \tilde{x}_k)^2}{(1 - \Sigma_{jk})^3} + \frac{\phi(\tilde{\gamma}_{r-1}^j) (\gamma_{r-1}^j \Sigma_{jk} - \tilde{x}_k)}{(1 - \Sigma_{jk})^2} - \frac{3 \Sigma_{jk}}{1 - \Sigma_{jk}^2} (1 - (\Sigma_{jk})^2)^{-\frac{3}{2}} \phi(\tilde{\gamma}_{r-1}^j) (\gamma_{r-1}^j \Sigma_{jk} - \tilde{x}_k)
\]

By similar arguments to Lemma 3.1, when taking the sum over all possible states all terms except the ones involving \(\phi(\tilde{\gamma}_j^r)\) and \(\phi(\tilde{\gamma}_{l_{x_j}}^j)\) 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 3.3.** For all \(|\Sigma_{jk}| \leq 1 - \delta\) we have

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

i.e.
\[
\frac{\partial}{\partial \Sigma_{jk}} \int_{S \times \mathbb{R}} \log L(\Sigma_{jk}, x_j^r, x_k) L(\Sigma_{jk}^*, x_j^r, x_k) d\varepsilon(x_j^r, x_k) = \int_{S \times \mathbb{R}} \frac{\partial}{\partial \Sigma_{jk}} \log L(\Sigma_{jk}, x_j^r, x_k) L(\Sigma_{jk}^*, x_j^r, x_k) d\varepsilon(x_j^r, x_k)
\]

where \(\varepsilon\) is the product measure on \(S \times \mathbb{R}\) defined by
\[
\varepsilon := \mu \otimes \lambda
\]

with \(\mu\) denoting the counting measure on the corresponding discrete space \(S\) and \(\lambda\) the Lebesgue measure on the corresponding Euclidean space.
2. 
\[ \frac{\partial}{\partial \Sigma_{jk}} \mathbb{E}_{\Sigma_{jk}^*} [\ell(\Sigma_{jk}, x^j, x^k)] = \mathbb{E}_{\Sigma_{jk}} \left[ \frac{\partial}{\partial \Sigma_{jk}} \ell(\Sigma_{jk}, x^j, x^k) \right], \]

i.e.
\[ \frac{\partial}{\partial \Sigma_{jk}} \int_{S \times S'} \log L(\Sigma_{jk}, x^j, x^k) L(\Sigma_{jk}^*, x^j, x^k) d\varpi(x^j, x^k) = \]
\[ \int_{S \times S'} \frac{\partial}{\partial \Sigma_{jk}} \log L(\Sigma_{jk}, x^j, x^k) L(\Sigma_{jk}^*, x^j, x^k) d\varpi(x^j, x^k), \]

where \( \varpi \) is the product measure on \( S \times S' \) defined by
\[ \varpi := \mu \otimes \mu', \]

with \( \mu \) and \( \mu' \) 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:
\[ \int_{S \times \mathbb{R}} \frac{\partial}{\partial \Sigma_{jk}} \log L(\Sigma_{jk}, x^j, x_k) L(\Sigma_{jk}^*, x^j, x_k) d\varepsilon(x^j, x_k) \]
\[ = \int_{\mathbb{R}} \sum_{r=1}^{l_{x_j}} \frac{\partial}{\partial \Sigma_{jk}} \log p(x^j_r, x_k, \Sigma_{jk}) p(x^j_r, x_k, \Sigma_{jk}^*) dx_k. \]

The left hand side corresponds to
\[ \frac{\partial}{\partial \Sigma_{jk}} \int_{S \times \mathbb{R}} \log L(\Sigma_{jk}, x^j, x_k) L(\Sigma_{jk}^*, x^j, x_k) d\varepsilon(x^j, x_k) \]
\[ = \frac{\partial}{\partial \Sigma_{jk}} \int_{\mathbb{R}} \sum_{r=1}^{l_{x_j}} \log p(x^j_r, x_k, \Sigma_{jk}) p(x^j_r, x_k, \Sigma_{jk}^*) dx_k. \]

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, a well known consequence of Lebesgue’s Dominated Convergence Theorem. Indeed the latter requirement has already been shown in Condition 2.1. Now case III, that is 2. follows by the same arguments as in Lemma 3.3 and the bound on the second derivative of the log likelihood functions in Condition 2.2, respectively.

**Corollary 3.4.** For all \( |\Sigma_{jk}| \leq 1 - \delta \) we have
\[ \frac{\partial^2}{\partial \Sigma_{jk}^2} \mathbb{E}_{\Sigma_{jk}^*} [\ell(\Sigma_{jk}, x^j, x_k)] = \mathbb{E}_{\Sigma_{jk}} \left[ \frac{\partial^2}{\partial \Sigma_{jk}^2} \ell(\Sigma_{jk}, x^j, x_k) \right], \text{ for case II and} \]
\[ \frac{\partial^2}{\partial \Sigma_{jk}^2} \mathbb{E}_{\Sigma_{jk}^*} [\ell(\Sigma_{jk}, x^j, x^k)] = \mathbb{E}_{\Sigma_{jk}} \left[ \frac{\partial^2}{\partial \Sigma_{jk}^2} \ell(\Sigma_{jk}, x^j, x^k) \right], \text{ for case III.} \]

**Proof.** This follows immediately by the same arguments as in Lemma 3.3 and the bound on the second derivative of the log likelihood functions in Condition 2.2, respectively.
4 Proof of Lemma 3.1 of the Manuscript

First, note that Φ⁻¹(·) is Lipschitz on [Φ(−2G), Φ(2G)] with a Lipschitz constant \( L_1 \) such that under the event \( A^1_j = \{ |γ^j_r| \leq 2G \} \)

\[
|γ^j_r - γ^j_{r'}| = |Φ⁻¹ \left( \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_{ji} \leq x^j_r) \right) - Φ⁻¹ \left( \Phi(γ^j_r) \right)|
\]

\[
\leq L_1 \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_{ji} \leq x^j_r) - Φ(γ^j_r) \right|
\]

We may then bound the probability of the complementary event as

\[
P \left( A^{c^2} \right) = P \left( |\hat{g}^j_r - g^j_r| > 2G \right)
\]

\[
= P \left( |\hat{g}^j_r - g^j_r| > 2G - |γ^j_r| \right)
\]

\[
\leq P \left( |\hat{g}^j_r - g^j_r| > G \right)
\]

\[
\leq P \left( \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_{ji} \leq x^j_r) - Φ(γ^j_r) \right| > \frac{G}{L_1} \right)
\]

\[
\leq 2 \exp \left( - \frac{2G^2 n}{L_1^2} \right),
\]

where the last step follows from Hoeffding’s inequality. Further define the event \( A^2 = \bigcap_{r=1}^{l_{X_j}} A^1_r \) and observe that

\[
P \left( A^{c^2} \right) = P \left( \bigcup_{r=1}^{l_{X_j}-1} A^{c^2}_r \right) \leq \sum_{r=1}^{l_{X_j}-1} P(A^{c^2}_r) \leq 2(l_{X_j} - 1) \exp \left( - \frac{2G^2 n}{L_1^2} \right),
\]

as desired.

