LATENT GAUSSIAN COPULA MODEL FOR HIGH DIMENSIONAL MIXED DATA, AND ITS APPLICATIONS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Xiaoyun Quan

May 2020

LATENT GAUSSIAN COPULA MODEL FOR HIGH DIMENSIONAL MIXED DATA, AND ITS APPLICATIONS

Xiaoyun Quan, Ph.D.

Cornell University 2020

Due to the advent of "big data" technologies, mixed data that consist of both categorical and continuous variables are encountered in many application areas. We present a framework to estimate the correlation among variables of mixed data types via a rank-based approach under a latent Gaussian copula model. Theoretical properties of the correlation matrix estimator are also established. With the correlation matrix estimate $\hat{\mathbf{\Sigma}}$, we are able to further extend the topic to other problems, such as graphical models, regression, and classification. In particular, we propose a family of methods for prediction with high dimensional mixed data that involves a shrunken estimate of the inverse matrix of $\hat{\mathbf{\Sigma}}$. By maximizing the log likelihood of the data subject to a penalty on the elements of $\hat{\mathbf{\Sigma}}^{-1}$, we demonstrate that higher prediction accuracy can be achieved, compared to other popular existing methods. We also show that several existing methods are special cases of the family. In addition, we consider the classification problem via a covariance-based approach analogous to linear discriminant analysis.

BIOGRAPHICAL SKETCH

Xiaoyun Quan is a Ph.D. candidate in Statistics at Cornell University in Ithaca, New York. She attained her Bachelor of Science in Biometry and Statistics from Cornell University in 2014.

To my family

ACKNOWLEDGEMENTS

This thesis would not be possible without the help of many people.

First and foremost, I would like to thank my advisors James Booth and Martin Wells, for their endless support and countless ideas, and for their patience, motivation, and immense knowledge. I feel unbelievably lucky to have been inspired by the same two smart minds throughout both undergraduate and PhD studies. They guided me from the most basics of Statistics all the way to the state-of-the-art of it. Besides academics, they also have been generously supportive of my career pursuit. They are above and beyond advisors.

I also want to express gratitude to Yang Ning, for serving on my committee with many insightful feedback at various points, and for being a great friend with a caring heart.

I am also grateful to all other faculties and staffs at our department, for teaching me as early as when I was an undergraduate, and for many conversations we had later in the lounge or hallway of Comstock Hall.

And thanks to all my friends and fellow doctoral students, for being intellectually stimulating and emotionally supportive, and for the joy and fun we had together.

Lastly, my profound gratitude goes to my parents. They gave me the opportunity, knowledge, and strength through this long journey. Without their everlasting love and unfailing support, this work would not become a reality.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Due to the advent of "big data" technologies, mixed data that consist of both categorical and continuous variables are often encountered in many application areas. For example, in questionnaires and surveys, we commonly see categorical data such as rating scales to measure attitudes, as well as continuous data such as income and age that might be possibly related to the categorical variables. Another example is the data collected from genetics studies that consist of SNP (single nucleotide polymorphism) data of categorical values and gene expression levels of continuous values.

Estimating the associations between mixed data types is of great importance to gain insights about dependence between the variables, particularly for conditional dependencies and potential causal pathways. With this motivation, we propose a novel method to estimate associations between multilevel ordinal and continuous data using a latent Gaussian copula model approach. We assume that the multilevel ordinal variable is obtained by discretizing a latent variable, and estimate the correlation/covariance matrix underlying the Gaussian copula model via a rank-based approach. The detailed framework is described in Chapter 2, where theoretical properties are also established.

With a good estimator of covariance/correlation matrix, it is possible to expand the topic to other contexts. One natural extension is the regression problem where one would like to predict a continuous response variable using a set of predictors that consists of both categorical and continuous data types. High-dimenisonal regression has been intensively studied in the past, and popular methods often perform some regularization, such as the famous ridge regression that estimates

the linear coefficients by shrinking the inverse sample covariance matrix $(\mathbf{X^T X})^{-1}$, and the remarkably successful lasso method that gives a shrunken linear coefficient estimate subject to an ℓ_1 penalty, and the elastic net that imposes both penalties in the ridge and the lasso weighted by a certain ratio. But they all rest upon two assumptions: 1) the existence of explicit linear relationship between observable predictors and the response and 2) the observable predictors and response variables jointly follow multivariate Gaussian distribution. These two assumptions are often violated for high-dimensional mixed data. On the other hand, several approaches have been proposed that do not require the unrealistic assumptions, such as the nonparametric sparse additive models [50], the semi-parametric single index model [48], and more recently the latent Gaussian copula regression model by [10]. However, these approaches are only tailored for continuous data but not mixed data.

To bridge the gap, a latent Gaussian copula model with an ℓ_1 penalty (GC-Lasso) is proposed in Chapter 3. Under latent Gaussian copula model, the observable data are not required to be Gaussian but rather Gaussian after transformation, and this gives rise to a linear relationship between the (latent) marginally transformed data. Recall that one can rewrite the least squares solution as $\hat{\beta} = (\mathbf{X^T X})^{-1}\mathbf{X^T Y} = \hat{\Sigma}_{XX}^{-1}\hat{\Sigma}_{XY}$, this opens the door to estimating the latent linear coefficients using the covariance matrix estimator $\mathbf{\Sigma}$. We will see in Chapter 3 that under latent Gaussian copula model, one could obtain a shrunken coefficients estimator $\hat{\beta}$ for the latent linear relationship without knowing the marginal transformation. Another advantage of the GC-Lasso method is that the widely applied data transformation such as log transformation is readily adapted by the GC-Lasso model with no need to prespecify the transformation.

We then further generalize the GC-Lasso to a family of methods in Chapter 4, where we impose two layers of regularization on the inverse of the covariance matrix estimator. The shrunken inverse covariance matrix is obtained such that it maximizes the log likelihood of the latent data, under marginal normality, subject to a penalty. This shrunken inverse covariance matrix estimator is then used to to compute the regularized regression coefficients. We will see that besides GC-Lasso, this family of methods also includes several other existing methods, including the lasso and the ridge.

In each following Chapter, numerical studies are conducted to investigate the performance of the proposed methods. Case studies on real data are also demonstrated to accompany the numerical studies.

All of the research described in this dissertation was primarily performed by the author, and Chapters 2, 3 and 4 are reproductions of papers coauthored by James G. Booth and Martin T. Wells.

CHAPTER 2

RANK-BASED APPROACH FOR ESTIMATING CORRELATIONS IN MIXED ORDINAL DATA

2.1 Introduction

High-dimensional multilevel ordinal and continuous mixed data are now routinely collected in many research areas. For example, in questionnaires and surveys, we commonly see categorical data such as rating scales to measure attitudes, as well as continuous data such as income and age that might be possibly related to the categorical variables. Another example is the data collected from genetics studies that consist of SNP (single nucleotide polymorphism) data of categorical values and gene expression levels of continuous values. Estimating the associations between mixed data types is of great importance to gain insights about dependence between the variables, particularly for conditional dependencies and potential causal pathways. There are several classical rank-based methods for analyzing association among ordinal variables [3]. Specifically, those measures are all based on the numbers of concordant and discordant pairs of observations. A pair of observations, say (X_i, Y_i) and $(X_{i'}, Y_{i'})$, is concordant if the subject that has a higher ranking on X also has a higher ranking on Y, and on the other hand this pair is called discordant if the subject ranking higher on X ranks lower on Y. Kendall's *tau-a* (τ^a) was first proposed by [29] as a measure that quantifies the difference between proportions of concordant and discordant pairs among all pairs, which is essentially a correlation coefficient for sign scores. Later in 1945, a revised version called *tau-b* (τ^b) was introduced [30] that took tied pairs into consideration. [20] proposed the *gamma* measure as the difference between proportions of concordant

and discordant pairs among all concordant and discordant pairs. Other similar measures such as Somers' d [53] also considers the difference between proportions of concordant and discordant pairs, just with a different base as its denominator. More contemporarily, multilevel ordinal variables are often seen as a result of discretizing latent continuous variables [47]. [14] propose a generative latent Gaussian copula model for binary and mixed data, assuming the binary data are obtained by dichotomizing a continuous latent variable. Yet it remains an open question to measure the association between *multilevel* ordinal and continuous data.

Driven by this motivation, we propose a general framework to estimate associations between multilevel ordinal and continuous data via a latent Gaussian copula model approach. We assume that the multilevel ordinal variable is obtained by discretizing a latent variable, and estimate the correlation/covariance matrix underlying the Gaussian copula model via a rank-based approach. These results extend those for the latent Gaussian copula model for binary and continuous data proposed by [14].

In the next section we first review the concept of Gaussian copula model and define a new latent Gaussian copula model for ordinal-continuous mixed data, and then review the motivations for Kendall's rank correlation coefficient. In Section 2.3, we propose the rank-based correlation estimation for ternary-continuous mixed data and then generalize it to the estimation for ordinal-continuous mixed data. We derive explicit formulas for the bridge functions that connect the Kendall's τ^a of observed data to the latent correlation matrix for different combinations of data types. This requires derivation of new bridge functions, and those derivations are somewhat involved and more complex than in continuous/binary case. We then use these formulas to construct a rank-based estimator of the latent correlation

matrix for the mixed data. The significant advantage of bridge function technique is that it allows us to estimate the latent correlation structure of Gaussian copula without estimating marginal transformation functions. We also establish theoretical concentration bound results for the new rank-based estimators. In Section 2.4 we consider the case of tied data. Simulation results are presented in Section 2.5. In Section 2.6 we give two real data analysis that highlight the our proposed techniques applies them to the construction graphical models for mixed (binary, continuous, and ordinal) data. The first is a well example in the algorithmic fairness literature about ProPublica's journalistic investigation on the apparent biases of machine learning based predictive analytics tool, COMPAS, in recidivism risk assessment [5]. The second example is another well know example first analyzed by [9] and subsequently by [25] consisting of 12 mixed type measurements for prostate cancer patients who were diagnosed as having either stage III or IV prostate cancer. We conclude with some discussion in Section 2.7. The proofs of the main results are given in Appendix A.

2.2 Background

2.2.1 Variations of the Gaussian copula model

In recent years, the Gaussian copula model has received a lot of attention due to the ability to relax the normality assumptions of a fully Gaussian model. Formally the Gaussian copula model is defined as follows [64, 37, 36]:

Definition 1 (Gaussian copula model)**.** *A random vector* $\mathbf{X} = (X_1, \ldots, X_d)^T$ *follows Gaussian copula model if there exists a set of monotonically increasing*

transformation functions $f = (f_j)_{j=1}^d$, *such that* $f(\mathbf{X}) = (f_1(X_1), \ldots, f_d(X_d))^T \sim N_d(\mathbf{0}, \Sigma)$ *with* $diag(\Sigma) = 1$.

A random vector $\mathbf{X}$ with these properties is said to follow a nonparanormal distribution denoted by $\text{NPN}(0, \Sigma, f)$. The distribution is much more flexible that a Gaussian model. In particular, individual components of $\mathbf{X}$ can have skewed or even multimodal distributions.

Note that the Gaussian copula model only applies to continuous data. We now extend the latent Gaussian copula model to ordinal-continuous mixed data. Following the notation in the binary-continuous mixed case we consider a mixed-data random vector as $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2})$, where $\mathbf{X_1}$ is d_1-dimensional vector of p-level discrete variables (with each component of X_1 taking values in $\{0, 1, \ldots, p-1\}$) and $\mathbf{X_2}$ is a d_2-dimensional vector of continuous variables.

Definition 2 (Latent Gaussian copula models for ordinal-continuous data)**.** *The random vector* $\mathbf{X}$ *follows the extended latent Gaussian copula model if there exists a* d_1*-dimensional random vector of latent variables* $\mathbf{Z_1} = (Z_1, \ldots, Z_{d_1})$ *such that* $X_j = l$ *if* $Z_j \in (C_j^l, C_j^{l+1})$ *for* $l = 0, 1, \ldots, p-1$ *and* $j = 1, \ldots, d_1$*, where the cutoff vector is given by* $\mathbf{C} = (\mathbf{C}_1, \ldots, \mathbf{C}_{d_1})$ *and* $\mathbf{C}_j = (C_j^0 = -\infty, C_j^1, \ldots, C_j^{p-1}, C_j^p = \infty)$ *is an increasing sequence of* $(p-1)$ *constants, and* $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{X}_2) \sim NPN(\mathbf{0}, \mathbf{\Sigma}, f)$.

The latent Gaussian copula model for binary-continuous data is just a special case of the above latent Gaussian copula model with $p = 2$. Alternatively, the binary case is retrieved if $C_j^2 = \infty$ for $j = 1, \ldots, d_1$. In fact, by setting $C_j^{k_j} = \infty$, where $2 \leq k_j \leq p$ for $j = 1, \ldots, d_1$, we can handle situations with ordinal variables with differing numbers of levels. [14] proposed the following latent Gaussian copula model as an extension to binary and mixed binary-continuous data:

Definition 3 (Latent Gaussian copula model for binary-continuous mixed data)**.** *Consider a mixed-data random vector* $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2})$ *where* $\mathbf{X_1}$ *is a* d_1*-dimensional vector of binary variables and* $\mathbf{X_2}$ *is a* d_2*-dimensional vector of continuous variables. Then* $\mathbf{X}$ *follows a latent Gaussian copula model if there exists a* d_1*-dimensional random vector of latent variables* $\mathbf{Z_1} = (Z_1, \ldots, Z_{d_1})$ *such that* $X_j = I(Z_j > C_j)$ *for* $j = 1, \ldots, d_1$ *where* $\mathbf{C} = (C_1, \ldots, C_{d_1})$ *is a* d_1*-dimensional vector of constants, with* $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{X}_2) \sim NPN(\mathbf{0}, \mathbf{\Sigma}, f)$*.*

Our interest is in estimating the correlation matrix $\mathbf{\Sigma}$ or the precision matrix $\mathbf{\Omega} = \mathbf{\Sigma^{-1}}$ with for latent Gaussian copula models for ordinal-continuous data. Furthermore, under the Gaussian copula model, the sparsity pattern of the precision matrix $\mathbf{\Omega}$ reveals the conditional dependencies between $X_j's$ for $j = 1, 2, ..., d$. Hence the graph structure could also be recovered by estimating $\mathbf{\Sigma}^{-1}$ as in the prostate cancer diagnostic example in Section 2.6.2.

2.2.2 Kendall's rank correlation coefficients

Kendall's τ^a (Kendall's rank correlation coefficient) is a nonparametric measure of nonlinear dependence between two random variables. It is similar to Spearman's ρ and Pearson's r, in that is measures the relationship between two variables. Even though Kendall's τ^a is a similar to Spearman's ρ in that it is a nonparametric measure of relationship it differs in the interpretation of the correlation value. Spearman's ρ and Pearson's r magnitude are similar, however, Kendall's τ^a is the difference between the probability that the observed data are in the same order versus the probability that the observed data are not in the same order.

Suppose the data consists of n independent d-dimensional random vectors,

$\mathbf{X}_1, \ldots, \mathbf{X}_n$, from a latent Gaussian copula model. The rank-based estimation framework for $\mathbf{\Sigma}$, depending on the data type. Specifically, estimation is based on the "bridge function" that relates Kendall's τ^a parameter, τ^a_{jk}, for each variable pair (j,k), $1 < j < k < d$, with the correlation, σ_{jk}, between them. Here, the parameter τ^a_{jk} is given by

$$\tau^a_{jk} = \mathbb{E}\Big[\text{sign}\{(X_{ij} - X_{i'j})(X_{ik} - X_{i'k})\}\Big], \tag{2.1}$$

which can be estimated unbiasedly by the corresponding τ^a statistic

$$\hat{\tau}^a_{jk} = \binom{n}{2}^{-1} \sum_{1 \le i < i' \le n} \text{sign}(X_{ij} - X_{i'j})(X_{ik} - X_{i'k}), \tag{2.2}$$

or equivalently by

$$\hat{\tau}^a_{jk} = \frac{C - D}{\binom{n}{2}} \tag{2.3}$$

where C and D are the number of concordant and discordant pairs among $(X_{1j}, X_{1k}), \ldots, (X_{nj}, X_{nk})$.

A variation of Kendall's τ^a that accounts for the important case of ties is τ^b. Binary and ordinal data are very likely to have a large number of ties in ranking and, as a result, Kendall's τ^a is likely to under-estimate the sample correlation. Therefore we consider a modified version, known as Kendall's τ^b.

$$\hat{\tau}^b_{jk} = \hat{\tau}^a_{jk} \frac{\binom{n}{2}}{\sqrt{\Big[\binom{n}{2} - t_{X_j}\Big]\Big[\binom{n}{2} - t_{X_k}\Big]}} \tag{2.4}$$

where $t_{X_j} = \sum_{1 \le i < i' \le n} I(X_{ij} = X_{i'j})$ is the number of pairs of tied values of the jth response, and similarly $t_{X_k} = \sum_{1 \le i < i' \le n} I(X_{ik} = X_{i'k})$.

Since Kendall's τ^b is a ratio of random terms (and the denominator involves a square root), the population bridge function linking it to σ_{jk} is intractable. We therefor consider 1st-order and 2nd order Taylor series approximation instead of

directly computing its expectation. However, we find there is almost no difference between the 1st- and 2nd-order Taylor series approximations, or between them and a Monte-Carlo approximation of the exact expectation..

2.3 Methodology

Suppose the data consists of n independent d-dimensional random vectors, $\mathbf{X}_1, \ldots, \mathbf{X}_n$, from a latent Gaussian copula model. In this section, we propose a rank-based estimation framework for $\mathbf{\Sigma}$, depending on the data type. Specifically, estimation is based on the "bridge function" that relates Kendall's τ^a parameter, τ^a_{jk}, for each variable pair (j, k), $1 \leq j < k \leq d$, with the correlation, σ_{jk}, between them. The main idea behind our alternative procedure is to exploit Kendall's τ^a statistics to directly estimate the unknown correlation matrix, without explicitly calculating the marginal transformation functions f_j. Recall that the Kendall τ^a statistics are invariant under monotonic transformations. For Gaussian random variables there is a one-to-one mapping between these two statistics. For Gaussian copula distributions Kendall's τ^a is connected to the covariance matrix in Definition 1 by $\sigma_{jk} = \sin(\frac{\pi}{2}\tau^a_{ij})$.

2.3.1 Estimate correlation between ternary and ternary data

We begin by considering ternary (3-level) data, and then extend to the general p-level case in Section 2.3.3. Now suppose $\mathbf{X}_j$, $j = 1, \ldots, d_1$ are discrete data with 3 categories, taking values $\{0, 1, 2\}$. Then under latent Gaussian copula model, we

have $p = 3$, and the data are obtained by trichotomizing the latent variable Z_j at cutoffs (C_j^1, C_j^2), $C_j^1 < C_j^2$, such that

$$X_{ij} = \begin{cases} 0 & \text{if } f(Z_{ij}) \leq \Delta_j^1 \\ 1 & \text{if } \Delta_j^1 < f(Z_{ij}) \leq \Delta_j^2 \\ 2 & \text{if } f(Z_{ij}) > \Delta_j^2 \end{cases}$$

where $\Delta_j^l = f(C_j^l)$, for $l = 1, 2$.

To estimate $\mathbf{\Sigma}$, we divide this into 3 cases where: (i) for $1 \leq j, k \leq d_1$, σ_{jk} is the correlation between ternary variables; (ii) for $1 \leq j \leq d_1 < k \leq d$, σ_{jk} is the correlation between ternary and continuous variables; and (iii) for $d_1 < j, k \leq d$, σ_{jk} is the correlation between continuous variables. In case (iii) it has been shown by [31] that $r_{jk} = \sin\left(\frac{\pi}{2}\hat{\tau}_{jk}^a\right)$. In the remainder of this section, we confine our attention to cases (i) and (ii) respectively.