5 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 \( \hat{f}(x) = Φ⁻¹ \left( W_{\delta_n}(\hat{F}_{X_k}(x)) \right) \) where \( W_{\delta_n}(u) = \delta_n I(u < \delta_n) + uI(\delta_n \leq u \leq (1 - \delta_n)) + (1 - \delta_n)I(u > (1 - \delta_n)) \) with the truncation constant \( \delta_n = 1/(4n^{1/4}π \log n) \). Further, \( f(x) = Φ⁻¹(F_{X_k}(x)) \), and let \( q = f⁻¹ \).

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

\[
\hat{σ}_{j}^{(n)} = \frac{r_{f(X_k),X_j}^{(n)} \sigma_{X_j}^{(n)}}{\sum_{r=1}^{l_{X_j}-1} \phi(γ^j_r)(x_{r+1} - x^j_r)} = \frac{r_{f(X_k),X_j}^{(n)} \sigma_{X_j}^{(n)}}{\sum_{r=1}^{l_{X_j}-1} \phi(γ^j_r)},
\]

for all \( 1 < j < d_1 + 1 \leq k \leq d_2 \). \( S_{f(X_k),X_j} \) denotes the sample covariance between the \( \hat{f}(X_k) \) and the \( X_j \), i.e.

\[
S_{f(X_k),X_j} = \frac{1}{n} \sum_{i=1}^{n} \left( \hat{f}(X_{ki}) - \mu_n(\hat{f}) \right) \left( X_{ji} - \mu_n(X_j) \right),
\]

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

\[
\sigma^{(n)}_{f(X_k)} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left( \hat{f}(X_{ki}) - \mu_n(\hat{f}) \right)^2}.
\]
Recall that we treat the thresholds estimates as given. In particular, we have here \(\phi(\gamma^j_1)\), therefore note further that \(\phi(\cdot)\), namely the density function of the standard normal is Lipschitz with Lipschitz constant \(L_0 = (2\pi)^{-1/2}e^{-1/2}\), s.t.

\[
|\phi(\gamma^j_2) - \phi(\gamma^j_1)| \leq L_0 |\gamma^j_2 - \gamma^j_1| = |\gamma^j_1 - \gamma^j_1|,
\]

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

The outline of the proof will be as follows: We start by forming concentration bounds for both 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.

\[
M_n ≡ (g(-\sqrt{\beta\log n}), g(\sqrt{\beta\log n}))
\]

\[
E_n ≡ [g(-\sqrt{\log n}), g(-\sqrt{\beta\log n})) \cup (g(\sqrt{\beta\log n}), g(\sqrt{M\log n})].
\]

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

\[
P\left(\max_{j,k} |S_{f(X_k)X_j} - S_{f(X_k)X_j}| > 2t\right)
\]

\[
= P\left(\max_{j,k} \frac{1}{n} \sum_{i=1}^{n} \left|\hat{f}(X_{ki})X_{ji} - f(X_{ki})X_{ji} \right|
\]

\[
= P\left(\max_{j,k} \left|\frac{1}{n} \sum_{i=1}^{n} \left(\hat{f}(X_{ki}) - f(X_{ki})\right)X_{ji} \right| > t\right)
\]

\[
+ P\left(\max_{j,k} \left(\mu_n(\hat{f}) - \mu_n(f)\right)\mu_n(X_j) > t\right).
\]

Let us take a closer look at the second term

\[
P\left(\max_{j,k} \left|\mu_n(\hat{f}) - \mu_n(f)\right|\mu_n(X_j) > t\right)
\]

\[
= P\left(\max_{j,k} \left|\frac{1}{n} \sum_{i=1}^{n} \left(\hat{f}(X_{ki}) - f(X_{ki})\right)X_{ji} \right| > t\right)
\]

\[
= P\left(\max_{j,k} \left|\frac{1}{n} \sum_{i=1}^{n} \left(\hat{f}(X_{ki}) - f(X_{ki})\right)X_{ji} \right| > t\right)
\]

\[
< P\left(\max_{j,k} \left|\frac{1}{n} \sum_{i=1}^{n} \left(\hat{f}(X_{ki}) - f(X_{ki})\right)X_{ji} \right| > \frac{t}{l_{\text{max}}}\right),
\]

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

Now, define

\[
\triangle_i(j,k) \equiv (\hat{f}(X_{ki}) - f(X_{ki}))X_{ji}
\]

and

\[
\hat{\triangle}_{r,s} \equiv (\hat{f}(s) - f(s))r,
\]

for \(r = 1, \ldots, l_{X_j}\). Furthermore, consider the event \(\mathcal{A}_n\), where

\[
\mathcal{A}_n \equiv \{g(-\sqrt{M\log n}) \leq X_{k1}, \ldots, X_{kn} \leq g(\sqrt{M\log n}), k = d_1 + 1, \ldots, d\}.
\]
The bound for the complement arises from the Gaussian maximal inequality \cite[Lemma 13]{4}, i.e.,
\[
P(A_n^c) \leq P \left( \max_{i,k} \frac{1}{n} \sum_{X_k \in E_n} |\hat{f}(X_{ki}) - f(X_{ki})| > t \right) \leq \frac{1}{2} \sqrt{\frac{2}{\pi} \log(n d_2)}.
\]

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

**Lemma 5.1.** On the event $A_n$, consider $\beta = \frac{1}{2}$, $t \geq C_M \sqrt{\frac{\log d_2 \log^2 n}{n^2 \pi}}$ and $A = \sqrt{\frac{2}{\pi} (\sqrt{M} - \sqrt{\beta})}$, then
\[
P \left( \max_{j,k} \frac{1}{n} \sum_{X_k \in E_n} |\hat{f}(X_{ki}) - f(X_{ki})| X_{ji} > \frac{t}{2} \right) \leq \exp \left( - \frac{k_1 n^{3/4} \sqrt{\log n}}{k_2 + k_3} \right),
\]
and
\[
P \left( \max_{j,k} \frac{1}{n} \sum_{X_k \in E_n} |\hat{f}(X_{ki}) - f(X_{ki})| > \frac{t}{2} \right) \leq \exp \left( - \frac{k_1 n^{3/4} \sqrt{\log n}}{k_2 + k_3} \right),
\]
where $k_i, i \in \{1, 2, 3\}$ are generic constants independent of sample size and dimension.

**Proof.** Let $\theta_1 = \frac{n^{3/4} \sqrt{\log n}}{4 A \sqrt{n} \sqrt{\pi}}$ and let us first consider the bound for the first inequality.
\[
P \left( \max_{j,k} \frac{1}{n} \sum_{X_k \in E_n} |\hat{f}(X_{ki}) - f(X_{ki})| X_{ji} > \frac{t}{2} \right)
\]
\[
= P \left( \max_{j,k} \frac{1}{n} \sum_{i : X_k \in E_n} |\hat{f}(X_{ki}) - f(X_{ki})| X_{ji} > \frac{t}{2} \right.
\]
\[
\cap \max_{j,k} \sup_{r \in \{1, \ldots, l_{x_j}\}} \left. \sup_{s \in E_n} \left| \hat{f}(t) - f(t) \right| \right| r > \theta_1 \right)
\]
\[
+ P \left( \max_{j,k} \frac{1}{n} \sum_{i : X_k \in E_n} |\hat{f}(X_{ki}) - f(X_{ki})| X_{ji} > \frac{t}{2} \right)
\]
\[
\cap \max_{j,k} \sup_{r \in \{1, \ldots, l_{x_j}\}} \left. \sup_{s \in E_n} \left| \hat{f}(s) - f(s) \right| \right| r \leq \theta_1 \right)
\]
\[
\leq P \left( \max_{j,k} \sup_{r \in \{1, \ldots, l_{x_j}\}, s \in E_n} \left| \hat{f}(s) - f(s) \right| \right| r > \theta_1 \right) + P \left( \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i \in E_n\}} > \frac{t}{2 \theta_1} \right).
\]

Similarly, for the bound of the second inequality we have
\[
P \left( \max_{k} \frac{1}{n} \sum_{i : X_k \in E_n} |\hat{f}(X_{ki}) - f(X_{ki})| > \frac{t}{2} \right)
\]
\[
= P \left( \max_{k} \frac{1}{n} \sum_{i : X_k \in E_n} |\hat{f}(X_{ki}) - f(X_{ki})| > \frac{t}{2} \cap \max_{k} \sup_{s \in E_n} \left| \hat{f}(s) - f(s) \right| > \theta_1 \right)
\]
\[
+ P \left( \max_{k} \frac{1}{n} \sum_{i : X_k \in E_n} |\hat{f}(X_{ki}) - f(X_{ki})| > \frac{t}{2} \cap \max_{k} \sup_{s \in E_n} \left| \hat{f}(s) - f(s) \right| \leq \theta_1 \right)
\]
\[
\leq P \left( \max_{k} \sup_{s \in E_n} \left| \hat{f}(s) - f(s) \right| \right| > \theta_1 \right) + P \left( \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i \in E_n\}} > \frac{t}{2 \theta_1} \right).
\]

Recall that $\sup\{1, \ldots, l_{x_j}\} = l_{x_j} > 0$. Furthermore, Lemma 16 in Liu et al. \cite{4} states that for all $n$
\[
\sup_{s \in E_n} \left| \hat{f}(s) - f(s) \right| < \sqrt{2(M + 2) \log n}
\]
(19)

With this in mind, we have
\[
P \left( \max_{k} \sup_{s \in E_n} \left| \hat{f}(s) - f(s) \right| > \theta_1 \right) \leq d_2 P \left( \sup_{s \in E_n} \left| \hat{f}(s) - f(s) \right| > \theta_1 \right).
\]
Recall that $C_M = \frac{8}{\sqrt{\pi}}(\sqrt{2M - 1}(M + 2)$ and since $t \geq C_M \sqrt{\frac{\log d_2 \log^2 n}{n^2}}$, we have

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

Consequently, we have

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

as well as

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

such that

$$P\left(\sup_{t \in \mathbb{E}_n} |f(t) - f(t)| > \theta_1\right) = P\left(\sup_{r \in \{1, \ldots, l_{x_j}\}, s \in \mathbb{E}_n} |f(s) - f(s)| r > \theta_1\right) = 0.$$ 

Now let us turn to the second term which is equivalent in both cases. We have

$$P\left(\frac{1}{n} \sum_{i=1}^{n} 1_{\{X_{ki} \in \mathbb{E}_n\}} > \frac{t}{2 \theta_1}\right) = P\left(\sum_{i=1}^{n} 1_{\{X_{ki} \in \mathbb{E}_n\}} > \frac{nt}{2 \theta_1}\right)$$

$$= P\left(\sum_{i=1}^{n} (1_{\{X_{ki} \in \mathbb{E}_n\}} - P(X_{k1} \in \mathbb{E}_n)) > \frac{nt}{2 \theta_1} - P(X_{k1} \in n\mathbb{E}_n)\right)$$

$$\leq P\left(\sum_{i=1}^{n} (1_{\{X_{ki} \in \mathbb{E}_n\}} - P(X_{k1} \in \mathbb{E}_n)) > \frac{nt}{2 \theta_1} - nA \sqrt{\frac{\log n}{n^2}}\right).$$

Choosing $\theta_1$ this way guarantees that

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

Then, using Bernstein’s inequality we get

$$P\left(\frac{1}{n} \sum_{i=1}^{n} 1_{\{X_{ki} \in \mathbb{E}_n\}} > \frac{t}{2 \theta_1}\right)$$

$$\leq P\left(\sum_{i=1}^{n} (1_{\{X_{ki} \in \mathbb{E}_n\}} - P(X_{k1} \in \mathbb{E}_n)) > nA \sqrt{\frac{\log n}{n^2}}\right)$$

$$\leq \exp\left(-\frac{k_1 n^{2-\beta} \log n}{k_2 n^{1-\beta/2} \sqrt{\log n} + k_3 n^{1-\beta/2} \sqrt{\log n}}\right),$$

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

$$P\left(\max_{j,k} \left|\frac{1}{n} \sum_{i=1}^{n} (\hat{f}(X_{ki}) - f(X_{ki})) X_{ji}\right| > t\right)$$

$$\leq P\left(\max_{j,k} \left|\frac{1}{n} \sum_{i=1}^{n} \Delta_i(j,k)\right| > t, A_n\right) + P(A_n^c)$$