We first consider Kendall's τ^a for two tenary variables. There are only four cases that need to be considered in order to determine concordance and discordance:

$$(X_{ij} \leq 1, X_{ik} \leq 1); \quad (X_{ij} \geq 1, X_{ik} \geq 1); \quad (X_{ij} \leq 1, X_{ik} \geq 1); \quad (X_{ij} \geq 1, X_{ik} \leq 1)\,.$$

Combining the first two will give a concordant pair and combining the last two will give a discordant pair. So using equations (2.1) and (2.3) we can directly calculate

the "bridge function" between τ^a_{jk} and σ_{jk} as

$$\begin{aligned}
\tau^a_{jk} &= \mathbb{P}(C) - \mathbb{P}(D) \\
&= \mathbb{P}(X_{ij} \le 1; X_{ik} \le 1)\mathbb{P}(X_{i'j} \ge 1; X_{i'k} \ge 1) \\
&\quad + \mathbb{P}(X_{ij} \ge 1; X_{ik} \ge 1)\mathbb{P}(X_{i'j} \le 1; X_{i'k} \le 1) \\
&\quad - \mathbb{P}(X_{ij} \le 1; X_{ik} \ge 1)\mathbb{P}(X_{i'j} \ge 1; X_{i'k} \le 1) \\
&\quad - \mathbb{P}(X_{ij} \ge 1; X_{ik} \le 1)\mathbb{P}(X_{i'j} \le 1; X_{i'k} \ge 1)
\end{aligned}$$

(all the tied pairs cases in the first line will cancel out from those in the last two lines)

$$\begin{aligned}
&= 2\Phi_2(\Delta^2_j, \Delta^2_k, \sigma_{jk})\Phi_2(-\Delta^1_j, -\Delta^1_k, \sigma_{jk}) \\
&\quad - 2\Big[\Phi(\Delta^2_j) - \Phi_2(\Delta^2_j, \Delta^1_k, \sigma_{jk})\Big]\Big[\Phi(\Delta^2_k) - \Phi_2(\Delta^1_j, \Delta^2_k, \sigma_{jk})\Big], \qquad (2.5)
\end{aligned}$$

where the last step follows from

$$\mathbb{P}(X_{ij} \le 1, X_{ik} \le 1) = \mathbb{P}(f_j(Z_{ij}) \le \Delta^2_j, f_k(Z_{ik}) \le \Delta^2_k) = \Phi_2(\Delta^2_j, \Delta^2_k, \sigma_{jk});$$

$$\mathbb{P}(X_{ij} \le 1, X_{ik} \ge 1) = \mathbb{P}(X_{ij} \le 1) - \mathbb{P}(X_{ij} \le 1, X_{ik} \le 1) = \Phi(\Delta^2_j) - \Phi_2(\Delta^2_j, \Delta^2_k, \sigma_{jk}).$$

The notation $\Phi_2(u, v, r)$ denotes the CDF of standard bivariate normal distribution with correlation r, namely $\Phi_2(u, v, r) = \int_{x_1<u}\int_{x_2<v} \phi_2(x_1, x_2; r)dx_1dx_2$ where $\phi_2(x_1, x_2; r)$ is the probability density function of the standard bivariate normal distribution with correlation r.

It follows that the bridge function for the population Kendall's τ^a for variable pair (j, k), is given by $\tau^a_{jk} = F(\sigma_{jk}; \Delta^1_j, \Delta^2_j, \Delta^1_k, \Delta^2_k)$ where

$$\begin{aligned}
F_a(\sigma_{jk}; \Delta^1_j, \Delta^2_j, \Delta^1_k, \Delta^2_k) &= 2\Phi_2(\Delta^2_j, \Delta^2_k, \sigma_{jk})\Phi_2(-\Delta^1_j, -\Delta^1_k, \sigma_{jk}) \\
&\quad - 2\Big[\Phi(\Delta^2_j) - \Phi_2(\Delta^2_j, \Delta^1_k, \sigma_{jk})\Big]\Big[\Phi(\Delta^2_k) - \Phi_2(\Delta^1_j, \Delta^2_k, \sigma_{jk})\Big]. \qquad (2.6)
\end{aligned}$$

It will be shown in Lemma 1 that, for fixed $\Delta^1_j, \Delta^2_j, \Delta^1_k, \Delta^2_k$, the function $F_a(\sigma_{jk}; \Delta^1_j, \Delta^2_j, \Delta^1_k, \Delta^2_k)$ is an invertible function of σ_{jk}.

Simple moment estimators can be derived for the cutoffs using the relations

$$\mathbb{E}(\mathbb{1}\{X_{ij}=0\}) = \Phi(\Delta_j^1) \quad \text{and} \quad \mathbb{E}(\mathbb{1}\{X_{ij}=2\}) = 1-\Phi(\Delta_j^2)\,.$$

Specifically, these motivate the estimators

$$\hat{\Delta}_j^1 = \Phi^{-1}\left(\frac{\sum_i \mathbb{1}\{X_{ij}=0\}}{n}\right) \quad \text{and} \quad \hat{\Delta}_j^2 = \Phi^{-1}\left(1-\frac{\sum_i \mathbb{1}\{X_{ij}=2\}}{n}\right).$$

Thus a rank-based estimator of σ_{jk} is given by

$$\hat{R}_{jk} = F_a^{-1}(\hat{\tau}_{jk}^a; \hat{\Delta}_j^1, \hat{\Delta}_j^2, \hat{\Delta}_k^1, \hat{\Delta}_k^2). \tag{2.7}$$

As will be seen from the following lemma, the bridge function $F_a(\sigma_{jk}; \hat{\Delta}_j^1, \hat{\Delta}_j^2, \hat{\Delta}_k^1, \hat{\Delta}_k^2)$ is strictly increasing in σ_{jk}, thus there exists a unique root for the equation $F_a(\sigma_{jk}; \hat{\Delta}_j^1, \hat{\Delta}_j^2, \hat{\Delta}_k^1, \hat{\Delta}_k^2) = \hat{\tau}_{jk}^a$ which can be efficiently solved by Newton's method.

Lemma 1. *For any fixed $\Delta_j^1, \Delta_j^2, \Delta_k^1, \Delta_k^2$, $F_a(r; \Delta_j^1, \Delta_j^2, \Delta_k^1, \Delta_k^2)$ in equation (2.6) is a strictly increasing function on $r \in (-1, 1)$. Thus, the inverse function $F_a^{-1}(\tau^a; \Delta_j^1, \Delta_j^2, \Delta_k^1, \Delta_k^2)$ exists.*

The proof of Lemma 1 is given in Appendix A.

We note here that the bridge functions for the binary-ternary and binary-binary cases can be derived directly from (2.6), by setting $\Delta_j^2 = \infty$ and both $\Delta_j^2 = \infty$ and $\Delta_k^2 = \infty$ respectively. Using the identities $\Phi_2(\infty, v, r) = \Phi(v)$, $\Phi_2(u, \infty, r) = \Phi(u)$, and $\Phi_2(-u, -v, r) = 1 - \Phi(u) - \Phi(v) + \Phi_2(u, v, r)$, we find

$$\begin{aligned} F_a(\sigma_{jk}; \Delta_j^1, \infty, \Delta_k^1, \Delta_k^2) = {} & 2\Phi_2(\Delta_j^1, \Delta_k^2, \sigma_{jk}) \left(1-\Phi(\Delta_k^1)\right) \\ & - 2\Phi(\Delta_k^2) \left(\Phi(\Delta_j^1) - \Phi_2(\Delta_j^1, \Delta_k^1, \sigma_{jk})\right), \end{aligned} \tag{2.8}$$

and

$$F_a(\sigma_{jk}; \Delta_j^1, \infty, \Delta_k^1, \infty) = 2\left(\Phi_2(\Delta_j^1, \Delta_k^1, \sigma_{jk}) - \Phi(\Delta_j^1)\Phi(\Delta_j^2)\right), \tag{2.9}$$

the latter agreeing with equation (3) of [14].

2.3.2 Estimate correlation between ternary and continuous data

We now consider the random vector pairs (X_{ij}, X_{ik}) where variable j is ternary and variable k continuous. The latent Gaussian copula model assumptions imply that the corresponding latent vector pairs, (Z_{ij}, X_{ik}), satisfy

$$(U_{ij}, V_{ik}) \equiv (f_j(Z_{ij}), f_k(X_{ik})) \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \sigma_{jk}\\ \sigma_{jk} & 1\end{bmatrix}\right)$$

independently, for $i = 1, \ldots, n$, where σ_{jk} is the correlation between U_{ij} and V_{ik}.

It follows that

$$(U_{ij}, U_{i'j}, \frac{V_{ik} - V_{i'k}}{\sqrt{2}})^T \sim N_3\left(\begin{bmatrix}0\\0\\0\end{bmatrix}, \begin{bmatrix}1 & 0 & \sigma_{jk}/\sqrt{2}\\ 0 & 1 & -\sigma_{jk}/\sqrt{2}\\ \sigma_{jk}/\sqrt{2} & -\sigma_{jk}/\sqrt{2} & 1\end{bmatrix}\right).$$

Let Φ_3 denote the CDF for $(U_{ij}, U_{i'j}, \frac{V_{ik}-V_{i'k}}{\sqrt{2}})$,

$$\Phi_3(a, b, c) = \mathbb{P}(U_{ij} < a, U_{i'j} < b, \frac{V_{ik} - V_{i'k}}{\sqrt{2}} < c). \tag{2.10}$$

Now we are ready to build the bridge function of the population Kendall's τ^a for ternary and continuous variables as follows.

Lemma 2. *When X_{ij} is ternary and X_{ik} is continuous, $\tau^a_{jk} = E(\hat{\tau}^a_{jk})$ is given by $\tau^a_{jk} = F_a(\sigma_{jk}; \Delta^1_j, \Delta^2_j)$ where*

$$F_a(\sigma_{jk}; \Delta^1_j, \Delta^2_j) = 4\Phi_2(\Delta^2_j, 0, \sigma_{jk}/\sqrt{2}) - 2\Phi(\Delta^2_j) + 4\Phi_3(\Delta^1_j, \Delta^2_j, 0) - 2\Phi(\Delta^1_j)\Phi(\Delta^2_j)\,. \tag{2.11}$$

The next lemma shows that, for fixed Δ^1_j, Δ^2_j, $F(r; \Delta^1_j, \Delta^2_j)$ is an invertible function of r, which implies that the equation has unique solution $\hat{r} = F_a^{-1}(\hat{\tau}^a_{jk}; \hat{\Delta}^1_j, \hat{\Delta}^2_j)$ where the unknown cut-offs Δ^1_j, Δ^2_j can be estimated with no bias by considering their expectations: $\hat{\Delta}^1_j = \Phi^{-1}\left(\frac{\sum_i \mathbb{1}\{X_{ij}=0\}}{n}\right)$ and $\hat{\Delta}^2_j = \Phi^{-1}\left(1 - \frac{\sum_i \mathbb{1}\{X_{ij}=2\}}{n}\right)$.

Lemma 3. *For any fixed Δ^1_j, Δ^2_j, $F_a(r; \Delta^1_j, \Delta^2_j)$ in equation (2.11) is a strictly increasing function on $r \in (-1, 1)$. Thus, the inverse function $F_a^{-1}(\tau^a; \Delta^1_j, \Delta^2_j)$ exists.*

The proof of Lemma 3 can be found in Appendix A.3.

Combining all three lemmas, we have constructed the rank-based estimate of $\mathbf{\Sigma}$ as follows:

$$r_{jk} = \begin{cases} F_a^{-1}(\hat{\tau}^a_{jk}; \hat{\Delta}^1_j, \hat{\Delta}^2_j, \hat{\Delta}^1_k, \hat{\Delta}^2_k) & \text{for} 1 \le j, k \le d_1 \\ F_a^{-1}(\hat{\tau}^a_{jk}; \hat{\Delta}^1_j, \hat{\Delta}^2_j) & \text{for } 1 \le j \le d_1 < k \le d \\ \sin\left(\frac{\pi}{2}\hat{\tau}^a_{jk}\right) & \text{for } d_1 < j, k \le d \end{cases}$$

2.3.3 Generalized rank-based estimate for p-level discrete-continuous mixed data

We now generalize the rank-based estimate to p-level discrete-continuous mixed data. Suppose that $\mathbf{X}_j$ is a p-level ordinal variable and $\mathbf{X}_k$ continuous, then the bridge function for p-level discrete-continuous mixed data is established in the following lemma:

Lemma 4. *When X_{ij} is p-level discrete taking value in $\{0, 1, \ldots, p-1\}$, and X_{ik} is continuous, the population version of Kendall's τ^a is given by $\tau^a_{jk} = F_a(\sigma_{jk}; \mathbf{\Delta}_j)$, where*

$$F_a(\sigma_{jk}; \mathbf{\Delta}_j) = \sum_{l=1}^{p-1} 4\Phi_3(\Delta_j^l, \Delta_j^{l+1}, 0) - 2\Phi(\Delta_j^l)\Phi(\Delta_j^{l+1}). \tag{2.12}$$

Moreover, if we consider the entire $\mathbb{Z}^+$ space, we can extend the estimates $\hat{\mathbf{\Delta}}_\mathbf{j}$ as

$$\hat{\Delta}_j^l = \Phi^{-1}\left(\frac{\sum_{i=1}^n I(X_{ij} \leq l-1)}{n}\right) \qquad \text{for } l \in \mathbb{Z}^+$$

so that for p-level mixed data ranging from $0, \ldots, p-1$ for $l \geq p$ we have $\hat{\Delta}_j^l = \infty$. Then for $l \geq p$, $4\Phi_3(\Delta_j^l, \Delta_j^{l+1}, 0) - 2\Phi(\Delta_j^l)\Phi(\Delta_j^{l+1}) = 4 \times \frac{1}{2} - 2 \times 1 \times 1 = 0$. Therefore we can write the ∞-form bridge function as:

$$F_a(r; \mathbf{\Delta}_j) = \sum_{l=1}^{\infty} 4\Phi_3(\Delta_j^l, \Delta_j^{l+1}, 0) - 2\Phi(\Delta_j^l)\Phi(\Delta_j^{l+1}).$$

2.3.4 Theoretical results

We now are ready to establish a theoretical result concerning the convergence rate of the correlation estimate. As mentioned in [14], these two assumptions impose little restrictions in practice.

*A*ssumption 1: (bounded correlations) There is a constant $\delta \geq 0$ such that $|\sigma_{jk}| \leq 1 - \delta$ for $1 \leq j < k \leq d$.

*A*ssumption 2: (bounded cut-offs) There is a constant M such that $|\Delta_j^1| \leq M$ and $|\Delta_j^2| \leq M$ for any $j = 1, \ldots, d$.

In the case of the estimate of correlation between p-level ordinal and continuous data we have the following concentration result.

Theorem 1. *Under Assumptions 1 and 2, at fixed p, for any $t > 0$ we have the following property*

$$P\Big(\big|r_{jk} - \sigma_{jk}\big| > t\Big) \leq 2d^2 p \exp\left(-\frac{2M^2 n}{L_1^2}\right) + 2d^2 \exp\Big(-\frac{nt^2}{2L_2^2}\Big) + 2d^2 p \exp(-\frac{4nt^2\pi}{24^2 L_1^2 L_2^2 p^2})$$

implying that with probability greater than $1 - d^{-1}$

$$\sup_{1 \leq j < k \leq d} |\Sigma - R| < C\sqrt{\frac{\log d}{n}}$$

where L_1 and L_2 are positive constants defined in Appendix A5. Essentially, Theorem 3.1 implies that for some constant ω independent of n and d, $\sup_{1 \leq j < k \leq d} ||R - \Sigma|| \leq \omega\sqrt{(\log d)/n}$ with probability $1 - d^{-1}$.

We have a similar concentration rate for the correlation estimator of ternary-continuous mixed data.

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

$$\begin{aligned} P\Big(\big|r_{jk} - \sigma_{jk}\big| > t\Big) \leq 4\exp\left(-\frac{2M^2 n}{L_1^2}\right) + 2\exp\Big(-\frac{nt^2}{2L_2^2}\Big) \\ + 2\exp(-\frac{nt^2\pi}{48^2 L_1^2 L_2^2}) + 2\exp(-\frac{nt^2\pi}{24^2 L_1^2 L_2^2}). \end{aligned}$$

2.4 Kendall's τ^b for tied data

Here we propose another correlation estimate for binary data as a variant to the one proposed by [14], with the bridge function given by (2.9).

2.4.1 Kendall's τ^b estimate for binary and binary variables

Lemma 5. *When X_{ij} and X_{ik} are both binary discrete random variables, the 1st-order Taylor series approximation of the population version of Kendall's τ^b, given by $\tau^b_{jk} = E(\hat{\tau}^b_{jk})$, is*

$$F_b(\sigma_{jk}; \Delta_j, \Delta_k) = \frac{\Phi_k(\Delta_j, \Delta_k, \sigma_{jk}) - \Phi(\Delta_j)\Phi(\Delta_k)}{\sqrt{(\Phi(\Delta_j) - \Phi(\Delta_j)^2)(\Phi(\Delta_k) - \Phi(\Delta_k)^2)}}.$$

We can easily see that $F_b(\sigma_{jk}; \Delta_j, \Delta_k)$ is strictly increasing in σ_{jk} since the denominator is independent of σ_{jk} and the numerator is the bridge function for Kendall's τ^a. Therefore the equation $r_{jk} = F_b^{-1}(\hat{\tau}^b_{jk}; \hat{\Delta}_j, \hat{\Delta}_k)$ has a unique solution.

2.4.2 Kendall's τ^b estimate for binary and continuous variables

Lemma 6. *When X_{ij} is binary and X_{ik} is continuous, the 1st-order Taylor series approximation of the population version of Kendall's τ^b, given by $\tau^b_{jk} = E(\hat{\tau}^b_{jk})$, is*

$$F_b(\sigma_{jk}; \Delta_j) = \frac{4\Phi_2(\Delta_j, 0, \sigma_{jk}/\sqrt{2}) - 2\Phi(\Delta_j)}{\sqrt{2(\Phi(\Delta_j)) - 2(\Phi(\Delta_j))^2}}.$$

This bridge function is also strictly increasing in $\sigma_{jk} \in (-1, 1)$ because the denominator does not involve σ_{jk} and the numerator has been shown to be monotonically

increasing by [14].

We also derived the 2nd-order Taylor approximation of the bridge function in this case.