$$\leq P\left(\max_{j,k} \left|\frac{1}{n} \sum_{i=1}^{n} \Delta_i(j,k)\right| > t \cap A_n\right) + \frac{1}{2 \sqrt{\pi \log(nd_2)}}.$$
Further, we have
\[
P\left(\max_{j,k} \left| \frac{1}{n} \sum_{i=1}^{n} \triangle_i(j,k) \right| > t \cap A_n \right)
\]
\[
\leq P\left(\max_{j,k} \frac{1}{n} \sum_{X_k \in M_n} \left| \triangle_i(j,k) \right| > \frac{t}{2}\right) + P\left(\max_{j,k} \frac{1}{n} \sum_{X_k \in E_n} \left| \triangle_i(j,k) \right| > \frac{t}{2}\right)
\]
\[
+ \frac{1}{2\sqrt{\pi} \log(nd_2)}
\]
\[
\leq P\left(\max_{j,k} \frac{1}{n} \sum_{X_k \in M_n} \left| \triangle_i(j,k) \right| > \frac{t}{2}\right) + \exp\left(\frac{-k_1n^{3/4}\sqrt{\log n}}{k_2+k_3}\right)
\]
\[
+ \frac{1}{2\sqrt{\pi} \log(nd_2)}
\]
where \(X_k \in M_n\) is shorthand notation for \(i : X_{ki} \in M_n\). The bound of the second term is derived in Lemma 5.1. Let us continue with the first term
\[
P\left(\max_{j,k} \frac{1}{n} \sum_{X_k \in M_n} \left| \triangle_i(j,k) \right| > \frac{t}{2}\right)
\]
\[
\leq d^2P\left(\sup_{r \in \{1, \ldots, l_{X_j}\}, s \in M_n} \left| \tilde{\Delta}_{r,s} \right| > \frac{t}{2}\right)
\]
\[
= d^2P\left(\sup_{r \in \{1, \ldots, l_{X_j}\}, s \in M_n} \left| (\hat{f}(s) - f(s)) \right| > \frac{t}{2}\right)
\]
\[
= d^2P\left(\sup_{s \in M_n} \left| (\hat{f}(s) - f(s)) \right| > \frac{t}{2l_{X_j}}\right),
\]
where clearly \(\sup(\{1, \ldots, l_{X_j}\}) = l_{X_j} > 0\). Define the event
\[
\mathbb{B}_n \equiv \{\delta_n \leq \hat{F}_{X_k}(g_j(s)) \leq 1 - \delta_n, \quad k = d_1 + 1, \ldots, d\}.
\]
Now, from the definition of the Winsorized estimator, we observe that
\[
d^2P\left(\sup_{s \in M_n} \left| (\hat{f}(s) - f(s)) \right| > \frac{t}{2l_{X_j}}\right)
\]
\[
\leq d^2P\left(\sup_{s \in M_n} \left| \Phi^{-1}(W_{\delta_n}[\hat{F}_{X_k}(s)]) - \Phi^{-1}(F_{X_k}(s)) \right| > \frac{t}{2l_{X_j} \cap \mathbb{B}_n}\right) + P(\mathbb{B}_n^c)
\]
\[
\leq d^2P\left(\sup_{s \in M_n} \left| \Phi^{-1}(\hat{F}_{X_k}(s)) - \Phi^{-1}(F_{X_k}(s)) \right| > \frac{t}{2l_{X_j}}\right)
\]
\[
+ 2 \exp\left(2 \log d - \frac{\sqrt{n}}{8\pi \log n}\right),
\]
where the expression for \(P(\mathbb{B}_n^c)\) follows directly from Lemma 19 in Liu et al. [4]. Now, define
\[
T_{1n} \equiv \max \left\{ F_{X_k}(g(\sqrt{\beta \log n})), 1 - \delta_n \right\}
\]
\[
T_{2n} \equiv 1 - \min \left\{ F_{X_k}(g(-\sqrt{\beta \log n})), \delta_n \right\},
\]
where it follows directly that $T_{1n} = T_{2n} = 1 - \delta_n$. Consequently, we apply the mean value theorem and get

$$P\left( \sup_{s \in S_n} \left| \Phi^{-1}(\hat{F}_{X_k}(s)) - \Phi^{-1}(F_{X_k}(s)) \right| > \frac{t}{2l_{X_j}} \right)$$

$$\leq P\left( (\Phi^{-1})'(1 - \delta_n) \sup_{s \in S_n} |\hat{F}_{X_k}(s) - F_{X_k}(s)| > \frac{t}{2l_{X_j}} \right)$$

$$= P\left( \frac{1}{\phi(\Phi^{-1}(1 - \delta_n))} \frac{1}{\sqrt{2 \log \frac{n}{\delta_n}}} = \sqrt{2 \pi} \left( \frac{1}{\delta_n} \right) = 8\pi n^{1/2} \sqrt{\beta \log n} \right)$$

Therefore,

$$d^2 P\left( \sup_{s \in S_n} \left| \Phi^{-1}(\hat{F}_{X_k}(s)) - \Phi^{-1}(F_{X_k}(s)) \right| > \frac{t}{2l_{X_j}} \right)$$

$$\leq 2 \exp \left( - \frac{\sqrt{m^2}}{4l_{X_j}^2 \pi^2 \log n} \right).$$

Collecting the remaining terms we have

$$P\left( \max_{j,k} \frac{1}{n} \sum_{X_k \in S_n} \left| \Delta_i(j,k) \right| > \frac{t}{2} \right) \leq 2 \exp \left( - \frac{\sqrt{m^2}}{64l_{X_j}^2 \pi^2 \log n} \right)$$

$$+ 2 \exp \left( 2 \log d - \frac{\sqrt{n}}{8\pi \log n} \right).$$

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

$$P\left( \max_{j,k} \frac{1}{n} \sum_{i=1}^{\frac{n}{X_j}} \left| \hat{f}(X_{ki}) - f(X_{ki}) \right| > \frac{t}{2} \right)$$

$$\leq P\left( \max_{j,k} \frac{1}{n} \sum_{X_k \in S_n} \left| \Delta_i(j,k) \right| > \frac{t}{2} \right)$$

$$+ \exp \left( - \frac{k_1 n^{3/4} \sqrt{\log n}}{k_2 + k_3} \right) + \frac{1}{2\sqrt{\pi} \log(2d)}$$

$$\leq 2 \exp \left( 2 \log d - \frac{\sqrt{m^2}}{64l_{X_j}^2 \pi^2 \log n} \right) + 2 \exp \left( 2 \log d - \frac{\sqrt{n}}{8\pi \log n} \right)$$

$$+ \exp \left( - \frac{k_1 n^{3/4} \sqrt{\log n}}{k_2 + k_3} \right) + \frac{1}{2\sqrt{\pi} \log(2d)}.$$
Let us turn back to the second term of the first sample covariance decomposition, i.e.

\[
P \left( \max_{j,k} \left| \mu_n(\hat{f}) - \mu_n(f) \right| \mu_n(X_j) > t \right) 
\leq P \left( \max_k \left| \frac{1}{n} \sum_{i=1}^{n} (\hat{f}(X_{ki}) - f(X_{ki})) \right| > \frac{t}{l_{\text{max}} \cap A_n} \right) + \frac{1}{2\sqrt{\pi \log(nd_2)}}.
\]

Now, analogous to before we find

\[
P \left( \max_k \frac{1}{n} \sum_{i=1}^{n} |\hat{f}(X_{ki}) - f(X_{ki})| > \frac{t}{l_{\text{max}} \cap A_n} \right) 
\leq d_2 P \left( \sup_{t \in M_n} |\hat{f}(t) - f(t)| > \frac{t}{2l_{\text{max}}} \right) + \exp \left( - \frac{k_1 n^{3/4} \sqrt{\log n}}{k_2 + k_3} \right).
\]

Let us take a closer look at

\[
d_2 P \left( \sup_{t \in M_n} |\hat{f}(t) - f(t)| > \frac{t}{2l_{\text{max}}} \right) 
\leq d_2 P \left( \sup_{t \in M_n} \left| \Phi^{-1}(W_{\delta_n} [\hat{F}_{X_k}(t)]) - \Phi^{-1}(F_{X_k}(t)) \right| > \frac{t}{2l_{\text{max}} \cap B_n} \right) + d_2 P(\overline{B}_n)
\]

\[
\leq d_2 P \left( \sup_{t \in M_n} \left| \Phi^{-1}(\hat{F}_{X_k}(t)) - \Phi^{-1}(F_{X_k}(t)) \right| > \frac{t}{2l_{\text{max}}} \right) + 2 \exp \left( - \frac{k_1 n^{3/4} \sqrt{\log n}}{k_2 + k_3} \right).
\]