Lemma 7. *Let* $T_j = \sqrt{\binom{n}{2} - t_{X_j}}$. *Then, the 2nd-order Taylor approximation of* $E(\hat{\tau}^b_{jk})$ *is given by*

$$\mathbb{E}(\hat{\tau}^b_{jk}) \approx \frac{\sqrt{\binom{n}{2}}\mathbb{E}(\hat{\tau}^a_{jk})}{\mathbb{E}(T_j)} + [\mathbb{E}(T_j)]^{-2}\left[\binom{n}{2} var(T_j)\frac{\sqrt{\binom{n}{2}}\mathbb{E}[\hat{\tau}^a_{jk}]}{\mathbb{E}(T_j)} - cov\Big(\sqrt{\binom{n}{2}}\hat{\tau}^a_{jk}, T_j\Big)\right]$$

where

$$\mathbb{E}[\hat{\tau}^a_{jk}] = 4\Phi_2(\Delta_j, 0, r/\sqrt{2}) - 2\Phi(\Delta_j);$$

$$\mathbb{E}(T_j) = \sum_{n_0=0}^{n}\left[\sqrt{\binom{n}{2} - \binom{n_0}{2} - \binom{n-n_0}{2}}\right]\binom{n}{n_0}\Big(\Phi(\Delta_j)\Big)^{n_0}\Big(1-\Phi(\Delta_j)\Big)^{n-n_0};$$

$$var(T_j) = \binom{n}{2}\Big(2\Phi(\Delta_j) - 2[\Phi(\Delta_j)]^2\Big) - \mathbb{E}(T_j)^2;$$

$$cov\left(\sqrt{\binom{n}{2}}\tau^a_{jk}, T_j\right) = \sum_{(C,D)\in S}\left\{(C-D)\sqrt{(C+D)}\frac{\sqrt{\binom{n}{2}}}{C!D!\left(\binom{n}{2}-C-D\right)!}\cdot p_C^C p_D^D(1-p_C-p_D)^{\binom{n}{2}-C-D}\right\} - \sqrt{\binom{n}{2}}\mathbb{E}(\hat{\tau}^a_{jk})\mathbb{E}(T_j)$$

with the sample space of (C, D) *being* $S = \{(C, D) : C \in \mathbb{Z}^+, D \in \mathbb{Z}^+, C + D \leq n\}$, *the probability of concordance and discordance respectively as* $p_C = 2(\Phi_2(\Delta_j, 0, \sigma_{jk}/\sqrt{2}) - \Phi_3(\Delta_j, \Delta_j, 0))$ *and* $p_D = 2(\Phi_2(\Delta_j, 0, -\sigma_{jk}/\sqrt{2}) - \Phi_3(\Delta_j, \Delta_j, 0))$.

The 1st-order and 2nd-order Taylor and Monte Carlo approximations to τ^b are

plotted in Figure 2.1 (right panel) for $n = 84$, and $\Delta_j = 0$. The difference between the two Taylor approximations is shown in the right panel.

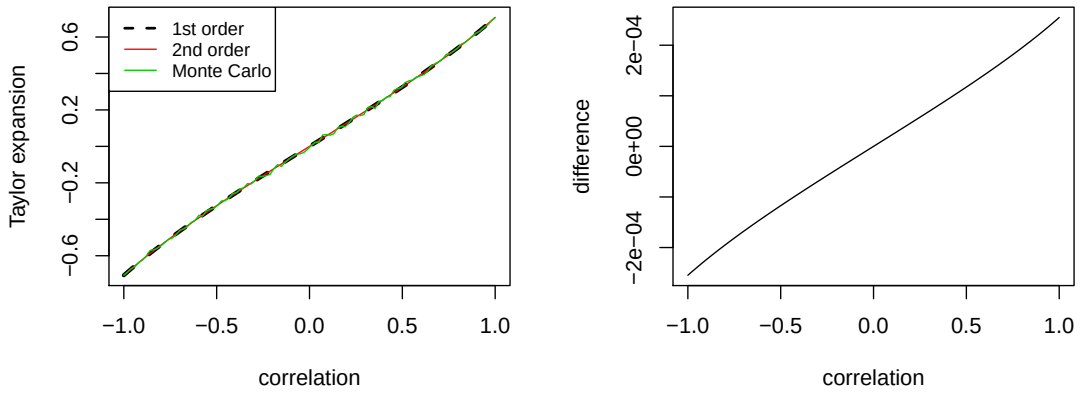

Figure 2.1: Left: 1st-order and 2nd-order Taylor approximations visually overlap with Monte-carlo simulated averages. Right: the difference between 1st-order and 2nd-order Taylor approximations are negligible.

2.5 Simulation results for generalized p-level mixed data

In this section, we show some simulation results for p-level mixed data where $p = 2, 3, \ldots, 16$. We conducted two scenarios here:

*S*cenario 1: Starting with $p = 2$, we dichotomize the data equally by setting the cutoff $\Delta_j = 0$. With p increasing, we discretize the data by setting the cutoff $\Delta_j^p = \Phi^{-1}(1/p)$ so that we will have equal counts of each level.

*S*cenario 2: Starting with $p = 16$, we discretize the continuous Gaussian copula data equally so that each level has about the same number of counts. As p decreases, we combine the highest level with one level lower: e.g. when $p = 15$, we collapse "16"s into "15"s. The motivation is that in Genetics research, when

encountering ternary data, people sometimes combine "1"s and "2"s to make the data binary. As we can see in the following plot, this will lead to an increased estimation error (see leftmost plots in Figure 2.2).

For each scenario, we first simulate bivariate Gaussian copula data of size $n = 100$, $d = 2$ and $f(x) = x$, with the correlation/covariance $r = \{0, 0.01, \ldots, 1\}$, and we estimate the correlations using the continuous data. Then we discretize the first dimension of the data into p level in the way described by each Scenario, and estimate the correlation following the bridge function in equation (2.12). For each r, the same process is repeated by 80 times and we take their mean of the squares as the error measure. We further smooth the curve by averaging the errors over $r \in [0, 0.1)$, $r \in [0.1, 0.2)$, etc.

We can see from the following plot that as p increases, the estimation error approaches to the one in raw continuous data. However, notice that how much estimation error will be introduced by combining levels as we can see in Figure 2.2.

2.6 Real data analysis

In this section, we present two studies of real data analysis. We start with applying our correlation estimation method to two sets of real data that have been studied intensively in the past, and then pass the correlation estimator to graph estimation procedures in next step. In the graph estimation procedure, we adopt the modified graphical lasso estimation method as in [14], which essentially consists of two steps: first we project the correlation estimator $\hat{\mathbf{R}}$ into the cone of positive semidefinite matrices to facilitate the optimization algorithms in [17], denoted as $\hat{\mathbf{R}}_p$; second we pass $\hat{\mathbf{R}}_p$ to the graphical lasso estimation to replace the sample covariance matrix,

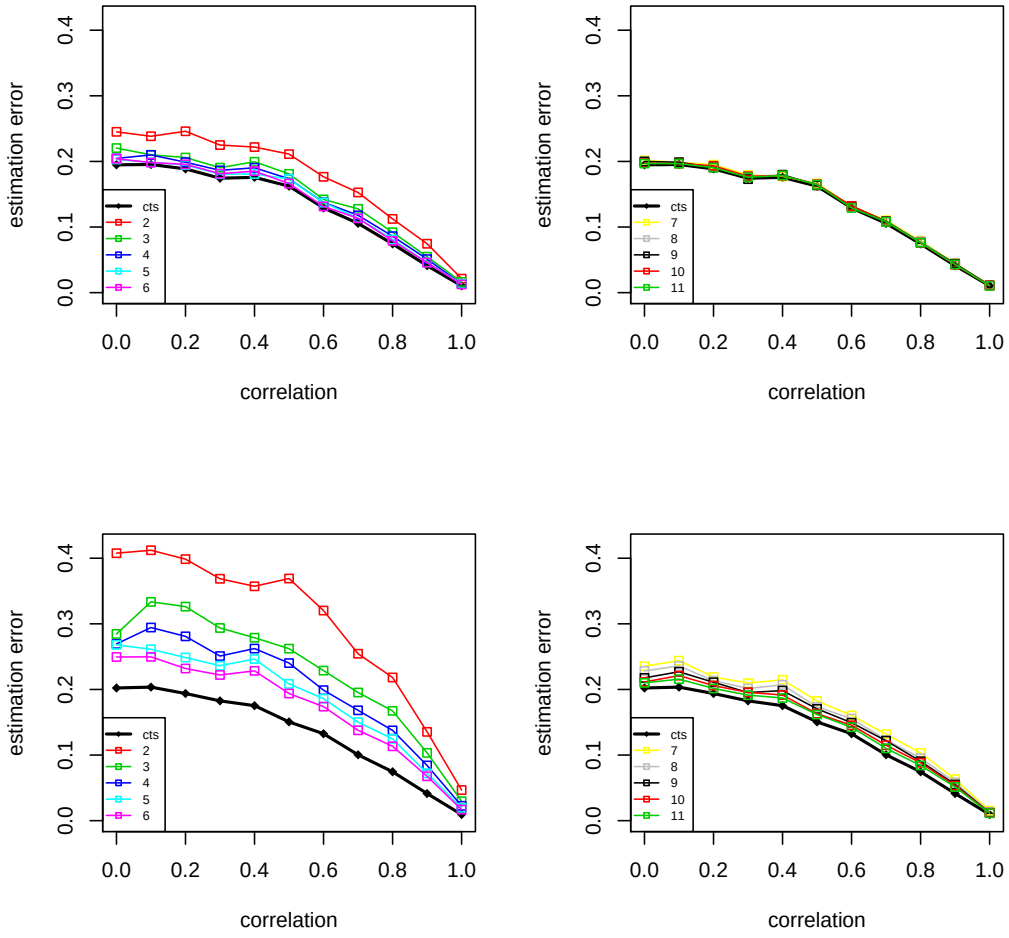

Figure 2.2: Top: Simulation results for Scenario 1. For every p, each level of data has about the same size. As p increase, the estimation error gets close to the one without discretizing the data. Bottom:Simulation results for Scenario 2. For every p, each level of data has about the same size. As p increase, the estimation error gets close to the one without discretizing the data.

to obtain the following precision matrix estimator:

$$\hat{\mathbf{\Omega}} = \underset{\mathbf{\Omega} \succeq 0}{\arg\min} \{ \text{tr}(\hat{\mathbf{R}}_p \mathbf{\Omega}) - \log |\mathbf{\Omega}| + \lambda \sum_{j \neq k} |\Omega_{jk}| \}.$$

We set the path of tuning parameter to be the vector of length 10 starting from $\frac{\max |\hat{\mathbf{R}}_p|}{10}$ to $\max |\hat{\mathbf{R}}_p|$, as suggested by [17]. Furthermore, we did not penalize the diagonal of inverse covariance matrix. We used high-dimensional BIC score (HBIC) as selection criterion, defined in [14]. The estimated graphs are then presented to reveal conditional independence relationships.

2.6.1 COMPAS Data

ProPublica [5] carried out a journalistic investigation on possible biases of machine learning based predictive analytic tools used in criminal justice. The ProPublica article examined whether black-box risk assessment tools disproportionately recommend nonrelease of African-American defendants. COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a proprietary software tool developed by Northpointe, Inc. that gives a prediction score for a defendant's likelihood of failing to appear in court or reoffending. [5] compiled criminal records from the criminal justice system in Broward County, Florida, combining detailed individual level criminal histories with predictions from the COMPAS risk assessment tool. This data set has served as a key example in the algorithmic fairness literature (e.g. [2, 7, 11, 28, 32, 55, 67]).

The COMPAS score is computed by a black-box algorithm and produces a decile score (deciles of the predicted probability of rearrest) as well as a (ordinal) categorical score consisting of three levels of risk (low, medium, and high). [13] suggest that a medium and high COMPAS scores garner more interest from

supervision agencies than low scores. In order to assess the accuracy of the recidivism predictions, [33] compared individual COMPAS score based predictions to a ground truth indicator of whether that particular individual had indeed been rearrested within two years of release. [33] developed a binary logistic regression model (low versus medium or high) that considered race, age, criminal history, future recidivism, and charge degree, they analyzed both the COMPAS scores for risk of overall and violent recidivism and used their model to assess the odds of getting a higher COMPAS score for certain subgroups.

The ProPublica article [5] mentions three African-Americans that had a medium risk COMPAS score and no subsequent offenses whereas non-African Americans had low risk score but had subsequent serious offenses. So it is of interest to examine the three level COMPAS score (low, medium, and high) rather than the binary classification in [33]. We use our proposed graphical model approach to examine the conditional independence relationship between two-year recidivism (binary) and the three level (both overall and violent) COMPAS score (low, medium, and high) with gender (binary), recorded misdemeanor (binary), age category (¡25, 25-45, and ¿45), number of priors, and juvenile criminal history (felony, misdemeanor, and other – all binary). To better understand the underlying relationships, we separate the data into three race groups: we estimated the underlying correlation matrices for African-American, Caucasian and Hispanic respectively, and also repeat the same procedure for the three races pooled together. Also, these analysis are done for overall COMPAS score categories and violent COMPAS score categories separately. Our estimated conditional independence graphs are in Figures 2.3 and 2.4. A first interesting finding is that we notice the graphical structures vary across African-American, Caucasian and Hispanic groups. In Figure 2.3 for the overall COMPAS score, note that the overall COMPAS score

has a direct effect on two-year recidivism for African-Americans but is conditionally independent for Caucasians and Hispanics, however has a quite indirect effect in the pooled model. Conversely, in Figure 2.4, the violent COMPAS score is conditionally independent of two-year recidivism for African-Americans but has a direct effect for Caucasians and an indirect effect for Hispanics and the pooled groups. It is also interesting how the various juvenile criminal history measures have different associations across the race groups and two COMPAS scores. There is common structure seen across all three races too, misdemeanor and number of priors are consistently connected to two-year recidivism for all races. In contrast, this in not the case for misdemeanor and number of priors and the two COMPAS scores. Also, the graphical models for pooled group are the same for both sets of variables involving score category and violent score category respectively.

2.6.2 Prostate cancer data analysis

This data set was first analyzed by [9] and subsequently by [25]. It consists of 12 mixed type measurements for 475 prostate cancer patients who were diagnosed as having either stage 3 or 4 prostate cancer. Among the 12 variables, 8 are continuous, 3 are ordinal and 1 is nominal (list of variables and corresponding abbreviations can be found in Table 2.1). More details of the data can be found at [4]. We are interested in how the 'Survival Status' is correlated with the other 11 variables after removing the nominal variable 'Electrocardiogram code' since it is not appropriate to infer latent variable for nominal variable. The 'Survival Status' is transformed into binary variable as either survived or died, regardless of causes of death. Also, we combined performance rating's level 2 and level 3 as one level since these patients are in bed more than 50% of daytime. The correlation/covariance

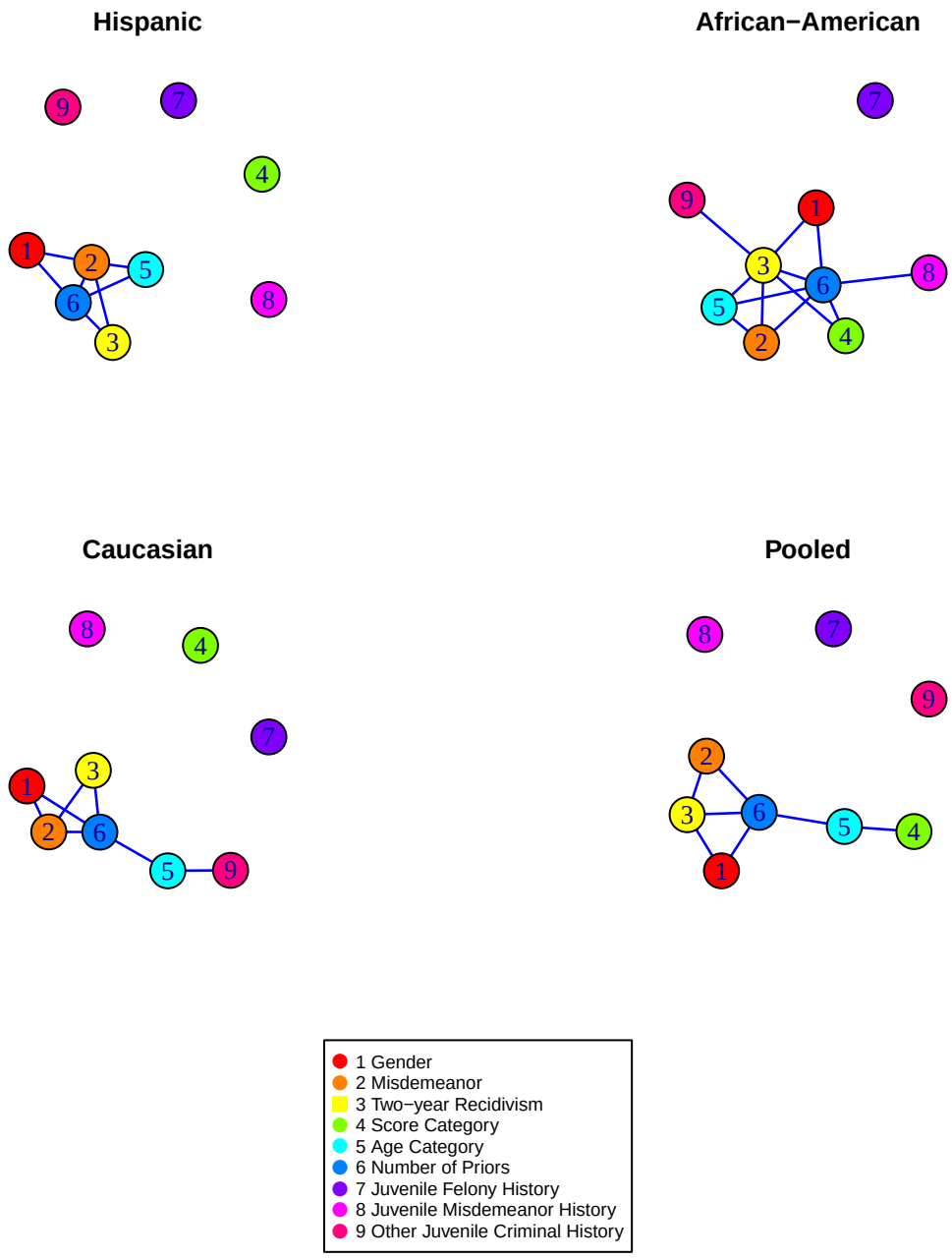

Figure 2.3: Mixed data graphical model for the three level ordinal overall COMPAS score data set by African-American, Caucasian, Hispanic and pooled groups.

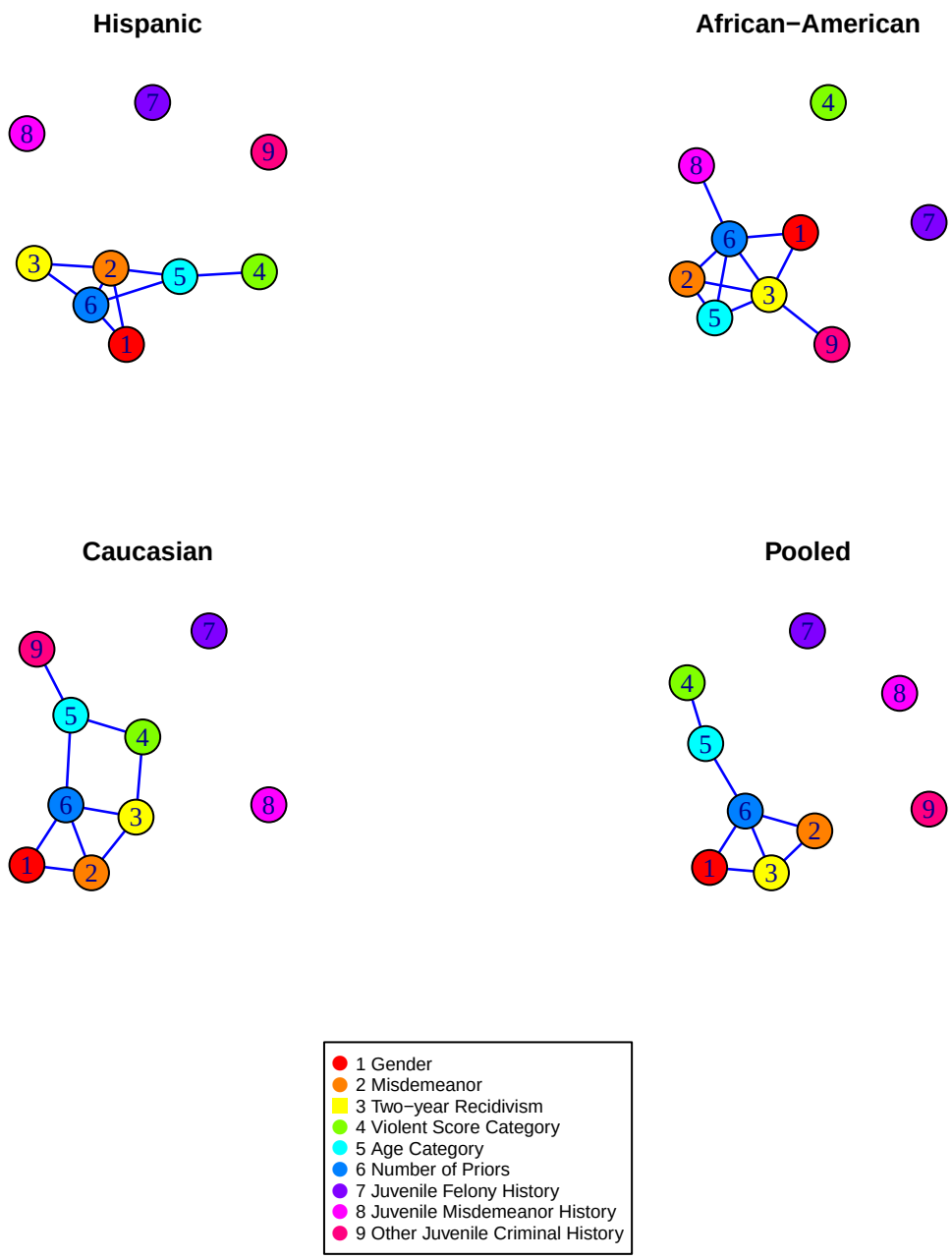

Figure 2.4: Mixed data graphical model for the three level ordinal violent COMPAS score data set by African-American, Caucasian, Hispanic and pooled groups.

matrices are given in Table 2.2 and 2.3 for Stage 3 and 4 patients respectively. Figure 2.5 illustrates the recovered graph for the 12 variables for Stage 3 and Stage 4 patients respectively. It is interesting that the set of nodes connected to 'Survival Status' are different among Stage 3 and 4 patients. For Stage 3 patients, the 'Survival Status' node is of degree 3, with neighbors including 'Cardiovascular disease history' (HX), 'Bone metastases' (BM), and 'Performance Rating' (PF). Whilst for Stage 4 patients, 'Survival Status' node is of degree 2 instead, with its neighbors being 'Performance rating' and 'Serum prostatic acid phosphatase'. It is interesting that 'Performance rating' (PF) is adjacent to 'Survival Status' in both networks, which is reasonable since an active patient (Performance rating = 0 or 1) was probably able to move around hence survived. However, PF was not included in the best model found by [25], which we speculate as a result of mistreating the categorical variable PF. Also, we notice that some variables are highly correlated with Surv but not a neighbor of Surv on the network graph, such as the 'Age' variable for Stage 3 patients, and Bone Metastases (BM) variable for Stage 4. It's easy to see the reason after a closer look at the correlation tables in Table 2.2 and 2.3: for Stage 3, the 'Age' variable has a higher correlation with 'PF' than with 'Surv', implying that the high correlation between 'Age' and 'Surv' might be a result of the high correlation between it and 'PF'. It is similar for Stage 4: 'BM' variable sees a higher correlation with 'PF' and 'AP', 'HG' has a higher correlation with 'PF' than with 'Surv', and 'SZ' finds itself highly correlated with 'AP' and 'PF', namely the high correlations between those variables and 'Surv' can be due to their high correlations with 'AP' and/or 'PF', thus they are indirectly connected to 'Surv' node in the network graph (Figure 2.5) but rather directly connected to the neighbors of 'Surv'. Another interesting structure can also be discovered from the network graph (Figure 2.5) that agrees with [25]: they found

that the cluster consisting of variables 'BM', 'Wt', 'HG', 'SBP' and 'DBP' gave the second best likelihood; on the other hand, we found that those 5 variables are consistently clustered for both Stage 3 and 4 patients, which agrees with the finding by [25]. One might also notice that 'Size of primary tumor' node is isolated for only Stage 3 patients' network. This in fact agrees with the definition of Stage: stage 3 represents local extension of the disease whilst stage 4 represents distant metastasis as evidenced by elevated acid phosphatase and/or X-ray evidence [25]. In other words, for Stage 3 patients, 'SZ' (node 10) is not necessarily a good indicator of 'Index of tumor stage, histolic grade' (node 11) or 'Serum prostatic acid phosphatase' (node 12), but it might be a good one for stage 4 patients as we can see in the graph that node 10 is connected to node 11 and 12. Another interesting finding is that 'Size of primary tumor' and 'Serum prostatic acid phosphatase' are adjacent in the networks for Stage 4 patients, which agrees with the results in [41] that Stage 4 patients on average saw larger tumors and higher levels of serum prostatic acid phosphatase.

Table 2.1: List of variables and their abbreviations

Covariate	Abbreviation	Number of levels (if categorical)
Cardiovascular.disease.history	HX	2
Bone.metastases	BM	2
SurvStat	Surv	2
Performance.rating	PF	3
Age	Age	
Weight	Wt	
Systolic.Blood.pressure	SBP	
Diastolic.blood.pressure	DBP	
Serum.haemoglobin	HG	
Size.of.primary.tumour	SZ	
Index.of.tumour.stage.and.histolic.grade	SG	
Serum.prostatic.acid.phosphatase	AP	

Table 2.2: Correlation/covariance matrix for Stage 3 patients

Variable	HX	BM	Surv	PF	Age	Wt	SBP	DBP	HG	SZ	SG	AP
HX	1.00	-1.00	0.48	0.39	0.27	-0.01	0.24	0.09	-0.09	-0.07	-0.17	-0.17
BM	-1.00	1.00	1.00	-1.00	-0.09	-0.14	-0.67	-0.03	-0.76	0.06	0.55	0.91
Surv	0.48	1.00	1.00	0.26	0.22	-0.15	0.07	0.05	-0.06	0.18	0.12	-0.05
PF	0.39	-1.00	0.26	1.00	0.34	-0.05	0.14	0.05	-0.04	-0.05	0.26	0.01
Age	0.27	-0.09	0.22	0.34	1.00	0.00	0.03	-0.11	-0.13	-0.07	-0.03	-0.01
Wt	-0.01	-0.14	-0.15	-0.05	0.00	1.00	0.23	0.18	0.17	0.06	0.06	0.12
SBP	0.24	-0.67	0.07	0.14	0.03	0.23	1.00	0.58	0.04	0.04	-0.03	-0.05
DBP	0.09	-0.03	0.05	0.05	-0.11	0.18	0.58	1.00	0.14	-0.05	-0.05	0.01
HG	-0.09	-0.76	-0.06	-0.04	-0.13	0.17	0.04	0.14	1.00	-0.06	0.07	0.17
SZ	-0.07	0.06	0.18	-0.05	-0.07	0.06	0.04	-0.05	-0.06	1.00	0.18	0.09
SG	-0.17	0.55	0.12	0.26	-0.03	0.06	-0.03	-0.05	0.07	0.18	1.00	0.10
AP	-0.17	0.91	-0.05	0.01	-0.01	0.12	-0.05	0.01	0.17	0.09	0.10	1.00

Table 2.3: Correlation/covariance matrix for Stage 4 patients

Variable	HX	BM	Surv	PF	Age	Wt	SBP	DBP	HG	SZ	SG	AP
HX	1.00	-0.07	0.15	0.16	0.16	0.12	-0.02	-0.10	0.06	-0.09	-0.05	0.01
BM	-0.07	1.00	0.33	0.50	-0.07	-0.29	-0.07	-0.12	-0.42	0.28	0.11	0.33
Surv	0.15	0.33	1.00	0.52	0.15	-0.21	0.07	-0.01	-0.30	0.22	0.15	0.28
PF	0.16	0.50	0.52	1.00	0.09	-0.42	0.10	-0.13	-0.65	0.24	0.09	0.30
Age	0.16	-0.07	0.15	0.09	1.00	-0.10	0.09	-0.10	-0.15	0.04	0.01	0.09
Wt	0.12	-0.29	-0.21	-0.42	-0.10	1.00	0.15	0.25	0.36	-0.04	-0.11	-0.15
SBP	-0.02	-0.07	0.07	0.10	0.09	0.15	1.00	0.57	0.11	0.12	-0.01	-0.00
DBP	-0.10	-0.12	-0.01	-0.13	-0.10	0.25	0.57	1.00	0.17	0.04	-0.03	-0.09
HG	0.06	-0.42	-0.30	-0.65	-0.15	0.36	0.11	0.17	1.00	-0.15	-0.09	-0.20
SZ	-0.09	0.28	0.22	0.24	0.04	-0.04	0.12	0.04	-0.15	1.00	0.23	0.34
SG	-0.05	0.11	0.15	0.09	0.01	-0.11	-0.01	-0.03	-0.09	0.23	1.00	0.15
AP	0.01	0.33	0.28	0.30	0.09	-0.15	-0.00	-0.09	-0.20	0.34	0.15	1.00

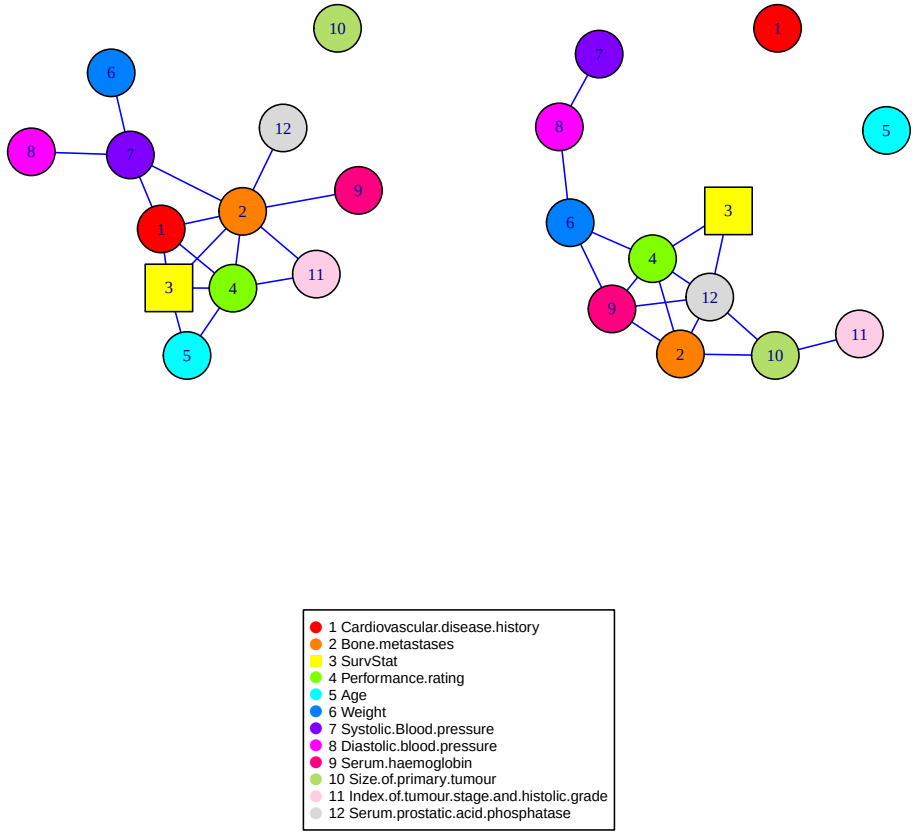

Figure 2.5: Prostate cancer data analysis, Plot of the connected components of the estimated graph for the prostate cancer data. Number 3 represents the 'Performance rating' variable. *Left*: Stage 3 patients, the 'Survival Status' node is of degree 4, with neighbors including 'Cardiovascular disease history', ' Bone metastases', 'Performance Rating' and 'Age'; *Right*: Stage 4 patients, 'Survival Status' node is of degree 2 instead, with its neighbors being 'Performance rating' and 'Serum prostatic acid phosphatase'.

2.7 Conclusion and Discussion

To sum up, we proposed a generalized rank-based method to estimate correlations for any p-level discrete-continuous mixed data. The method is under latent Gaussian copula model, assuming there is some latent variable that discretize the continuous data into categorical. There exists unique solution to the bridge function, which can be obtained easily by Newton's method. The theoretical properties of the estimates are well established. In our simulation studies, we see as p increases, the estimation becomes as accurate as the one using raw continuous data. This agrees with the intuition that as we obtain more information, the estimation will do a better job.

Correlation estimates for ternary-ternary data and binary-ternary data are also given, to help social science researches find associations among different types of data.

Also, we proposed a modified estimate based on Kendall's τ^b compared to the one based on Kendall's τ^a in [14], to account for occurrences of tied pairs. Since the Kendall's τ^b involves a square root term in its denominator, we did not compute its population version directly but rather obtained its 1st-order and 2nd-order Taylor approximations, which showed no visible difference from the Monte-Carlo simulated average.

Our method can further be applied to graph recovery by inverting the correlation matrix estimate (into the so-called Precision matrix). Conditional independence can also be inferred from the precision matrix. One practical advantage of our method is that it can estimate the correlations regardless of dimensions. For high-dimensional data, estimation can be done in parallel to reduce time expense.

In the next chapter, we will see how to apply the covariance matrix estimator to regression problem.

CHAPTER 3

HIGH DIMENSIONAL SEMIPARAMETRIC REGRESSION MODEL FOR MIXED DATA

3.1 Introduction

Regression analysis that investigates the relationship between a response variable of interest and a large set of predictors with mixed data is a common problem. Intensive statistical studies on mixed data have been conducted for the past decades, covering topics including clustering analysis [24, 23], covariance estimation and network analysis [14], and canonical correlation analysis [65]. However, high dimensional regression that investigates the relationship between a response variable and a large set of predictors with mixed data is a common problem but much less explored. Both conventional and modern regression methods do not fit well into this mixed data regression problem due to high-dimensionality and/or non-normality of the data. Ordinary least squares regression is not applicable for high dimensional data where the number of predictors is greater than the number of observations. The more advanced regularization methods for high-dimensional regression such as the lasso estimator [56] has been remarkably successful under the assumption of a linear relationship between the response and explanatory variables that jointly follow Gaussian distribution, but this assumption might not be realistic for real-world data, especially for mixed data. Several approaches have been proposed to address this issue, such as the nonparametric sparse additive models [50], the semi-parametric single index model [48], and more recently the latent Gaussian copula regression model by [10]. However, these approaches are tailored for continuous data but not mixed data.

To bridge the gap, we propose a general approach in this paper for high dimensional regression on mixed data that does not require the observed data to be normally distributed or linearly related. Our main contribution is the use of latent Gaussian copula model to uncover the latent linear relationship, which allows us to estimate the latent linear regression coefficients using a rank-based covariance matrix estimator without knowing the marginal transformation functions. In fact, the widely applied data transformations, such as log transformation are readily adapted by this model with no need to prespecify the transformation. Moreover, the estimator is ℓ_1 regularized, so variable selection is done automatically. Lastly, the estimator can then be used to predict a new response for a given value of the covariates.

The rest of this chapter is organized as follows: in Section 3.2, we formulate the problem in detail and review definitions related to latent Gaussian copula model. Section 3 presents our rank-based method for estimating the ℓ_1-penalized linear coefficient vector, $\boldsymbol{\beta}^{\mathrm{GC}}_{\mathrm{Lasso}}$, and illustrates how to make prediction of the response variable for a given value of covariates using the linear model coefficient estimate. Theoretical results are also established. In Section 4 we demonstrate numerical performance of the methodology on simulated data. In addition, Section 5 illustrates our approach on a real-world dataset from breast cancer patients.

3.2 Background

3.2.1 Notation

For the rest of this dissertation, we keep the following notations. For a d-dimensional vector $\mathbf{v} \in \mathbb{R}^d$, we denote v_j to be its j-th entry. In addition, its ℓ_q-norm is defined as $\|v\|_q = (\sum_{j=1}^d |v_j|^q)^{1/q}$. For a matrix $\mathbf{A} = (A_{jk}) \in \mathbb{R}^{d\times d}$, we denote its j-th column as $\mathbf{A}_j$. Let $\mathbf{A}[1:a, 1:b]$ be the submatrix of $\mathbf{A}$ with rows between 1 and a and columns between 1 and b. We define $\|A\|^1 = \sum_{1\le j,k\le d}|A_{jk}|$, and $\|A\|^2 = \sum_{1\le j,k\le d}(A_{jk})^2$. And $\Phi(\cdot)$ denotes the cumulative distribution function of a standard normal distribution.

3.2.2 Latent Gaussian copula model

Suppose we observe the $n \times p$ data matrix denoted as $\mathbf{X} = (X_1, \ldots, X_p)^T$, where n is the number of observations and p the number of predictors (each predictor can be ordinal or continuous). Let $\mathbf{Y} \in \mathbb{R}^n$ denote the responses for each observation. Under latent Gaussian copula model, the observed data $(\mathbf{X}, \mathbf{Y})$ is assumed to result from monotonically transforming some latent multivariate Gaussian data $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})$ (and one extra step of discretization for ordinal predictors). In particular, the marginal transformations are assumed to be monotonically increasing functions, and the latent variables satisfy $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) \sim N_{p+1}(\mathbf{0}, \mathbf{\Sigma})$ with $\mathrm{diag}(\mathbf{\Sigma}) = 1$ for identifiability. Then the latent multivariate Gaussian distribution of $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})$ implies the linear model

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \epsilon,$$

where $\epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Recall the least square solution is $\boldsymbol{\beta} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{Y}}$. Denoting $\text{cov}(\tilde{\mathbf{X}}, \tilde{\mathbf{X}}) = \boldsymbol{\Sigma}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}}$ and $\text{cov}(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) = \boldsymbol{\Sigma}_{\tilde{\mathbf{X}}\tilde{\mathbf{Y}}}$, the solution can be rewritten as $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}}\boldsymbol{\Sigma}_{\tilde{\mathbf{X}}\tilde{\mathbf{Y}}}$. Another consequence is $\sigma^2 = 1 - \boldsymbol{\Sigma}^T_{\tilde{\mathbf{X}}\tilde{\mathbf{Y}}}\boldsymbol{\Sigma}^{-1}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}}\boldsymbol{\Sigma}_{\tilde{\mathbf{X}}\tilde{\mathbf{Y}}}$.

Therefore, instead of assuming linear relationship between the observed $\mathbf{Y}$ and $\mathbf{X}$, we only assume that between some latent variables $\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{X}}$. We aim to estimate the latent linear coefficients $\boldsymbol{\beta}$ for high-dimensional setting, and use that estimate to make prediction of the observable response given a value of the covariates.

Our approach to estimation rests upon a rank-based covariance estimator under latent Gaussian copula model. We first introduce some notations and basic definitions to facilitate the approach development.

Formally, we say the observed random vector $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$ follows latent Gaussian copula model if there exist latent random vector $\tilde{\mathbf{Z}} = (\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) \sim N_{p+1}(\mathbf{0}, \boldsymbol{\Sigma})$ with $\text{diag}(\boldsymbol{\Sigma}) = 1$ and monotonically increasing functions $\boldsymbol{f} = (f_j)_{j=1}^{p+1}$ such that

(i) if Z_j is continuous, $Z_j = f_j(\tilde{Z}_j)$;

(ii) if Z_j is categorical taking values in $\{0, 1, \ldots, m\}$, we have a cut-off vector $\boldsymbol{\Delta}_j = (-\infty, \Delta_j^{(1)}, \ldots, \Delta_j^{(m)}, \infty)$ that discretizes the transformed variable: $Z_j = l$ if $f_j(\tilde{Z}_j) \in (\Delta_j^{(l)}, \Delta_j^{(l+1)}]$.