The definition of the event \( B_n \) is the same as above. Then applying once more the Dvoretzky–Kiefer–Wolfowitz inequality we end up with the following upper bound:

\[
d_2 P \left( \sup_{t \in M_n} \left| \Phi^{-1}(\hat{F}_{X_k}(t)) - \Phi^{-1}(F_{X_k}(t)) \right| > \frac{t}{2l_{\text{max}}} \right) 
\leq 2 \exp \left( - \frac{n^{3/4} \sqrt{\log n}}{8\pi \log n} \right).
\]

18
Collecting terms and simplifying yields

\[ P \left( \max_{j,k} \left| \frac{1}{n} \sum_{i=1}^{n} \left[ (\hat{f}(X_{ki}) - f(X_{ki}))X_{ji} \right] \right| > t \right) \]

\[ + P \left( \max_{j,k} \left| (\mu_n(\hat{f}) - \mu_n(f))\mu_n(X_j) \right| > t \right) \]

\[ \leq 2 \exp \left( 2 \log d - \frac{\sqrt{n}t^2}{64 l^2_{\max} \pi^2 \log n} \right) + 2 \exp \left( \log d - \frac{\sqrt{n}}{8\pi \log n} \right) \]

\[ + \exp \left( - \frac{k_1 n^{3/4} \sqrt{\log n}}{k_2 + k_3} \right) + \frac{1}{2\sqrt{\pi} \log(nd_2)} \]

\[ + 2 \exp \left( \log d - \frac{\sqrt{n}t^2}{64 l^2_{\max} \pi^2 \log n} \right) + 2 \exp \left( \log d - \frac{\sqrt{n}}{8\pi \log n} \right) \]

\[ + \exp \left( - \frac{k_1 n^{3/4} \sqrt{\log n}}{k_2 + k_3} \right) + \frac{1}{2\sqrt{\pi} \log(nd_2)} \]

\[ \leq 4 \exp \left( 2 \log d - \frac{\sqrt{n}t^2}{64 l^2_{\max} \pi^2 \log n} \right) + 4 \exp \left( 2 \log d - \frac{\sqrt{n}}{8\pi \log n} \right) \]

\[ + 2 \exp \left( - \frac{k_1 n^{3/4} \sqrt{\log n}}{k_2 + k_3} \right) + \frac{1}{\sqrt{\pi} \log(nd_2)}. \]

Then

\[ P \left( \max_{j,k} \left| S_{f(X_k)X_j} - S_{f(X_k)X_j} \right| > 2t \right) \]

\[ \leq 4 \exp \left( 2 \log d - \frac{\sqrt{n}t^2}{64 l^2_{\max} \pi^2 \log n} \right) + 4 \exp \left( 2 \log d - \frac{\sqrt{n}}{8\pi \log n} \right) \]

\[ + 2 \exp \left( - \frac{k_1 n^{3/4} \sqrt{\log n}}{k_2 + k_3} \right) + \frac{1}{\sqrt{\pi} \log(nd_2)}. \]

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

\[ \sigma_{f(X_k)}^{(n)} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left( \hat{f}(X_{ki}) - \mu_n(\hat{f}) \right)^2}. \]
Consider the following decomposition of the standard deviation of the Winsorized estimator,

\[
\left| \sigma_{f(X_k)}^{(n)} - \sigma_{f(X_k)}^{(n)} \right| = \left| \frac{1}{1/n} \sum_{i=1}^{n} \left( \hat{f}(X_{ki}) - \mu_n(\hat{f}) \right)^2 - \frac{1}{1/n} \sum_{i=1}^{n} \left( f(X_{ki}) - \mu_n(f) \right)^2 \right|
\]

\[
\leq \frac{1}{\sqrt{n}} \left| \left( \| \hat{f}(X_k) - \mu_n(\hat{f}) \|_2 - \| f(X_k) - \mu_n(f) \|_2 \right) \right|
\]

\[
\leq \frac{1}{\sqrt{n}} \| \hat{f}(X_k) - \mu_n(\hat{f}) - f(X_k) + \mu_n(f) \|_2
\]

\[
\leq \frac{1}{\sqrt{n}} \| \hat{f}(X_k) - \mu_n(\hat{f}) - f(X_k) + \mu_n(f) \|_\infty
\]

\[
= \sup_{i : X_{ki} \in \{1, \ldots, n\}} \left| \hat{f}(X_{ki}) - f(X_{ki}) + \mu_n(f) - \mu_n(\hat{f}) \right|
\]

\[
= \sup_{i : X_{ki} \in \{1, \ldots, n\}} \left| \hat{f}(X_{ki}) - f(X_{ki}) + \mu_n(f) - \mu_n(\hat{f}) \right|
\]

\[
\leq \sup_{i : X_{ki} \in \{1, \ldots, n\}} \left| \hat{f}(X_{ki}) - f(X_{ki}) + \frac{1}{n} \sum_{i=1}^{n} \left| \hat{f}(X_{ki}) - f(X_{ki}) \right| \right|
\]

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

\[
P\left( \max_k \left| \sigma_{f(X_k)}^{(n)} - \sigma_{f(X_k)}^{(n)} \right| > 2t \right)
\]

\[
\leq P\left( \max_k \sup_{i : X_{ki} \in \{1, \ldots, n\}} \left| \hat{f}(X_{ki}) - f(X_{ki}) \right| > t, A_n \right)
\]

\[
+ P\left( \max_k \frac{1}{n} \sum_{i=1}^{n} \left| \hat{f}(X_{ki}) - f(X_{ki}) \right| > t, A_n \right) + P(A_n^c).
\]

Note, that the second term is in effect equivalent to Eq. (20) above such that we can immediately conclude that

\[
P\left( \max_k \frac{1}{n} \sum_{i=1}^{n} \left| \hat{f}(X_{ki}) - f(X_{ki}) \right| > t \cap A_n \right)
\]

\[
\leq d_2 P\left( \sup_{t \in \mathbb{R}^n} \left| \hat{f}(t) - f(t) \right| > \frac{t}{2} \right) + \exp\left( - \frac{k_1 n^{3/4} \sqrt{\log n}}{k_2 + k_3} \right)
\]

\[
\leq 2 \exp\left( \log d_2 - \frac{\sqrt{n}}{64 \pi d_2 \log n} \right) + 2 \exp\left( \log d_2 - \frac{\sqrt{n}}{8 \pi \log n} \right)
\]

\[
+ \exp\left( - \frac{k_1 n^{3/4} \sqrt{\log n}}{k_2 + k_3} \right).
\]

The bound for the end region follows again from Lemma 5.1.

Similarly, we find that

\[
P\left( \max_k \sup_{i : X_{ki} \in \{1, \ldots, n\}} \left| \hat{f}(X_{ki}) - f(X_{ki}) \right| > t \cap A_n \right)
\]

\[
\leq d_2 P\left( \sup_{t \in \mathbb{R}^n} \left| \hat{f}(t) - f(t) \right| > \frac{t}{2} \right) + d_2 P\left( \sup_{t \in \mathbb{R}^n} \left| \hat{f}(t) - f(t) \right| > \frac{t}{2} \right)
\]

\[
= d_2 P\left( \sup_{t \in \mathbb{R}^n} \left| \hat{f}(t) - f(t) \right| > \frac{t}{2} \right),
\]
where the bound over the end region has been shown in Lemma 5.1. Thus, we only have to take care of
\[
P\left( \sup_{t \in \mathbb{M}_n} \left| \tilde{f}(t) - f(t) \right| > \frac{t}{2} \right)
\]
\[
\leq P\left( \sup_{t \in \mathbb{M}_n} \left| \Phi^{-1}(W_{\delta_n}[\tilde{F}_{X_k}(t)]) - \Phi^{-1}(F_{X_k}(t)) \right| > \frac{t}{2} \cap B_n \right) + P(B_n^c)
\]
\[
\leq P\left( \sup_{t \in \mathbb{M}_n} \left| \Phi^{-1}(\tilde{F}_{X_k}(t)) - \Phi^{-1}(F_{X_k}(t)) \right| > \frac{t}{2} \right) + 2 \exp \left( \log d_2 - \frac{\sqrt{n}}{8\pi \log n} \right).
\]

The definition of the event $B_n$ is the same as above. Then again, by the Dvoretzky–Kiefer–Wolfowitz inequality we end up with the following upper bound:
\[
P\left( \sup_{t \in \mathbb{M}_n} \left| \Phi^{-1}(\tilde{F}_{X_k}(t)) - \Phi^{-1}(F_{X_k}(t)) \right| > \frac{t}{2} \right) \leq 2 \exp \left( - \frac{\sqrt{nt^2}}{64\pi^2 \log n} \right).
\]

Collecting terms, the concentration bound for the sample standard deviation is given by
\[
P\left( \max_k \left| \sigma_{f(X_k)}(n) - \sigma_{f(X_k)}^* \right| > 2t \right)
\]
\[
\leq 4 \exp \left( \log d_2 - \frac{\sqrt{nt^2}}{64\pi^2 \log n} \right) + 4 \exp \left( \log d_2 - \frac{\sqrt{n}}{8\pi \log n} \right)
\]
\[
+ \exp \left( - \frac{k_1n^{3/4} \sqrt{\log n}}{k_2 + k_3} \right) + \frac{1}{2 \sqrt{\pi \log(\text{nd}_2)}}.
\]

With this intermediate result, we have shown that both the sample covariance (numerator) as well as the sample standard deviation (denominator) can be estimated accurately.

The following lemma provides us with the means to forming a probability bound for
\[
\max_{jk} \left| \Sigma_{jk}(n) - \Sigma_{jk}^* \right|.
\]

**Lemma 5.2.** Consider the polyserial ad hoc estimator $\Sigma_{jk}^{(n)}$ for $1 \leq j \leq d_1 < k \leq d_2$ and let $\epsilon \in \left[ C_M \sqrt{\frac{\log d_2 \log^2 n}{n^{1/3}}} \right], 8(1 + 4c^2)$, where $c$ is the corresponding sub-Gaussian parameter of the discrete variable. Both the numerator and the denominator are bounded, i.e.
\[
S_{f(X_k)X_j} \in [-1 + \epsilon, 1 + \epsilon],
\]
and
\[
\sigma_{f(X_k)}^{(n)} \in [1 - \epsilon, 1 + \epsilon].
\]

Consequently, $\hat{\Sigma}_{jk}^{(n)}$ is Lipschitz with constant $L$. The following decomposition holds
\[
\max_{jk} \left| \Sigma_{jk}^{(n)} - \Sigma_{jk}^* \right| = \max_{jk} \left| \hat{\Sigma}_{jk}^{(n)} - \Sigma_{jk}^* + \Sigma_{jk}^{(n)} - \Sigma_{jk}^* \right|
\]
\[
\leq \max_{jk} \left| \Sigma_{jk}^{(n)} - \Sigma_{jk}^* \right| + \max_{jk} \left| \Sigma_{jk}^{(n)} - \Sigma_{jk}^* \right|
\]
\[
\leq L \left( \max_{jk} \left| S_{f(X_k)X_j} - S_{f(X_k)X_j}^* \right| + C_\gamma \max_k \left| \sigma_{f(X_k)}^{(n)} - \sigma_{f(X_k)}^* \right| \right)
\]
\[
+ \max_{jk} \left| S_{f(X_k)X_j} - S_{f(X_k)X_j}^* \right| + C_\gamma \max_k \left| \sigma_{f(X_k)}^{(n)} - 1 \right| \right),
\]
where $C_\gamma = \sum_{r=1}^{L \gamma_j} \phi(\gamma_j^r)(x^r_{j+1} - x^r_j)$.  

21
Proof. Let us assume w.l.o.g. that \( X_j \) – the discrete variable – has zero mean and variance one. By the Cauchy-Schwarz inequality, the true covariance of the pair is bounded from above by 1, i.e.

\[
|S_{f(X_k)X_j}^*| \leq \sigma_{f(X_k)}^2 \sigma_{X_j}^2 = 1.
\]

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

\[
P\left( \max_{j,k} \left| S_{f(X_k)X_j} - S_{f(X_k)X_j}^* \right| > 2t \right) \leq 4 \exp \left( 2 \log d_2 - \frac{\sqrt{n}t^2}{64 d_2^2 \max \pi^2 \log n} \right) + 4 \exp \left( 2 \log d_2 - \frac{\sqrt{n}}{8 \pi \log n} \right) + 2 \exp \left( - \frac{k_1 n^{3/4} \log n}{k_2 + k_3} \right) + \frac{1}{\sqrt{\pi \log(n d_2)}}.
\]

(22)

According to Lemma 1 in Ravikumar et al. [5] 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

\[
P\left( \max_{j,k} \left| S_{f(X_k)X_j} - S_{f(X_k)X_j}^* \right| \geq t \right) \leq 4d^2 \exp \left\{ - \frac{nt^2}{128(1 + 4c^2)^2} \right\},
\]

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

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

and since

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

with high probability we have

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

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

\[
P\left( \max_k \left| \sigma_{f(X_k)}^{(n)} - \sigma_{f(X_k)}^{(n)} \right| > 2t \right)
\]

\[
\leq 4 \exp \left( \log d_2 - \frac{\sqrt{n}t^2}{64 \pi^2 \log n} \right) + 4 \exp \left( \log d_2 - \frac{\sqrt{n}}{8 \pi \log n} \right)
\]

\[
+ \exp \left( - \frac{k_1 n^{3/4} \log n}{k_2 + k_3} \right) + \frac{1}{2 \sqrt{\pi \log(n d_2)}}.
\]