For the rest of the Chapter, we suppose the first p_1 dimensions of $\mathbf{X}$, that is, X_i for $i = 1, \ldots, p_1$, are categorical variables taking values $\{0, 1, 2\}$ for simplicity. Also, to avoid confusion with Chapter 2, by τ we refer to the general version of Kendall's tau, namely τ^a, for the following sections and next chapter as well.

3.2.3 Rank-based Correlation Matrix Estimator

We now review the rank-based estimator of covariance/correlation matrix $\mathbf{\Sigma}$ as in Chapter 2. Our estimation method is based on a "bridge function" that relates Kendall's tau, τ_{jk}, for each observed variable pair (Z_j, Z_k), $1 \leq j, k \leq p+1$, with the correlation, $\mathbf{\Sigma}_{jk}$, between the corresponding latent Gaussian pair $(\tilde{Z}_j, \tilde{Z}_k)$. The population version of τ_{jk} is given by

$$\tau_{jk} = \mathbb{E}\left[\text{sign}\{(Z_{ij} - Z_{i'j})(Z_{ik} - Z_{i'k})\}\right], \tag{3.1}$$

an unbiased estimate of which is given by the corresponding τ statistic

$$\hat{\tau}_{jk} = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < i' \leq n} \text{sign}[(Z_{ij} - Z_{i'j})(Z_{ik} - Z_{i'k})]. \tag{3.2}$$

Depending on the data type, different bridge functions are involved [46]:

(i) if both $\mathbf{Z}_j$ and $\mathbf{Z}_k$ are continuous, the bridge function is given by

$$\tau_{jk} = F(\Sigma_{jk}) = \frac{\pi}{2} \sin^{-1}(\Sigma_{jk})\,; \tag{3.3}$$

(ii) if both $\mathbf{Z}_j$ and $\mathbf{Z}_k$ are ternary, associated with cutoff vectors $\mathbf{\Delta}_j$ and $\mathbf{\Delta}_k$, the bridge function is given by

$$\begin{aligned}
\tau_{jk} &= F(\Sigma_{jk}; \Delta_j^{(1)}, \Delta_j^{(2)}, \Delta_k^{(1)}, \Delta_k^{(2)}) \\
&= 2\Phi_2(\Delta_j^{(2)}, \Delta_k^{(2)}, \Sigma_{jk})\Phi_2(-\Delta_j^{(1)}, -\Delta_k^{(1)}, \Sigma_{jk}) \\
&\quad - 2\left[\Phi(\Delta_j^{(2)}) - \Phi_2(\Delta_j^{(2)}, \Delta_k^{(1)}; \Sigma_{jk})\right]\left[\Phi(\Delta_k^{(2)}) - \Phi_2(\Delta_j^{(1)}, \Delta_k^{(2)}; \Sigma_{jk})\right];
\end{aligned} \tag{3.4}$$

(iii) if $\mathbf{Z}_j$ is ternary associated with cutoff vector $\mathbf{C}_j$, and $\mathbf{Z}_k$ is continuous, then the bridge function is given by

$$\begin{aligned}\tau_{jk} &= F(\Sigma_{jk}; \Delta_j^{(1)}, \Delta_j^{(2)}) \\ &= 4\Phi_2(\Delta_j^{(2)}, 0; \Sigma_{jk}/\sqrt{2}) - 2\Phi(\Delta_j^{(2)}) \\ &\quad + 2[\Phi_3(\Delta_j^{(1)}, \Delta_j^{(2)}, 0; \Sigma_{jk}) - \Phi_3(\Delta_j^{(2)}, \Delta_j^{(1)}, 0; \Sigma_{jk})], \end{aligned} \tag{3.5}$$

where $\Phi(t)$ denotes $P(z \le t)$ for $z \sim N(0,1)$, and $\Phi_2(t_1, t_2; r) = P(z_1 \le t_1, z_2 \le t_2)$ for $(z_1, z_2) \sim N_2\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & r\\ r & 1\end{bmatrix}\right)$, and $\Phi_3(t_1, t_2, t_3; r) = P(z_1 \le t_1, z_2 \le t_2, z_3 \le t_3)$ for $(z_1, z_2, z_3) \sim N_3\left(\begin{bmatrix}0\\0\\0\end{bmatrix}, \begin{bmatrix}1 & 0 & r/\sqrt{2}\\ 0 & 1 & -r/\sqrt{2}\\ r/\sqrt{2} & -r/\sqrt{2} & 1\end{bmatrix}\right)$.

Note that for case (ii) and (iii) the cut-offs are involved, but they can be easily estimated by their moment estimators

$$\hat{\Delta}_j^1 = \Phi^{-1}\left(\frac{\sum_i \mathbb{1}\{Z_{ij} = 0\}}{n}\right) \quad \text{and} \quad \hat{\Delta}_j^2 = \Phi^{-1}\left(1 - \frac{\sum_i \mathbb{1}\{Z_{ij} = 2\}}{n}\right).$$

It has been shown that all the bridge functions above are monotonic [46], therefore Σ_{jk} can be estimated as the unique solution to the equation of sample Kendall's tau and the bridge function: $\hat{\tau}_{jk} = F(\Sigma_{jk}; ...)$ for bridge function $F(\cdot)$ as in (3.3), (3.4) or (3.5) corresponding to the data types. Therefore we have the rank-based estimator of covariance/correlation matrix as

$$\hat{\mathbf{\Sigma}} = (\hat{\Sigma}_{jk})_{(p+1)\times(p+1)} \quad \text{with } \hat{\Sigma}_{jk} = F^{-1}(\hat{\tau}_{jk}; \ldots) \text{ for } j \ne k, \tag{3.6}$$

which has the block structure

$$\hat{\mathbf{\Sigma}} = \begin{pmatrix} \hat{\Sigma}_{\tilde{X}\tilde{X}} & \hat{\Sigma}_{\tilde{X}\tilde{Y}} \\ \hat{\Sigma}_{\tilde{X}\tilde{Y}}^T & \hat{\Sigma}_{\tilde{Y}\tilde{Y}} \end{pmatrix} \tag{3.7}$$

and $\text{diag}(\hat{\mathbf{\Sigma}}) = 1$.

This semiparametric rank-based estimator has three advantages: since the estimation is done pairwisely for the $(p+1)$ variables, it can handle high dimensionality well; also the estimator can be obtained without knowing the marginal transformations; moreover, due to the rank-based nature, the estimator is robust to extreme values that are not rarely seen in practice.

3.3 Methodology

3.3.1 Estimation of $\boldsymbol{\beta}$

We now introduce the approach to estimate $\boldsymbol{\beta}$ under latent Gaussian copula model. Recalling from Section 2 that we observe i.i.d. pairs $(\mathbf{X}_i^T, Y_i)$, $i = 1, \ldots, n$, that follows latent Gaussian copula model, where $Y_i \in \mathbb{R}$ and $\mathbf{X}_i \in \mathbb{R}^p$ with its first p_1 dimensions being categorical ordinal and the rest continuous. Then there exist some latent variables $(\tilde{\mathbf{X}}_i^T, \tilde{Y}_i) \sim N_{p+1}(\mathbf{0}, \mathbf{\Sigma})$ for $i = 1, \ldots, n$ that satisfy the linear relationship

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \epsilon,$$

where $\epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, $\sigma^2 = 1 - \mathbf{\Sigma}_{\tilde{\mathbf{X}}\tilde{\mathbf{Y}}}^T \mathbf{\Sigma}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}}^{-1} \mathbf{\Sigma}_{\tilde{\mathbf{X}}\tilde{\mathbf{Y}}}$.

The least square solution gives $\hat{\boldsymbol{\beta}} = \mathbf{\Sigma}_{\tilde{X}\tilde{X}}^{-1} \mathbf{\Sigma}_{\tilde{X}\tilde{Y}}$. In high-dimensional context where we have $n < p$, we solve the problem under ℓ_1-penalization with the objective function defined as:

$$\hat{\boldsymbol{\beta}}_{\text{Lasso}} = \underset{\beta \in \mathbb{R}^d}{\arg\min} \{ \frac{1}{2n} ||\tilde{\mathbf{Y}} - \tilde{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1 \}.$$

Notice that

$$||\tilde{\mathbf{Y}} - \tilde{X}\boldsymbol{\beta}||_2^2 = \boldsymbol{\beta}^T\tilde{X}^T\tilde{X}\boldsymbol{\beta} - 2\tilde{\mathbf{Y}}^T\tilde{X}\boldsymbol{\beta},$$

and it is natural to substitute $\tilde{X}^T\tilde{X}$ with $n\Sigma_{\tilde{X}\tilde{X}}$ and $\tilde{\mathbf{Y}}^T\tilde{X}$ with $n\Sigma_{\tilde{X}\tilde{Y}}$. We then can rewrite the objective function as

$$\hat{\boldsymbol{\beta}} = \underset{\beta\in\mathbb{R}^d}{\arg\min}\{\frac{1}{2}\Big(\boldsymbol{\beta}^T\Sigma_{\tilde{X}\tilde{X}}\boldsymbol{\beta} - 2\Sigma_{\tilde{X}\tilde{Y}}\boldsymbol{\beta}\Big) + \lambda||\boldsymbol{\beta}||_1\} \tag{3.8}$$

Therefore, even though $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})$ is not directly accessible, we are still be able to estimate β in (3.8) by replacing $\Sigma_{\tilde{X}\tilde{X}}$ with $\hat{\Sigma}_{\tilde{X}\tilde{X}}$ and $\Sigma_{\tilde{X}\tilde{Y}}$ with $\hat{\Sigma}_{\tilde{X}\tilde{Y}}$, which can be obtained following (3.6). We summarize this process in Algorithm 1 as follows.

Algorithm 1: Compute rank-based estimator of Σ and $\hat{\boldsymbol{\beta}}^{\mathrm{GC}}_{\mathrm{Lasso}}$

Data: Observed data matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ and response $\mathbf{Y} \in \mathbb{R}^n$, parameter $\lambda > 0$.

Result: Regularized estimator $\hat{\boldsymbol{\beta}}^{\mathrm{GC}}_{\mathrm{Lasso}}$.

for $j = 1$ *to* $p+1$ **do**
  for $k = j+1$ *to* $p+1$ **do**
    Find sample Kendall's tau $\hat{\tau}_{jk}$ according to equation (2.2);
    Compute the rank-based estimator of correlation $\hat{\Sigma}_{jk}$ as described in (3.6);
    $\hat{\Sigma}_{jk}$ is then the (j,k)-th and (k,j)-th entries in $\hat{\mathbf{\Sigma}}$;
  end
end
Extract the block components $\hat{\Sigma}_{\tilde{X}\tilde{X}}$ and $\hat{\Sigma}_{\tilde{X}\tilde{Y}}$ from $\hat{\Sigma}_{p\times p}$ as in (3.7);
Compute

$$\hat{\boldsymbol{\beta}}^{\mathrm{GC}}_{\mathrm{Lasso}} = \underset{\beta\in\mathbb{R}^p}{\arg\min}\{\frac{1}{2}\Big(\boldsymbol{\beta}^T\hat{\Sigma}_{\tilde{X}\tilde{X}}\boldsymbol{\beta} - 2\hat{\Sigma}_{\tilde{X}\tilde{\mathbf{Y}}}\boldsymbol{\beta}\Big) + \lambda||\boldsymbol{\beta}||_1\} \tag{3.9}$$

---

[10] established the concentration rate of $\hat{\boldsymbol{\beta}}^{\mathrm{GC}}_{\mathrm{Lasso}}$ based on the Restricted Strong

Convexity (RSC) condition which was first introduced by [44], for its relation to restricted eigenvalue condition [8] under high-dimensional linear regression context. Inspired by [10], we also establish the theoretical properties of $\hat{\boldsymbol{\beta}}^{\mathrm{GC}}_{\mathrm{Lasso}}$ based on RSC with definition as follows:

Definition 4 (Restricted Strong Convexity (RSC))**.** *For a given sparsity level $s \leq p$ and constant $\alpha \geq 1$, a matrix $\Sigma \in \mathbb{R}^{p\times p}$ satisfies the restricted strong convexity (RSC) condition with constants (γ_1, s, α) if*

$$\theta^T\Sigma\theta \geq \gamma_1||\theta||_2^2 \quad \textit{for all } \theta \in \{\theta \in \mathbb{R}^p : ||\theta_{S^c}||_1 \leq \alpha||\theta_S||_1, S \subset \{1,\ldots,p\}, |S| \leq s\}$$

The following theorem establishes concentration rate of $\hat{\boldsymbol{\beta}}$ under the assumption on Σ's RSC condition.

Theorem 2. *Assuming β has sparsity of s. Suppose that the condition number of Σ is bounded by M for some $M > 0$, namely $\kappa(\Sigma) \leq M$, and $\Sigma_{\tilde{X}\tilde{X}}$ satisfies the RSC with constants $(\gamma_1, s, 3)$. Let $\hat{\boldsymbol{\beta}}(\lambda)$ be defined as (3.9). If $s = o(\frac{n}{\log p})$, and $\lambda = C_1\sqrt{\frac{\log p}{n}}$ for some $C_1 > 2M$, then with probability at least $1 - 2p^{-1}$,*

$$||\hat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}||_2 \leq c_2\sqrt{\frac{s\log p}{n}} \quad \textit{and} \quad ||\hat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}||_1 \leq c_3 s\sqrt{\frac{\log p}{n}} \tag{3.10}$$

for some c_2, c_3 independent of n.

The above theorem implies the minimax rate optimality of $\hat{\boldsymbol{\beta}}(\lambda)$ under latent Gaussian copula regression model. Moreover, under further regularity conditions, $\hat{\boldsymbol{\beta}}(\lambda)$ is also shown to be sign consistent by [10].

3.3.2 Prediction

With $\boldsymbol{\beta}$ well estimated by $\hat{\boldsymbol{\beta}}^{\text{GC}}_{\text{Lasso}}$, we are able to predict y given $\mathbf{x}^\star = (x_1^\star, \ldots, x_p^\star)$. Under latent Gaussian copula model as described in Section 2.1, we have

$$\tilde{Y}^\star = \tilde{\mathbf{x}}^\star \boldsymbol{\beta} = \sum_{i=1}^{d} f_i(x_i^\star)\beta_i \quad \text{and} \quad \tilde{Y}^\star = f_Y(Y^\star)$$

so the optimal prediction of the response is

$$\mu^\star = f_Y^{-1}(\sum_{i=1}^{d} f_i(x_i^\star)\beta_i).$$

Notice that $\mu^\star$ is a function involving $\boldsymbol{f} = \{f_Y, f_1, \ldots, f_p\}$. Therefore, in order to estimate $\mu^\star$ based on the observed data $(Y_i, \mathbf{X}_i^T)$, $i = 1, \ldots, n$, we will need to approximate $\boldsymbol{f}$. We consider the following procedure that approximate $\boldsymbol{f}$ involving empirical Cumulative Distribution Functions (CDF) based on the observations.

Let F_Y be the CDF of Y, and $\hat{F}_Y$ be the empirical CDF of Y based on the observed $Y_1, \ldots, Y_n$. Likewise, let F_i denote the CDF of i-th dimension of $\mathbf{X}$,and $\hat{F}_j$ be the empirical CDF of the observed $\{X_{1,j}, \ldots, X_{n,j}\}$, for $j = 1, \ldots, p$. Since $f_j(x_{i,j}) \sim N(0,1)$ and $F_j(x_{i,j}) = \Phi(f_j(x_{i,j}))$, [10] proposed to approximate f_j by

$$\hat{f}_j(t) = \Phi^{-1}(\hat{F}_j(t)),$$

for $j = 1, \ldots, p$, which works well with continuous data. However, for mixed data, infinite values might arise leading to extreme predictions. As an alternative, we consider the Winsorized empirical CDF $\tilde{F}_j(\cdot)$

$$\tilde{F}_j(t) = \delta_n \cdot I_{\{\hat{F}_i(t) < \delta_n\}} + \hat{F}_i(t) \cdot I_{\{\delta_n \le \hat{F}_i(t) \le 1-\delta_n\}} + (1-\delta_n) \cdot I_{\{\hat{F}_i(t) > 1-\delta_n\}} \tag{3.11}$$

where the Winsorization parameter δ_n helps not only avoid infinite values but also achieve a better bias-variance tradeoff [15]. It is suggested by [38] to set $\delta_n = 1/n^2$

for good practical performance. Therefore, we approximate f_j by

$$\hat{f}_j(t) = \Phi^{-1}(\tilde{F}_j(t)),$$

for $j = 1, \ldots, p$; and likewise

$$\hat{f}_Y(t) = \Phi^{-1}(\tilde{F}_Y(t)).$$

And subsequently, we estimate $\mu^\star$ by

$$\hat{\mu}^\star = \hat{f}_Y^{-1}(\sum_{j=1}^{d} \hat{f}_j(x_j^\star)\hat{\beta}_j) \tag{3.12}$$

where $\hat{f}_Y^{-1}(t) = \inf\{x \in \mathbb{R} : \hat{f}_Y(x) \geq t\}$.

Algorithm 2: Prediction for y given a new set of predictors $\mathbf{x}^\star$ using $\hat{\boldsymbol{\beta}}^{\mathrm{GC}}_{\mathrm{Lasso}}$

Data: Observed data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ and response $\mathbf{Y} \in \mathbb{R}^n$, a new observed set of covariates $\mathbf{x}^\star$, coefficients estimate $\hat{\boldsymbol{\beta}}^{\mathrm{GC}}_{\mathrm{Lasso}}$, parameter $\lambda > 0$.

Result: $\hat{\mu}^\star$.

for $j = 1$ *to* p **do**

  Compute the approximated functions $\hat{f}_j(\cdot) = \Phi^{-1}(\tilde{F}_j(\cdot))$, where $F_j(\cdot)$ is the Winsorized empirical CDF of X_j as in Equation (3.11);

  Approximate $\tilde{x}_j^\star$ by $\tilde{x}_j^\star = \hat{f}_j(x_j^\star)$;

end

Compute $\widehat{\tilde{y}^\star} = \tilde{\mathbf{x}}^\star \hat{\boldsymbol{\beta}}^{\mathrm{GC}}_{\mathrm{Lasso}}$;

Approximate f_Y by $\hat{f}_Y(\cdot) = \Phi^{-1}(\tilde{F}_Y(\cdot))$;

Obtain the prediction

$$\hat{\mu}^\star = \hat{f}_Y^{-1}(\sum_{j=1}^{d} \hat{f}_j(x_j^\star)\hat{\beta}_j) \tag{3.13}$$

The theoretical result of $\hat{\mu}^\star$ is established as follows:

Theorem 3. *Suppose the conditions in Theorem 2 still hold. Also suppose that f_Y is Lipschitz-continuous with some positive constant c_0, also suppose $\tilde{Y^\star} = f_Y(\mu^\star) < M$ for some $M > 0$ and $F_i(x_i^\star)$ is bounded by some positive constant $\delta^\star \in (0,1)$ such that $F_i(x_i^\star) \in (\delta^\star, 1-\delta^\star)$ for $i \in \{1,\ldots,p\}$. If $s = o(\sqrt{\frac{n}{\log d}})$, then the predictor $\hat{\mu}^\star$ in (3.12) satisfies, with probability at least $1 - p^{-1} - n^{-1}$,*

$$|\hat{\mu}^\star - \mu^\star| \lesssim s\sqrt{\frac{\log d}{n}}.$$

3.4 Numerical studies

In this section, we demonstrate the numerical performance of the proposed estimator in (3.9) in six simulation scenarios. For comparison, we also consider the performance of the regular Lasso [56] and the elastic net [68] methods. All the simulations are based on the latent model that $\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \epsilon$ where $\epsilon \sim N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I})$.