Furthermore, we use again Lemma 1 in Ravikumar et al. [5] to bound the variance. Since \( f(X_k) \) is standard Gaussian and hence sub-Gaussian with parameter \( c = 1 \) we immediately get

\[
P\left( \max_k \left( \sigma_{f(X_k)}^{(n)} \right)^2 - 1 \right) \geq t \right) \leq 4d_2 \exp \left\{ - \frac{nt^2}{128(1 + 4)^2} \right\}.
\]

Put differently, with high probability

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

and consequently we also have with high probability

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

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

\[
\sigma_{f(X_k)}^{(n)} \in [1 - t, 1 + t].
\]
Putting these things together, we obtain that with high probability
\[ \sigma_{f(X_k)}^{(n)} \in [1 - 3t, 1 + 3t]. \]

In order to finish the proof consider the following function \( h : \mathbb{R} \times \mathbb{R}_+ \rightarrow \mathbb{R} \) defined by
\[ h(u, v) = \frac{u}{v}, \]
with \( \nabla h = (1/v, -u/v^2)^T \).
As we have just shown in case of the polyserial ad hoc estimator, \( \nabla h = (1/v, -u/v^2)^T \) is bounded with
\[ \sup_{uv} \| \nabla h \|_2 = \sqrt{\left( \frac{1}{1 - 3t} \right)^2 + \left( \frac{1 + 3t}{1 - 3t} \right)^2 } := L. \]
Consequently \( h \) is Lipschitz and we have the following decomposition
\[ |h(u, v) - h(u', v')| = |h(u, v) - h(\bar{u}, \bar{v}) + h(\bar{u}, \bar{v}) - h(u', v')| \]
\[ \leq |h(u, v) - h(\bar{u}, \bar{v})| + |h(\bar{u}, \bar{v}) - h(u', v')| \]
\[ \leq L(|u - \bar{u}| + |v - \bar{v}|) + L(|\bar{u} - u'| + |\bar{v} - v'|). \]
Finally, taking \( \epsilon = 3t \) finishes the proof. \(
\]

At last, collecting terms, we find that for \( j \in 1, \ldots, d_1 \) and \( k \in d_1 + 1, \ldots, d \) and any \( \epsilon \in \left[ C_M \sqrt{\frac{\log d \log^2 n}{\sqrt{n}}}, 8(1+4\epsilon^2) \right] \)
the following bound holds
\[ P \left( \max_{jk} |\Sigma_{jk}^{(n)} - \Sigma_{jk}^*| \geq \epsilon \right) \]
\[ \leq P \left( \max_{jk} |S_{f(X_k)X_j} - S_{f(X_k)X_j}^*| > \frac{\epsilon}{4L} \right) + P \left( \max_k |\sigma_{f(X_k)}^{(n)} - \sigma_{f(X_k)}^*| > \frac{\epsilon}{4LC}\right) \]
\[ + P \left( \max_{jk} |S_{f(X_k)X_j} - S_{f(X_k)X_j}^*| \geq \frac{\epsilon}{4L} \right) + P \left( \max_k |(\sigma_{f(X_k)}^{(n)})^2 - 1| \geq \frac{\epsilon}{4LC}\right) \]
The conclusion of Theorem 3.3 follows by plugging in the corresponding concentration bounds and simplifying.

6 Additional simulation setup and results

In the simulations carried out we start by constructing the true latent graph \( \Omega^* \). We set \( \Omega_{ij}^* = 1 \) and \( \Omega_{jk}^* = sb_{jk} \) if \( j \neq k \), where \( s \) is the constant signal strength so as to assure positive definiteness. Furthermore, \( b_{jk} \) are realizations of a Bernoulli random variable with corresponding success probability \( p_{jk} = (2\pi)^{-1/2} \exp \left[ -\|v_j - v_k\|_2^2/(2c) \right] \). In particular \( v_j = (v_{j1}, v_{j2}) \) are independent realizations of a bivariate uniform \( [0, 1] \) distribution and \( c \) controls the sparsity of the graph. Throughout the experiments \( \hat{\Omega} \) is chosen by minimizing the eBIC according to the procedure outlined in Section 3.6 of the Manuscript with \( \theta = 0.1 \) for the low and medium, \( \theta = 0.5 \) for the high dimensional graphs.

We set \( s = 0.15 \) and incrementally increase the dimensionality of each graph: \( d = 50, 250, 750 \) representing a transition from small to large scale graphs. We let \( \Sigma^* = (\Omega^*)^{-1} \) rescaled such that all diagonal elements are equal to 1. Equipped with \( \Sigma^* \) we draw \( n \) iid. samples from \( N_d(0, \Sigma^*) \) obtaining realizations for the case of the latent Gaussian model. For the nonparanormal family, we sample from \( \text{NPN}(0, \Sigma^*, \hat{f}) \) where \( f_j(x) = x^j \) for all \( 1 \leq j \leq d \).

In order to agree with the latent setup according to Definition 2.2 let \( X_1 \) be partitioned into equally sized collections of binary, ordinal and Poisson distributed random variables i.e. \( X_1 = (X_1^{\text{bin}}, X_1^{\text{ord}}, X_1^{\text{pois}}) \) where the generative procedure is according to Eq. (1) of the main Manuscript.
Recall, that for any continuous random variable $X$ with corresponding cumulative distribution function (CDF) $F_X$, $Y := F_X(X)$ is a standard uniformly distributed random variable. Then, given $Y$ and with the aid of the inverse probability integral transform we can generate random samples from any cumulative distribution function [6]. Incidentally, this corresponds exactly to the relationship described in Eq. (1) of the main Manuscript. For $X^{bn}_i$ the aforementioned transformation is employed with success probability drawn from Uniform($[0.4, 0.6]$) for 80% of $X^{bn}_i$. The remaining 20% mimic unbalanced classes and success probability is drawn from Uniform($[0.05, 0.1]$).

Regarding $X^{end}_i$, the inverse probability integral transform is used to generate samples from the multinomial distribution. To that end, the state space is drawn from Uniform($[3, 10]$) and the corresponding probability of falling into one of these states is chosen to be proportional to the amount of states. Lastly, $X^{pos}_i$ is generated with the inverse probability integral transform and the rate parameter set equal to 6.

### 6.1 Ternary mixed data results

We now compare our method against Quan et al.’s (2018) generalization of the bridge function approach by Fan et al. [8] when a mix of ternary, binary, and continuous data is present. Recall that in this case $\binom{2+2}{2} = 6$ bridge functions are needed.

The $(d, n)$-setup is analogous to the binary mixed case. Overall, when additionally to binary also ternary data is present both according to Table 1 estimation error and graph recovery error reduce slightly. This is unsurprising as we now have more information about the underlying latent variables. Note that we expect the oracle $\hat{\Omega}$ to be roughly equivalent to the previous binary mixed case.

Starting with $\hat{\Omega}_{MLE}$ the pattern from the binary/continuous case continues to show in the ternary-binary mix: whenever $f(x) = x$ it generally performs best, in particular with respect to graph recovery.

However, when $f_j(x) = x^3$ performance drops notably which again is driven not by an increased FPR but by a decreased TPR. Again, results for $\hat{\Omega}_r$ and $\hat{\Omega}_r$ are similar although there appears to be a pattern in the sense of some evidence of smaller estimation error and better graph recovery all throughout the experiments by using our estimator $\hat{\Omega}_r$.

Overall, no performance reduction neither in terms of estimation error nor in graph recovery can be detected when comparing $\hat{\Omega}_r$ to the ternary $\hat{\Omega}_r$. In fact, the opposite is the case. Consequently, in both special cases of Definition 2.2 – the binary and the ternary mixed scheme – these results support the use of the polychoric and polyserial estimation strategies as a simpler and more general approach as compared with constructing bridge functions for every combination of variables and their state spaces.

### 7 Variable Description for real world data application

Table 2 gives an overview over the variables present in the UK Biobank data set.

| 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 |
| weight        | in kg |
| b.f. perc.    | body fat percentage in % |
| walking       | number of days per week walked 10+ minutes |
| Variable                  | Description                                                                                                                                 |
|---------------------------|--------------------------------------------------------------------------------------------------------------------------------------------|
| 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             | categorical, "Do you smoke tobacco now?" (yes, no, occasionally)                                                                           |
| past smoking              | categorical, "How often did you smoke tobacco?" (never, once/twice, occasionally, on most days)                                           |
| diet var.                 | categorical, "Does your diet change?" (never, sometimes, often)                                                                          |
| alc. freq.                | categorical, "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.                 | categorical, "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                                                                                                 |
| diabetes 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          | hayfever/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                                                                                               |
| endemotr.                 | 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                                                                                                |
| measeses                  | measesles/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. muscle/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. cataract extraction/lens implant, binary indicator with 0=no, 1=yes
red./fix. bone frac. reduction or fixation of bone fracture, binary indicator with 0=no, 1=yes
cholecysect. 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
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
\begin{table}[ht]
\centering
\begin{tabular}{|c|cccc|}
\hline
$d, n, f(x)$ & $\|\hat{\Omega} - \Omega\|_F$ & Oracle $\hat{\Omega}$ & ternary $\hat{\Omega}_r$ & $\hat{\Omega}_{MLE}$ & $\hat{\Omega}_r$ \\
\hline
50, 200, $x$ & 2.860 & (0.098) & 2.936 & (0.105) & 2.935 & (0.106) & 2.930 & (0.109) \\
 & 0.016 & (0.005) & 0.067 & (0.017) & 0.071 & (0.021) & 0.075 & (0.023) \\
 & 0.340 & (0.046) & 0.370 & (0.061) & 0.381 & (0.068) & 0.389 & (0.070) \\
 & 0.880 & (0.013) & 0.758 & (0.019) & 0.769 & (0.019) & 0.764 & (0.020) \\
\hline
50, 200, $x^3$ & 2.856 & (0.116) & 2.942 & (0.102) & 3.053 & (0.098) & 2.935 & (0.108) \\
 & 0.016 & (0.007) & 0.068 & (0.019) & 0.076 & (0.020) & 0.075 & (0.022) \\
 & 0.342 & (0.051) & 0.372 & (0.059) & 0.280 & (0.051) & 0.391 & (0.066) \\
 & 0.882 & (0.015) & 0.759 & (0.019) & 0.691 & (0.020) & 0.768 & (0.019) \\
\hline
250, 200, $x$ & 3.185 & (0.097) & 3.742 & (0.090) & 3.709 & (0.089) & 3.711 & (0.091) \\
 & 0.006 & (0.001) & 0.025 & (0.003) & 0.024 & (0.003) & 0.025 & (0.003) \\
 & 0.308 & (0.034) & 0.238 & (0.033) & 0.237 & (0.031) & 0.235 & (0.030) \\
 & 0.884 & (0.014) & 0.759 & (0.018) & 0.773 & (0.018) & 0.768 & (0.018) \\
\hline
250, 200, $x^3$ & 3.199 & (0.096) & 3.757 & (0.096) & 3.894 & (0.096) & 3.724 & (0.087) \\
 & 0.006 & (0.001) & 0.025 & (0.003) & 0.026 & (0.003) & 0.025 & (0.003) \\
 & 0.302 & (0.034) & 0.239 & (0.032) & 0.143 & (0.027) & 0.237 & (0.032) \\
 & 0.882 & (0.012) & 0.759 & (0.016) & 0.691 & (0.016) & 0.767 & (0.015) \\
\hline
750, 300, $x$ & 11.181 & (0.134) & 10.830 & (0.129) & 10.640 & (0.122) & 10.659 & (0.118) \\
 & 0.256 & (0.006) & 0.179 & (0.006) & 0.180 & (0.006) & 0.179 & (0.005) \\
 & 0.937 & (0.009) & 0.723 & (0.016) & 0.744 & (0.017) & 0.736 & (0.016) \\
 & 0.939 & (0.006) & 0.820 & (0.009) & 0.831 & (0.009) & 0.828 & (0.009) \\
\hline
750, 300, $x^3$ & 11.196 & (0.130) & 10.838 & (0.129) & 11.250 & (0.130) & 10.646 & (0.137) \\
 & 0.256 & (0.006) & 0.180 & (0.006) & 0.173 & (0.006) & 0.179 & (0.006) \\
 & 0.937 & (0.009) & 0.724 & (0.016) & 0.590 & (0.020) & 0.737 & (0.016) \\
 & 0.939 & (0.006) & 0.820 & (0.008) & 0.743 & (0.011) & 0.828 & (0.009) \\
\hline
\end{tabular}
\caption{Ternary mixed data structure learning; Simulated data with 100 simulation runs. Standard errors in brackets}
\end{table}
8 Additional Empirical Results

Empirical Results for Subset B

Figures 1 and 2 exemplify additional empirical results for subset B.

Empirical Results for Subset C

Figures 3 and 4 exemplify additional empirical results for subset C.
Figure 2: Plot of the estimated precision matrix of data set B.
Figure 3: Plot of the estimated adjacency matrix of data set C.
Figure 4: Plot of the estimated precision matrix of data set C.
References

[1] Ulf Olsson, Fritz Drasgow, and Neil J. Dorans. The polyserial correlation coefficient. *Psychometrika*, 47(3): 337–347, 1982.

[2] G. M. Tallis. The maximum likelihood estimation of correlation from contingency tables. *Biometrics*, 18:342–353, 1962.

[3] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58: 13–30, 1963.

[4] Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.*, 10:2295–2328, 2009.

[5] Pradeep Ravikumar, Martin J. Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing $\ell_1$-penalized log-determinant divergence. *Electron. J. Stat.*, 5:935–980, 2011.

[6] John E. Angus. The probability integral transform and related results. *SIAM Rev.*, 36(4):652–654, 1994.

[7] Xiaoyun Quan, James G. Booth, and Martin T. Wells. Rank-based approach for estimating correlations in mixed ordinal data, 2018. arXiv: 1809.06255.

[8] Jianqing Fan, Han Liu, Yang Ning, and Hui Zou. High dimensional semiparametric latent graphical model for mixed data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79(2):405–421, 2017.