Since the covariance matrix $\boldsymbol{\Sigma}$ for latent Gaussian copula model is a correlation matrix with all the diagonal entries 1's, we start with some correlation matrix $\Sigma_{\tilde{X}\tilde{X}}$, coefficient vector $\boldsymbol{\beta}$, and σ_ϵ of our own design choice, then obtain the full covariance/correlation matrix as $\boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{\tilde{X}\tilde{X}} & \Sigma_{\tilde{X}\tilde{X}}\beta/\sigma_{\tilde{Y}} \\ \beta^T\Sigma_{\tilde{X}\tilde{X}}/\sigma_{\tilde{Y}} & 1 \end{bmatrix}$ where $\sigma_{\tilde{Y}}^2 = \boldsymbol{\beta}^T\Sigma_{\tilde{X}\tilde{X}}\boldsymbol{\beta} + \sigma_\epsilon^2$. By doing so, we have a full matrix $\boldsymbol{\Sigma}$ with all 1's on its diagonal. In the following six simulation scenarios, we repeat the same procedure to construct such $\boldsymbol{\Sigma}$ for a latent Gaussian copula model, so that the latent variables $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, which will then be marginally transformed to observable values $(\boldsymbol{X}, \boldsymbol{Y})$ by functions of our choice. For each scenario, we simulate a training set for estimating β and a test set for prediction, and repeat this for 100 replicates.

The tuning parameters are all chosen by 5-folds cross-validation on the training data.

1. Inspired by the original lasso paper [56], we set $\Sigma_{\tilde{X}\tilde{X}}(i,j) = 0.5^{|i-j|}$, $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and $\sigma_\epsilon = 3$. Let $y = \tilde{y}$, $x = \tilde{x}^5$, then discretize the first 4 dimensions of x: namely trichotomize x_j by $\Delta_j^{(1)} = (\Phi^{-1}(0.5))^5, \Delta_j^{(2)} = (\Phi^{-1}(0.75))^5$ for $j \leq 4$. The rest 4 dimensions of x remain continuous. The training set is of size 20, and the test set consists of 180 observations.

2. Same as Simulation 1 except that $y = \exp(\tilde{y})$.

3. Each training set contains 40 observations and 11 predictors, and the test set is of size 40. Set $\Sigma_{\tilde{X}\tilde{X}}$ to be a toeplitz matrix such that $\Sigma_{\tilde{X}\tilde{X}}(i,j) = 1 - 0.1 \times |i-j|$. Let $\boldsymbol{\beta} = (2, 1, 0, 0, 0, 0, 0, 0, 0, -1, -2)^T$, and $\sigma_\epsilon = 3$. Let $y = \tilde{y}$, $x = \tilde{x}^5$, then trichotomize x_j by $\Delta_j^{(1)} = (\Phi^{-1}(0.5))^5, \Delta_j^{(2)} = (\Phi^{-1}(0.75))^5$ for $j \leq 6$.

4. Each training set contains 50 observations and 50 predictors, and the test set is of size 50. $\beta_i = 2$ for $i = 21, \ldots, 25, 46, \ldots, 50$ and $\beta_i = 0$ for all other i. $\Sigma_{\tilde{X}\tilde{X}}(i,j) = 0.5^{|i-j|}$ and $\sigma_\epsilon = 2$. Let $y = \exp(\tilde{y})$, and $x = 2\tilde{x}^5 + 1$ followed by trichotomization using cut-offs $\Delta_j^{(1)} = 2(\Phi^{-1}(0.5))^5 + 1, \Delta_j^{(2)} = 2(\Phi^{-1}(0.75))^5 + 1$ for $j = 1, \ldots, 5$.

5. Same as Simulation 4, except applying trichotomization to the first 15 dimensions of x (using the same cut-offs).

6. Each training set contains 100 observations and 100 predictors, and test set is of size 100. $\Sigma_{\tilde{X}\tilde{X}}(i,j) = 0.5^{|i-j|}$, $\beta_i = 1$ for $i = 1, \ldots, 5$ and $\beta_i = -1$ for $i = 51, \ldots, 55$, and $\beta_i = 0$ for all other i. Set $\sigma_\epsilon = 2$. Let $y = \exp(\tilde{y})$, and $x = 2\tilde{x}^5 + 1$ followed by trichotomization using cut-off $\Delta_j^{(1)} = 2(\Phi^{-1}(0.5))^5 + 1, \Delta_j^{(2)} = 2(\Phi^{-1}(0.75))^5 + 1$ for $j = 1, \ldots, 50$.

We compared the simulation results for the different methods using the metrics:

(i) the mean squared prediction error given by $\frac{1}{k}\sum_{i=1}^{k}(y_i - \hat{y}_i)^2$ where k denotes the size of test set;

(ii) the model selection accuracy measured by the following three metrics:

$$\text{Model Selection Error} = \frac{1}{d}\sum_{j=1}^{d} I(I(\beta_j = 0) \neq I(\hat{\beta}_j = 0)),$$

$$\text{True Positve Rate (TPR)} = \frac{1}{d}\sum_{j=1}^{d} I(\beta_j \neq 0) \cdot I(\hat{\beta}_j \neq 0),$$

$$\text{False Positve Rate (FPR)} = \frac{1}{d}\sum_{j=1}^{d} I(\beta_j \neq 0) \cdot I(\hat{\beta}_j = 0),$$

where $I(\cdot)$ is indicator function.

Here we don't consider the $\hat{\beta}$'s estimation error by the three methods because the latent Gaussian copula method estimates the coefficients in linear model between *latent* variables whilst regular Lasso and Elastic Net methods focus on that between *observed* variables. The performance of each method is summarized in Table 3.1,3.2, 3.3, 3.4 and 3.5. It can be clearly seen that compared to the other two methods, the latent Gaussian copula method exhibits smaller Mean Squared Error and model selection error across the suite of simulation scenarios. It also has at least similar or much better combinations of FPR and TPR.

We further investigate the performance of the estimators at different levels of discretization. Specifically, we vary the number of discrete dimensions in Simulation 6 for comparison. Each time, the first $p_1 = 10, \ldots, 50$ dimensions of X are discretized following the same procedure as described in Simulation 6. Figure 3.1 illustrates the performance of the three methods. It can be seen that as the number of discrete variables increases, the latent Gaussian copula Lasso estimator consistently outperforms regular Lasso and Elastic Net methods.

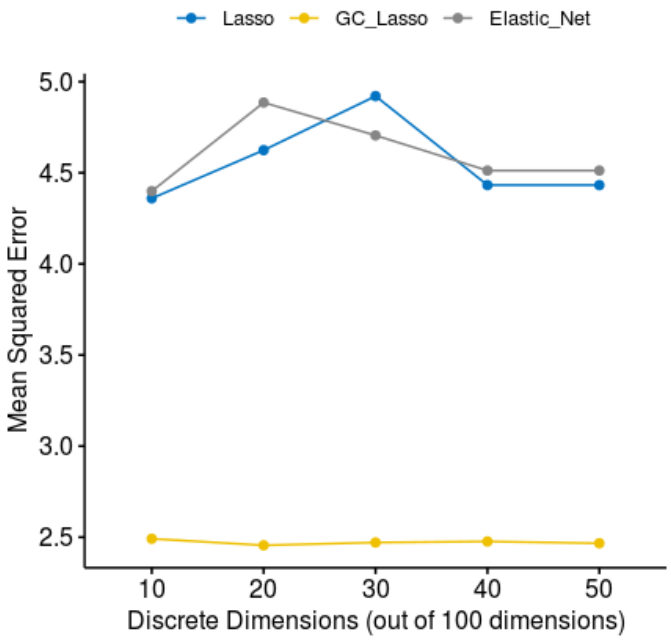

Figure 3.1: Mean squared error plot for different levels of discretization in Simulation 6. As number of discrete variables increases, the latent Gaussian copula Lasso estimator (yellow line) consistently outperforms regular Lasso (blue line) and elastic net (gray line) estimators.

Table 3.1: Mean squared error of predictions over 100 replicates for each simulation scenario. Standard errors are given in parentheses. Tuning parameters were chosen by cross-validation.

Simulation	Lasso	Elastic Net	GC-Lasso
1	4.69 (18.40)	4.07 (16.40)	0.85 (0.25)
2	16.30 (39.22)	9.37 (19.60)	4.51 (3.78)
3	5.22 (20.21)	3.73 (16.50)	0.55 (0.18)
4	10.10 (10.74)	5.59 (5.72)	2.09 (3.02)
5	17.40 (41.11)	5.32 (4.82)	1.71 (1.68)
6	4.92 (7.51)	4.70 (6.12)	2.47 (1.61)

Table 3.2: Model selection error summary over 100 replicates for each simulation scenario. Standard errors are given in parentheses.

Simulation	Lasso	Elastic Net	GC-Lasso
1	0.31 (0.17)	0.35 (0.17)	0.29 (0.15)
2	0.34 (0.18)	0.36 (0.18)	0.31 (0.17)
3	0.27 (0.19)	0.59 (0.09)	0.15 (0.08)
4	0.21 (0.10)	0.20 (0.07)	0.07 (0.04)
5	0.22 (0.13)	0.21 (0.12)	0.07 (0.04)
6	0.09 (0.04)	0.12 (0.08)	0.04 (0.02)

3.5 Case study

In this section, we illustrate our approach on a dataset collected from breast cancer patients. The data are publicly available at The Cancer Genome Atlas (TCGA) project database, a collection of genetic and clinical data from different high-throughput platforms.

The gene expression data are measures of RNAseq profiling which are of continuous values, whilst the clinical data contains both categorical variables such as "Pathologic stage", as well as continuous variables such as "Age". In this analysis, our goal is to study the relationship between survival time and genetic and clinical variables.

Specifically, we consider the set of gene expression data and clinical traits for

Table 3.3: True positive rates (TPR) summary over 100 replicates for each simulation scenario. Standard errors are given in parentheses.

Simulation	Lasso	Elastic Net	GC-Lasso
1	0.73 (0.23)	0.69 (0.27)	0.57 (0.33)
2	0.69 (0.26)	0.60 (0.31)	0.54 (0.31)
3	0.66 (0.23)	0.59 (0.27)	0.64 (0.16)
4	0.22 (0.13)	0.09 (0.13)	0.93 (0.08)
5	0.25 (0.15)	0.10 (0.13)	0.92 (0.08)
6	0.33 (0.20)	0.39 (0.17)	0.88 (0.10)

Table 3.4: False positive rates (FPR) summary over 100 replicates for each simulation scenario. Standard errors are given in parentheses.

Simulation	Lasso	Elastic Net	GC-Lasso
1	0.27 (0.18)	0.27 (0.22)	0.20 (0.24)
2	0.32 (0.27)	0.28 (0.25)	0.22 (0.26)
3	0.25 (0.29)	0.21 (0.21)	0.02 (0.06)
4	0.04 (0.04)	0.02 (0.03)	0.06 (0.04)
5	0.09 (0.12)	0.05 (0.09)	0.06 (0.04)
6	0.02 (0.02)	0.03 (0.04)	0.02 (0.02)

$n = 111$ deceased subjects. For computing stability and efficiency, we pre-screened 592 genes using the marginal screening method as suggested by [26]. We also excluded three single-valued variables from the study: "Vital status" (all deceased), "Gender" (all female), and "Ethnicity" (all 'not Hispanic or Latino'). In addition, the ordinal variables "Pathologic stage", "Tumor stage", "Lymph nodes status", and "Metastasis status" categorize each subject into a cancer stage indicating the extent of the cancer following the TNM staging system, and they are further divided to as many as 16 different sub-stages to provide more details to medical researchers. Here it might be necessary to group them into fewer categories for regression analysis given that we only have 111 subjects in the dataset. A common practice according to the National Cancer Institute is to group the stages into 3 categories representing 'carcinoma in situ' (abnormal cells are present but have not spread to nearby tissue), 'cancer is present', and 'cancer has spread to distant

Table 3.5: True positive rates (TPR) and False positive rates (FPR) summary over 100 replicates for each simulation scenario. FPR is in parentheses.

Simulation	Lasso	Elastic Net	GC-Lasso
1	0.73 (0.27)	0.69 (0.27)	0.57 (0.20)
2	0.69 (0.32)	0.60 (0.28)	0.54 (0.22)
3	0.66 (0.25)	0.59 (0.21)	0.64 (0.02)
4	0.22 (0.04)	0.09 (0.02)	0.93 (0.06)
5	0.25 (0.09)	0.10 (0.05)	0.92 (0.06)
6	0.33 (0.02)	0.39 (0.03)	0.88 (0.02)

parts of the body' respectively.

After the above data preprocessing steps, the data has 111 observations with complete data for 598 covariates of mixed data types, which contains 5 categorical variables: "Pathologic stage", "Tumor stage", "Lymph nodes status", "Metastasis status", and "Race"; and 593 continuous variables: "Age" and genes.

To investigate the performance of the proposed method, we randomly split the data 100 times: each time a training set of 100 observations is used for estimating $\hat{\beta}$, then we make predictions on a test set of 11 observations in order to evaluate the estimate's performance. In particular, we calculate the following metrics for prediction results: Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and the median of ℓ_2 errors between predicted and observed survival time (in years).

For comparison, we consider the regular Lasso estimator $\hat{\beta}_{\text{Lasso}}$ in addition to the proposed ℓ_1-regularized latent Gaussian copula estimator $\hat{\beta}^{\text{GC}}_{\text{Lasso}}$. For both methods, we chose the tuning parameter values via 5-folds cross-validation during training. The results are summarized in Table 3.6. It can be seen from Table 3.6 that in all three metrics, the latent Gaussian copula method outperforms regular Lasso, which is as expected due to the violation of normality and lack of linear

relationship.

Table 3.6: Cross-validation results for the survival time predictions on TCGA data

Metric	Lasso	GC-Lasso
RMSE	3.69 (1.22)	3.58 (1.17)
Median L_2 loss	6.60 (4.12)	5.51 (3.71)
MAPE	1.31 (0.601)	1.22 (0.577)
Number of variables selected	23 (18.4)	11 (3.1)

Applying the methods to the entire dataset, we obtained the results shown in Table 3.7. In particular, 15 variables are selected by the latent Gaussian copula method, including "Tumor stage", "Metastasis status", and genes including "CCL24", "ECHDC1", "EGR1", "FMO2", "FOSB", "GOLGA6L10", "GPR97", "GULP1", "MAPK10", "NR4A1", "SCN3A", "SFRS5" and "SH3BP2" that are all found to be related to breast cancer in biomedical literature, see [27], [19], [52], [54], [42], [34], [16], [45], [63], [66], [62], [1], [43]. It is noteworthy that all of the genes have highly skewed distributions, indicating a violation of normality assumption (see Figure 3.2). Existing models might require extra manual work such as a log transformation to assure the normality assumption, but the latent Gaussian copula model in this situation could handle it automatically.

Table 3.7: Summary of fitted survival time results in TCGA data

Metric	Lasso	GC-Lasso
RMSE	4.630	3.323
Median L_2 loss	7.106	3.553
MAPE	1.427	0.957
Number of variables selected	34	15

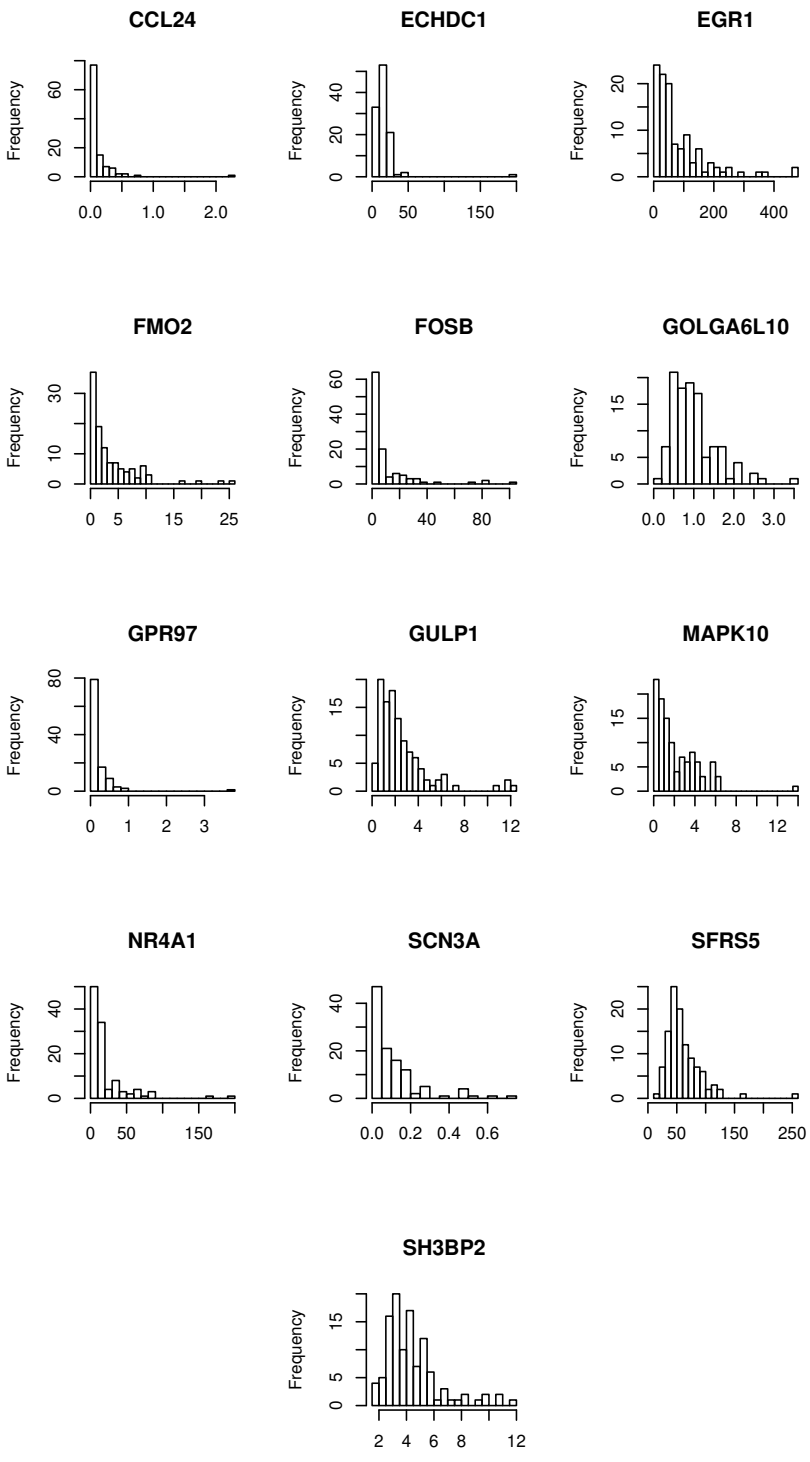

Figure 3.2: Histograms of selected genes by latent Gaussian copula model

3.6 Discussion

This chapter presents an approach for simultaneously estimating the latent linear model and performing variable selection without the unrealistic normality assumption on high-dimensional mixed data. Prediction for the response given a new set of covariates can also be made using the latent linear coefficients estimate. The proposed estimator and the corresponding predictions are found to outperform other existing methods in a simulation study and analysis of real data. Unlike other methods assuming normality, the proposed latent Gaussian copula regression model handles non-normal data very well with no need of data transformation, and it can be readily applied to high dimensional setting of mixed data because of pairwise covariance/correlation estimation plus regularized regression. One possible extension would be to allow the response to be categorical as well, namely a classification problem. It would be interesting to see if the covariance-based regression model can achieve satisfactory results in classfication tasks.

CHAPTER 4

COVARIANCE-REGULARIZED REGRESSION FOR HIGH-DIMENSIONAL MIXED DATA

4.1 Introduction

We now further extend the GC-Lasso method to a family of methods where the inverse covariance/correlation matrix is regularized, rather than the coefficients. Recall the least squares solution $\hat{\beta} = (\mathbf{\Sigma}_{\tilde{X}\tilde{X}})^{-1}\mathbf{\Sigma}_{\tilde{X}\tilde{Y}}$, and a zero entry in $(\mathbf{\Sigma}_{\tilde{X}\tilde{X}})^{-1}$ indicates conditional independence between the pair of variables, given all of the other covariates. In high-dimensional settings, the entries of $(\mathbf{\Sigma}_{\tilde{X}\tilde{X}})^{-1}$ can be noisy so some regularization is desired. One can impose a penalty on the matrix $(\mathbf{\Sigma}_{\tilde{X}\tilde{X}})^{-1}$ directly, for example the Ridge regression. More interestingly, [61] proposed the Scout Procedure that regularizes the inverse covariance matrix by maximizing the log likelihood of the data assuming normality. In many practices, sets of variables that truly are conditionally dependent will also be related to the response. One advantage of the Scout Procedure is its ability to distinguish between variables that truly are partially correlated with each other and those that in fact have zero partial correlation. This is particularly important, for example in the context of genetics research, that all the features/genes along the pathways related to the response can be found. However, the Scout Procedure also rests upon the two unrealistic assumptions discussed in the previous chapter. Therefore, we generalize the Scout Procedure to the GC-Scout under latent Gaussian copula model, where the benefit of distinguishing features is preserved while the assumptions are relaxed to a latent linear relationship and normality only after marginal transformations. The subsequent sections of this chapter is organized as follows: Section 2 presents

the GC-Scout framework for regression and classification, we also discuss some properties of certain members of the GC-Scout class; Section 3 investigates the performance of the GC-Scout methods on simulated data; followed by a case study on real data in Section 4; and we conclude with some discussion in Section 5.

4.2 Methodology

4.2.1 GC-Scout Procedure For Regression

Suppose we observe the $n \times p$ data matrix denoted as $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)^T$, where n is the number of observations and p the number of predictors (each predictor can be ordinal or continuous). Let $\mathbf{Y} \in \mathbb{R}^n$ denote the responses for each observation.

Under latent Gaussian copula model, the observed data $(\mathbf{X}, \mathbf{Y})$ is related to some multivariate Gaussian variables $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) \sim N_{p+1}(\mathbf{0}, \mathbf{\Sigma})$ with $\text{diag}(\mathbf{\Sigma}) = 1$ via marginal transformations (and discretization for ordinal predictors). Consequently, the log likelihood of the data is given by

$$\log(\det \mathbf{\Sigma}^{-1}) - \text{tr}(\mathbf{S}\mathbf{\Sigma}^{-1}) \tag{4.1}$$

where $\mathbf{S}$ is the sample covariance matrix of $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})$. But since we do not have direct access to the latent variables $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})$, we can replace $\mathbf{S}$ with the rank-based estimator $\hat{\mathbf{\Sigma}}$ as in (3.7) and (3.6), which has well established properties [46]. The linear regression

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \epsilon,$$

can be estimated via $\mathbf{\Sigma}^{-1}$ due to the fact that $\boldsymbol{\beta} = -\frac{\mathbf{\Theta}_{\tilde{\mathbf{X}}\tilde{\mathbf{Y}}}}{\mathbf{\Theta}_{\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}}}$ where $\mathbf{\Theta}$ denotes a

symmetric estimate of $\mathbf{\Sigma}^{-1}$ with the block structure $\mathbf{\Theta} = \begin{pmatrix} \mathbf{\Theta}_{\tilde{X}\tilde{X}} & \mathbf{\Theta}_{\tilde{X}\tilde{Y}} \\ \mathbf{\Theta}^T_{\tilde{X}\tilde{Y}} & \Theta_{\tilde{Y}\tilde{Y}} \end{pmatrix}$, see [39]. Despite that $\hat{\mathbf{\Sigma}}$ can readily handle high-dimensional data, in practice, some level of regularization has consistently been seen to be helpful for prediction [56] not only in high-dimensional setting but also in low-dimensions. Therefore, inspired by [61], one can consider the estimator of the inverse covariance/correlation matrix $\mathbf{\Sigma}^{-1}$ with ℓ_p penalty, denoted as $\mathbf{\Theta}$, such that

$$\hat{\mathbf{\Theta}} = \underset{\mathbf{\Theta}=\mathbf{\Theta}^T \succ 0;}{\arg\max} \Big\{ \log(\det \mathbf{\Theta}) - \operatorname{tr}(\hat{\mathbf{\Sigma}}\mathbf{\Theta}) - \|\mathbf{\Theta}\|^q \Big\}$$

where $\hat{\mathbf{\Sigma}}$ is the rank-based estimator as in (3.7) and (3.6). It is further suggested to solve the above problem via a two-stage maximization procedure where $\mathbf{\Theta}_{\tilde{X}\tilde{X}}$ and $\mathbf{\Theta}_{\tilde{X}\tilde{Y}}$ are regularized one-by-one, so that we will be able to first distinguish pairs of variables that truly are conditionally dependent given the other predictors (namely $\mathbf{X}_j, \mathbf{X}_k$ such that $(\mathbf{\Sigma}^{-1})_{\tilde{X}\tilde{X}}(j,k) \neq 0$) from those that are conditionally dependent due only to chance [61], and then distinguish the predictors that are conditionally dependent on the response given all other predictors from those that are not. Specifically, we propose the following Algorithm 3, the Scout procedure adapted for latent Gaussian Copula Model (GC-Scout Procedure).

The fundamental difference between *GC-Scout* and *Scout* Procedures is the choice of covarinace matrix estimator: unlike the *Scout* Procedure utlizing the empirical sample covariance matrix, the *GC-Scout* Procedure involves the rank-based estimator $\hat{\mathbf{\Sigma}}$ which brings two benefits: the restrictive normality assumption on observed variables is relaxed to latent normality, and the unrealistic assumption about a linear relationship between the response and observed covariates is no longer required. In fact, *GC-Scout* is a general extension of the *Scout* Procedure, since multivariate Gaussian is just a special case of latent Gaussian copula by simply setting $f_j(x) = x$ for all j.

Algorithm 3: The GC-Scout Procedure with ℓ_q penalties

Data: Observed data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ and response $\mathbf{Y} \in \mathbb{R}^n$, penalties q_1, q_2 and corresponding parameters $\lambda_1, \lambda_2 \geq 0$.

Result: Coefficients estimator $\hat{\boldsymbol{\beta}}$ for covariance-regularized regression.

1 Compute the rank-based correlation matrix estimator $\hat{\boldsymbol{\Sigma}}$ as in (3.7) and (3.6) ;

2 Compute $\hat{\boldsymbol{\Theta}}_{\tilde{X}\tilde{X}}$ that maximizes

$$\log(\det \boldsymbol{\Theta}_{\tilde{X}\tilde{X}}) - \mathrm{tr}(\hat{\boldsymbol{\Sigma}}_{\tilde{X}\tilde{X}}\boldsymbol{\Theta}_{\tilde{X}\tilde{X}}) - \lambda_1 \|\boldsymbol{\Theta}_{\tilde{X}\tilde{X}}\|^{q_1}; \tag{4.2}$$

3 Compute $\hat{\boldsymbol{\Theta}}$ that maximizes

$$\log(\det \boldsymbol{\Theta}) - \mathrm{tr}(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Theta}) - \lambda_2 \|\boldsymbol{\Theta}\|^{q_2} \tag{4.3}$$

subject to $\hat{\boldsymbol{\Theta}}[1:p, 1:p] = \hat{\boldsymbol{\Theta}}_{\tilde{X}\tilde{X}}$. Essentially the penalty is imposed upon $\hat{\boldsymbol{\Theta}}_{\tilde{X}\tilde{Y}}$ and $\hat{\Theta}_{\tilde{Y}\tilde{Y}}$ only;

4 Compute $\hat{\boldsymbol{\beta}}$, defined by

$$\hat{\boldsymbol{\beta}} = -\frac{\hat{\boldsymbol{\Theta}}_{\tilde{\mathbf{X}}\tilde{\mathbf{Y}}}}{\hat{\Theta}_{\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}}}. \tag{4.4}$$

Another minor difference is that the extra step of scaling in the *Scout* Procedure in which the solution is scaled by a factor of c obtained from fitting the simple linear regression $\tilde{\mathbf{Y}} = c(\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}})$ is no longer necessary. This is because the scaling factor is readily absorbed by the monotonic increasing marginal transformations $\mathbf{f}$ (more technical explanations can be found in proof to Claim 4).

Following the original Scout work in [61], we refer to the penalized log likelihoods in (4.2) and (4.3) as the first and second Scout criteria. For the rest of the chapter, we denote the solution to the above *GC-Scout* Procedure in Algorithm 3 as *GC-Scout*(q_1, q_2). If $\lambda_1 = 0$, then this will be indicated by *GC-Scout*$(\cdot, q_2)$, and if $\lambda_2 = 0$ then *GC-Scout*$(q_1, \cdot)$.

4.2.2 Maximization and Properties of GC-Scout

If $\lambda_1 = 0$ and $q_2 = 1$, then the solution to *GC-Scout*$(\cdot, 1)$ is just equivalent to the proposed estimator in Chapter 3:

$$\hat{\boldsymbol{\beta}}^{\text{GC}}_{\text{Lasso}} = \arg\min_{\beta \in \mathbb{R}^p} \{ \frac{1}{2} \Big(\boldsymbol{\beta}^T \hat{\Sigma}_{\tilde{X}\tilde{X}} \boldsymbol{\beta} - 2 \hat{\Sigma}_{\tilde{X}\tilde{\mathbf{Y}}} \boldsymbol{\beta} \Big) + \lambda \|\boldsymbol{\beta}\|^1 \}.$$

However, one interesting fact is that *GC-Scout*$(\cdot, 1)$ has no variable grouping effect. In other words, since *GC-Scout*$(\cdot, 1)$ only has penalty on $\hat{\Sigma}_{\tilde{X}\tilde{Y}}$, it will only select those variables that are partially independent given the response and other predictors. Therefore the solution $\hat{\boldsymbol{\beta}}^{\text{GC}}_{\text{Lasso}}$ is often much more sparse than the solution to *GC-Scout*$(p_1, 1)$ (see more numerical evidence in Section 4).

If $\lambda_1 > 0$ and $q_1 = 1$, then the maximization of the first GC-Scout criterion can be implemented via the famous Graphical Lasso algorithm [6, 18]. Improvments to the Graphical Lasso algorithm have also been recently studied, see for example [60], [40].

In the case that $\lambda_1 > 0$ and $q_1 = 2$, it turns out there exists a closed-form solution to *GC-Scout*$(2, q_2)$ as shown in the following Theorem 4, which is adopted from [61] for GC-Scout procedure.

Theorem 4. *For $q_1 = 2, \lambda_1 > 0$, the solution to GC-Scout Procedure Step 2 in Equation (4.2) is given by*

$$\hat{\boldsymbol{\Theta}}_{\tilde{X}\tilde{X}} = \mathbf{V}(\bar{\mathbf{D}})\mathbf{V}^{\mathbf{T}} \tag{4.5}$$

and the associated inverse of $\hat{\boldsymbol{\Theta}}_{\tilde{X}\tilde{X}}$ is

$$\hat{\boldsymbol{\Theta}}^{-1}_{\tilde{X}\tilde{X}} = \mathbf{V}(\mathbf{D} + \tilde{\mathbf{D}})\mathbf{V}^{\mathbf{T}}\,, \tag{4.6}$$

where $\mathbf{V}, \mathbf{D}$ are matrices of eigen-vectors and eigen-values of $\hat{\boldsymbol{\Sigma}}_{\tilde{X}\tilde{X}}$ respectively, and $\bar{\mathbf{D}}$ is a $p \times p$ diagonal matrix with i-th diagonal entry equal to

$2\Big(D_{ii} + \sqrt{D_{ii}^2 + 8\lambda_1}\Big)^{-1}$, *and* $\tilde{\mathbf{D}}$ *is a* $p \times p$ *diagonal matrix with* i*-th diagonal entry equal to* $\frac{1}{2}(-D_{ii} + \sqrt{D_{ii}^2 + 8\lambda_1})$.

The proof can be found in Appendix. With this property, one can compute the solution to Step2 in *GC-Scout*$(2, q_2)$ with time complexity as an eigen-value problem. This is remarkably computationally efficient compared to *GC-Scout*$(1, q_2)$, which is typically solved by popular methods such as Graphical Lasso.

It is trivial that if $\lambda_2 = 0$, then the solution to *GC-Scout*$(q_1, \cdot)$ is given by $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Theta}}_{\tilde{X}\tilde{X}}\hat{\Sigma}_{\tilde{X}\tilde{Y}}$ where $\hat{\boldsymbol{\Theta}}_{\tilde{X}\tilde{X}}$ is given in (4.5). In the case that $\lambda_2 > 0$ and $q_2 = 1$ or $q_2 = 2$, then we could maximize the second GC-Scout criterion via an L_{p_2} regression, due to the following result:

Theorem 5. *For* $q_2 \in \{1, 2\}, \lambda_2 > 0$*, the solution to GC-Scout Procedure Step 4 in Equation (4.4) is equal to the solution to the following:*

$$\hat{\boldsymbol{\beta}} = \underset{\beta \in \mathbb{R}^p}{\arg\min}\{\frac{1}{2}\Big(\boldsymbol{\beta}^T\hat{\boldsymbol{\Theta}}^{-1}_{\tilde{X}\tilde{X}}\boldsymbol{\beta} - 2\hat{\Sigma}_{\tilde{X}\tilde{Y}}\boldsymbol{\beta}\Big) + \lambda_2\|\boldsymbol{\beta}\|^{p_2}\} \tag{4.7}$$

where $\hat{\boldsymbol{\Theta}}^{-1}_{\tilde{X}\tilde{X}}$ *is the inverse of the solution to maximizing (4.3), Step 3 of the GC-Scout Procedure.*

The proof can be readily adopted from the Proof to Claim 2 in [61] by replacing S with $\hat{\boldsymbol{\Sigma}}$ and everything else remains the same.

Therefore, combining Theorem 4 and 5, one can easily see the following result:

Corollary 2. *If* $\lambda_1 > 0$ *and* $q_1 = 2$*,* $\lambda_2 > 0$ *and* $q_2 \in \{1, 2\}$*, then the solution to GC-Scout*(q_1, q_2) *is equal to the solution to the following:*

$$\hat{\boldsymbol{\beta}} = \underset{\beta \in \mathbb{R}^p}{\arg\min}\{\frac{1}{2}\Big(\boldsymbol{\beta}^T\hat{\boldsymbol{\Theta}}^{-1}_{\tilde{X}\tilde{X}}\boldsymbol{\beta} - 2\hat{\Sigma}_{\tilde{X}\tilde{Y}}\boldsymbol{\beta}\Big) + \lambda_2\|\boldsymbol{\beta}\|^{p_2}\} \tag{4.8}$$

where $\hat{\boldsymbol{\Theta}}^{-1}_{\tilde{X}\tilde{X}}$ *is the solution to maximizing (4.3), Step 3 of the GC-Scout Procedure, as given in (4.6).*

It is also noteworthy that if $\lambda_1 = 0$, then the case $q_2 = 2$ is the latent Gaussian Copula analogous to the ridge regression.

4.2.3 Prediction

In this section, we present the framework to predict the (observable) response for a given new set of covariates using the coefficients estimator obtained from *GC-Scout*(q_1, q_2) Procedure.

Let $\hat{\boldsymbol{\beta}}$ denote the coefficients estimator from the *GC-Scout*(q_1, q_2) Procedure with data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ and response $\mathbf{Y} \in \mathbb{R}^n$. Now suppose we observe a new set of predictors x^* and want to predict y^*. If we do have direct access to the marginal transformations $\mathbf{f}$, then this is easy: $\hat{y^*} = f_y^{-1}(\sum_{j=1}^p f_j(X_j)\hat{\beta}_j)$. Therefore, in order to make the prediction, we could approximate $\mathbf{f}$. One reasonable approximation proposed by [15] and [10] is $\hat{f}_j(t) = \Phi^{-1}(\hat{F}_j(t))$ where $\hat{F}_j(t)$ is the empirical Cumulative Distribution Function of X_j. Recall that $f_j(x_j) \sim N(0,1)$ for all j, then the Cumulative Distribution Function of X_j follows that $F_j(x_j) = \Phi(f_j(x_j))$, hence the estimate $\hat{f}_j(t) = \Phi^{-1}(\hat{F}_j(t))$. For stability reason, we consider the Winsorized empirical CDF:

$$\tilde{F}_j(t) = \delta_n \cdot I_{\{\hat{F}_i(t) < \delta_n\}} + \hat{F}_i(t) \cdot I_{\{\delta_n \leq \hat{F}_i(t) \leq 1-\delta_n\}} + (1-\delta_n) \cdot I_{\{\hat{F}_i(t) > 1-\delta_n\}}. \tag{4.9}$$

Following [38] and [14], we set $\delta_n = 1/n^2$. Therefore, we have the following Algorithm 4 for prediction.

[10] established the tight error bound of $\hat{y^\star}$ to be $s\sqrt{\frac{\log p}{n}}$ under certain conditions, where s is the sparsity level of the true $\boldsymbol{\beta}$ (number of zero entries).

Algorithm 4: Prediction for y given a new set of predictors $\mathbf{x}^\star$ using $\hat{\boldsymbol{\beta}}^{\mathrm{GC}}_{\mathrm{Scout}}$

Data: Prediction for y given a new set of predictors $\mathbf{x}^\star$ using $\hat{\boldsymbol{\beta}}^{\mathrm{GC}}_{\mathrm{Scout}}$
Result: $\hat{\mu}^\star$.

for $j = 1$ *to* p **do**
  Compute the approximated functions $\hat{f}_j(\cdot) = \Phi^{-1}(\tilde{F}_j(\cdot))$, where $F_j(\cdot)$ is the Winsorized empirical CDF of X_j as in Equation (4.9);
  Approximate $\tilde{x}^\star_j$ by $\tilde{x}^\star_j = \hat{f}_j(x^\star_j)$
end
Compute $\widehat{\tilde{y}^\star} = \tilde{\mathbf{x}}^\star \hat{\boldsymbol{\beta}}^{\mathrm{GC}}_{\mathrm{Scout}}$;
Approximate f_Y by $\hat{f}_Y(\cdot) = \Phi^{-1}(\tilde{F}_Y(\cdot))$;
Obtain the prediction

$$\hat{y^\star} = \hat{f}_Y^{-1}(\sum_{j=1}^{d} \hat{f}_j(x^\star_j)\hat{\beta}_j) \tag{4.10}$$

4.2.4 GC-Scout Procedure for Classification

We further extend the *GC-Scout*(q_1, q_2) Procedure to classification problems. In the case of $n > p$, one can consider linear discriminant analysis (LDA) for classification problem. When it comes to high dimensional setting, we need to regularize the within-class covariance matrix. [21] proposed a method to shrink the within-class covariance matrix by adding a multiple of the identity matrix to the empirical covariance matrix. On the other hand, [61] proposed to shrunk the within-class inverse covariance matrix with an ℓ_q penalty instead, by maximizing the log likelihood of the observed data assuming normality. Here, we are interested in the classification problem where the normality assumption on the observed data does not hold.

Suppose the response variable Y has K distinct classes, and we observe a sample of size n where each observation $X^{(i)} \in \mathbb{R}^p$ belongs to some class $k \in$

$\{1, \ldots, K\}$. Given the training set, we want to accurately classify observations in an independent test set.

For each class k, let $\hat{\mu}_k$ denote the mean vector of those observations in class k. We define the within-class correlation matrix as $\mathbf{\Sigma}_{\text{wc}} = \frac{1}{K}\sum_{k=1}^{K}\hat{\mathbf{\Sigma}}(\{X^{(i)} : Y^{(i)} = k\})$, where $\hat{\mathbf{\Sigma}}(\{X^{(i)} : Y^{(i)} = k\})$ denotes the correlation matrix estimate for the subset of data belonging to class k. Then the *GC-Scout*$(q_1, \cdot)$ Procedure for classification is as follows:

Algorithm 5: GC-Scout Procedure for classification

Data: Observed training data $\mathbf{X} \in \mathbb{R}^{n\times p}$ and associated classes $\mathbf{Y} \in \mathbb{R}^n$, penalty norm q and corresponding parameters $\lambda \geq 0$. Independent test data $\mathbf{X}_{\text{test}} \in \mathbb{R}^p$.

Result: $\hat{y}_{\text{test}}$.

1 Compute the winthin-class correlation matrix estimator $\hat{\mathbf{\Sigma}}_{\text{wc}} = \frac{1}{K}\sum_{k=1}^{K}\hat{\mathbf{\Sigma}}(\{X^{(i)} : Y^{(i)} = k\})$ where $\hat{\mathbf{\Sigma}}(\{X^{(i)} : Y^{(i)} = k\})$ denotes the correlation matrix estimate for the subset of data belonging to class k;

2 Compute the regularized within-class inverse covariance matrix $\hat{\mathbf{\Sigma}}_{\text{wc}}^{-1}$ such that

$$\hat{\mathbf{\Sigma}}_{\text{wc}}^{-1} = \underset{\mathbf{\Sigma}^{-1}}{\arg\max}\Big\{\log\det\mathbf{\Sigma}^{-1} - \text{tr}(\hat{\mathbf{\Sigma}}_{\text{wc}}\mathbf{\Sigma}^{-1}) - \lambda\|\mathbf{\Sigma}^{-1}\|^q\Big\} \tag{4.11}$$

3 Classify test set observation $\mathbf{X}_{\text{test}}$ to class k' if k' satisfies

$$k' = \arg\max_k\Big\{\mathbf{X}_{\text{test}}^T\hat{\mathbf{\Sigma}}_{\text{wc}}^{-1}\hat{\mu}_k - \frac{1}{2}\hat{\mu}_k^T\hat{\mathbf{\Sigma}}_{\text{wc}}^{-1}\hat{\mu}_k + \log\pi_k\Big\} \tag{4.12}$$

where π_k is the count of class k in the training set.

4.3 Simulation studies

To examine the performance of the proposed estimators in the *GC-Scout*(q_1, q_2) family, we study the results from eight simulation scenarios in this section. Besides the GC-Scout estimators, we also consider the regular Lasso [56] and the elastic

net [68] methods for comparison. All the simulations follows the latent model that $\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \epsilon$ where $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, and $\epsilon \sim N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I})$.

For each of the eight simulation scenarios, we construct the Correlation matrix $\boldsymbol{\Sigma}$ based upon $\Sigma_{\tilde{X}\tilde{X}} \in \mathbb{R}^{p\times p}$, $\boldsymbol{\beta} \in \mathbb{R}^p$, and $\sigma_\epsilon \in \mathbb{R}$ via the following formula

$$\boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{\tilde{X}\tilde{X}} & \Sigma_{\tilde{X}\tilde{X}}\beta/\sigma_{\tilde{Y}} \\ \beta^T\Sigma_{\tilde{X}\tilde{X}}/\sigma_{\tilde{Y}} & 1 \end{bmatrix}$$

where $\sigma_{\tilde{Y}}^2 = \boldsymbol{\beta}^T\Sigma_{\tilde{X}\tilde{X}}\boldsymbol{\beta} + \sigma_\epsilon^2$. Using the correlation matrix, we generate n i.i.d. samples, $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}) \in \mathbb{R}^{n\times(p+1)}$, from $N(\mathbf{0}, \boldsymbol{\Sigma})$, and marginally transform $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})$ to obtain the 'observable' samples $(\boldsymbol{X}, \boldsymbol{Y}) = (f_1(\tilde{X}_1), \ldots, f_p(\tilde{X}_p), f_y(\tilde{\mathbf{Y}}))$. For each scenario, we repeatedly simulate a training set to estimate β and a test set for prediction for 100 replicates. The tuning parameters are all chosen by 5-folds cross-validation on the training set. The simulations scenarios are described in details as follows, where the first two are inspired by the original Lasso paper [56], the third, fourth, and sixth are of the same design as in Chapter 3, and the fifth is based on the original Scout paper [61]:

1. Set $\Sigma_{\tilde{X}\tilde{X}}(i,j) = 0.5^{|i-j|}$, $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and $\sigma_\epsilon = 3$. Let $\mathbf{Y} = \tilde{\mathbf{Y}}$, $\mathbf{X} = \tilde{\mathbf{X}}^5$, then discretize the first 4 dimensions of x: namely trichotomize x_j by $\Delta_j^{(1)} = (\Phi^{-1}(0.5))^5, \Delta_j^{(2)} = (\Phi^{-1}(0.75))^5$ for $j \leq 4$, the rest 4 dimensions of x remain continuous. Training set is of size 20, and test set of 180 observations.

2. Same as Simulation 1 except that $\mathbf{Y} = \exp(\tilde{\mathbf{Y}})$.

3. Training set contains $n = 40$ observations and $p = 11$ predictors, and test set is of size 40. Set $\Sigma_{\tilde{X}\tilde{X}}$ to be a toeplitz matrix such that $\Sigma_{\tilde{X}\tilde{X}}(i,j) = 1 - 0.1 \times |i-j|$. Let $\boldsymbol{\beta} = (2, 1, 0, 0, 0, 0, 0, 0, 0, -1, -2)^T$, and $\sigma_\epsilon = 3$. Let $\mathbf{Y} = \tilde{\mathbf{Y}}$,

$\mathbf{X} = \tilde{\mathbf{X}}^5$, then trichotomize X_j by $\Delta_j^{(1)} = (\Phi^{-1}(0.5))^5, \Delta_j^{(2)} = (\Phi^{-1}(0.75))^5$ for $j \leq 6$.

4. Each training set contains $n = 50$ observations and $p = 50$ predictors, and test set is of size 50. $\beta_j = 2$ for $j = 21, \ldots, 25, 46, \ldots, 50$ and $\beta_j = 0$ for all other j. $\Sigma_{\tilde{X}\tilde{X}}(i,j) = 0.5^{|i-j|}$ and $\sigma_\epsilon = 2$. Let $\mathbf{Y} = \exp(\tilde{\mathbf{Y}})$, and $\mathbf{X} = 2\tilde{\mathbf{X}}^5 + 1$ followed by trichotomization using cut-off $\Delta_j^{(1)} = 2(\Phi^{-1}(0.5))^5 + 1, \Delta_j^{(2)} = 2(\Phi^{-1}(0.75))^5 + 1$ for $j = 1, \ldots, 5$.

5. As in Simulation 1, except $\boldsymbol{\beta} = (3, 1.5, 0, 0, 0, -1, -1)^T$.

6. Each training set contains 100 observations and 100 predictors, and test set is of size 100. $\Sigma_{\tilde{X}\tilde{X}}(i,j) = 0.5^{|i-j|}$, $\beta_j = 1$ for $j = 1, \ldots, 5$ and $\beta_j = -1$ for $j = 51, \ldots, 55$, and $\beta_j = 0$ for all other j. Set $\sigma_\epsilon = 2$. Let $\mathbf{Y} = \exp(\tilde{\mathbf{Y}})$, and $\mathbf{X} = 2\tilde{\mathbf{X}}^5 + 1$ followed by trichotomization using cut-off $\Delta_j^{(1)} = 2(\Phi^{-1}(0.5))^5 + 1, \Delta_j^{(2)} = 2(\Phi^{-1}(0.75))^5 + 1$ for $j = 1, \ldots, 30$.

These scenarios are designed to cover a variety of situations: Simulations 1,2,3, and 5 are in low dimensional settings with sparse β. Simulations 4 and 6 are in high dimensional setting. Simulations 1,2,4,5 and 6 have a sparse inverse covariance matrix, and Simulations 5 and 6 have β's non-zero entries with alternating signs. Also notice that in Simulations 1,3, and 5, $\mathbf{Y}$ is still Gaussian but the performance of Scout methods turns out to be much worse than the GC-Scout estimators, as can be seen in Table 3.1.

To compare the performance of various methods, we consider the following metrics:

(i) the mean squared prediction error given by $\frac{1}{k}\sum_{i=1}^{k}(y_i - \hat{y}_i)^2$ where k denotes the size of test set;

(ii) the model selection accuracy measured by the following three metrics:

$$\text{model selection error} = \frac{1}{d}\sum_{j=1}^{d} I(I(\beta_j = 0) \neq I(\hat{\beta}_j = 0)),$$

$$\text{True Positve Rate (TPR)} = \frac{1}{d}\sum_{j=1}^{d} I(\beta_j \neq 0) \cdot I(\hat{\beta}_j \neq 0),$$

$$\text{False Positve Rate (FPR)} = \frac{1}{d}\sum_{j=1}^{d} I(\beta_j \neq 0) \cdot I(\hat{\beta}_j = 0),$$

where $I(\cdot)$ is indicator function.

Here we don't consider the $\hat{\beta}$'s estimation error by the three methods because the latent Gaussian copula method estimates the coefficients in linear model between *latent* variables whilst regular Lasso and Elastic Net methods focus on that between *observed* variables. The performance of each method is summarized in Table 4.1,4.2, 4.3, 4.4 and 4.5. It can be clearly seen that compared to the other two methods, the latent Gaussian copula method shows smaller Mean Squared Error and model selection error across the suite of simulation scenarios, and has at least similar or much better combinations of FPR and TPR. From Table 4.5, it is interesting to notice that in the low dimensional settings (Simulation 1,2,3, and 5). *Scout* methods are doing similar to or even better than *GC-Scout*$(\cdot, 1)$, but worse than it in high-dimensional setting (Simulation 4 and 6). However, both *GC-Scout*$(1, 1)$ and *GC-Scout*$(2, 1)$ outperform other methods consistently in all simulation settings.

We further investigate the performance of the estimators at different levels of discretization. Specifically, keeping everything else the same as in Simulation scenario 6, we vary the number of discrete dimensions for comparison. We experiment with five different discretization levels: $10, 20, 30, 40$ and 50 dimensions of X are discretized following the same procedure as described in Simulation 6. We obtain Figure 4.1 to visualize the performance for comparison.

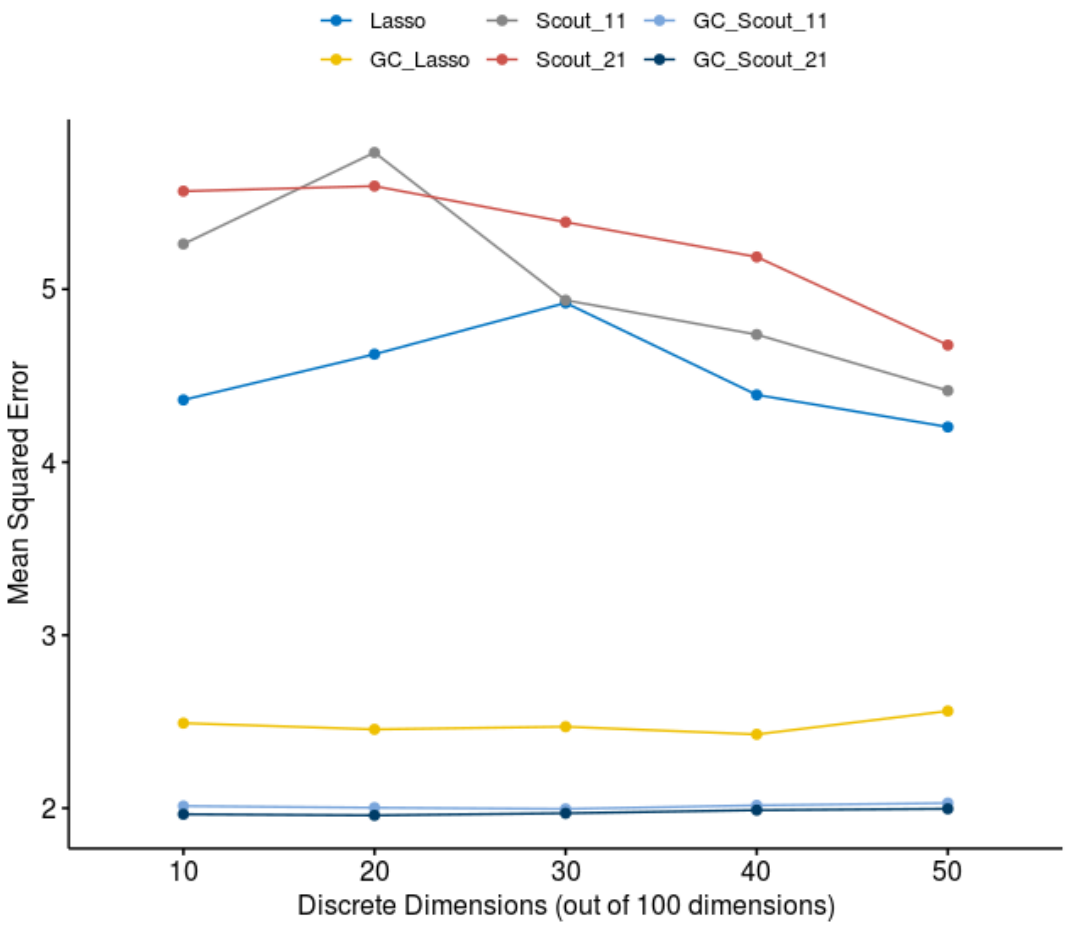

Figure 4.1: Mean squared error plot for different levels of discretization in Simulation 6. As number of discrete variables increases, the latent Gaussian copula estimators (three lines on the bottom: GC-Lasso (yellow), GC-Scout(1,1) (light blue), and GC-Scout(2,1) (dark blue)) consistently outperforms regular normality-assumed methods (three lines on top: Scout(2,1) (red), Scout(1,1) (gray) and regular lasso (blue)) .

Table 4.1: Mean squared error of predictions over 100 replicates for each simulation scenario. Standard errors are given in parentheses. Tuning parameters were chosen by cross-validation.

Simulation	GC-Lasso	Scout(1,1)	Scout(2,1)	GC-Scout(1, 1)	GC-Scout(2, 1)
1	0.85 (0.25)	1.04 (1.62)	1.04 (1.37)	0.60 (0.01)	0.57 (0.08)
2	4.51 (3.78)	5.00 (4.73)	4.71 (3.76)	4.10 (2.84)	3.50 (2.12)
3	0.55 (0.18)	11.30 (30.40)	10.70 (30.90)	4.06 (9.86)	2.30 (2.09)
4	2.09 (3.02)	10.37 (11.27)	8.83 (9.73)	3.20 (5.65)	2.21 (4.88)
5	7.92 (12.52)	10.56 (46.83)	10.73 (47.33)	3.77 (2.49)	3.93 (2.57)
6	2.47 (1.61)	4.94 (5.49)	5.39 (6.72)	2 (1.21)	1.97 (1.32)

Table 4.2: Variable Selection Error over 100 replicates for each simulation scenario. Standard errors are given in parentheses. Tuning parameters were chosen by cross-validation.

Simulation	GC-Lasso	Scout(1,1)	Scout(2,1)	GC-Scout(1, 1)	GC-Scout(2, 1)
1	0.29 (0.15)	0.32 (0.19)	0.30 (0.20)	0.31 (0.14)	0.32 (0.14)
2	0.31 (0.17)	0.34 (0.18)	0.33 (0.18)	0.3 0 (0.13)	0.26 (0.13)
3	0.15 (0.08)	0.36 (0.19)	0.36 (0.20)	0.29 (0.10)	0.27 (0.11)
4	0.07 (0.04)	0.28 (0.21)	0.25 (0.18)	0.32 (0.06)	0.25 (0.14)
5	0.27 (0.11)	0.34 (0.12)	0.35 (0.12)	0.26 (0.16)	0.21 (0.14)
6	0.04 (0.02)	0.22 (0.25)	0.16 (0.18)	0.12 (0.11)	0.11 (0.07)

4.4 Case study

In this section, we demonstrate two applications to real world data: one for regression and the other for classification.

4.4.1 Case study with TCGA data

For comparison reason, we study the same dataset as in Chapter 3 which consists of genetic and clinical data from 111 breast cancer patients. The data is publicly available at The Cancer Genome Atlas (TCGA) database.

The gene expression data are measures of RNAseq profiling which are continu-

Table 4.3: True positive rates (TPR) summary over 100 replicates for each simulation scenario. Standard errors are given in parentheses.

Simulation	GC-Lasso	Scout(1,1)	Scout(2,1)	GC-Scout$(1,1)$	GC-Scout$(2,1)$
1	0.57 (0.33)	0.81 (0.19)	0.79 (0.20)	0.97 (0.10)	1.00 (0.03)
2	0.54 (0.31)	0.79 (0.18)	0.77 (0.18)	0.95 (0.12)	0.97 (0.09)
3	0.64 (0.16)	0.69 (0.24)	0.69 (0.23)	0.98 (0.06)	1.00 (0.02)
4	0.93 (0.0817)	0.21 (0.13)	0.22 (0.13)	0.98 (0.05)	0.99 (0.03)
5	0.49 (0.18)	0.62 (0.18)	0.59 (0.20)	0.75 (0.20)	0.74 (0.19)
6	0.88 (0.10)	0.42 (0.18)	0.41 (0.18)	0.74 (0.37)	0.94 (0.2)

Table 4.4: False positive rates (FPR) summary over 100 replicates for each simulation scenario. Standard errors are given in parentheses.

Simulation	GC-Lasso	Scout(1,1)	Scout(2,1)	GC-Scout$(1,1)$	GC-Scout$(2,1)$
1	0.20 (0.24)	0.3 (0.22)	0.26 (0.21)	0.45 (0.23)	0.45 (0.22)
2	0.22 (0.26)	0.34 (0.27)	0.32 (0.27)	0.43 (0.22)	0.37 (0.20)
3	0.023 (0.06)	0.35 (0.30)	0.36 (0.31)	0.41 (0.17)	0.38 (0.17)
4	0.06 (0.04)	0.06 (0.07)	0.06 (0.07)	0.35 (0.08)	0.24 (0.15)
5	0.04 (0.09)	0.3 (0.27)	0.27 (0.28)	0.27 (0.22)	0.16 (0.17)
6	0.02 (0.02)	0.06 (0.08)	0.07 (0.11)	0.09 (0.08)	0.08 (0.05)

ous values, whilst clinical data contains both categorical variables such as "Pathologic stage", as well as continuous variables such as "Age". In this analysis, our goal is to study the relationship between survival time and genetic and clinical variables.

Upon completion of data preprocessing steps such as missing data removal, the data has 111 observations with complete data for 598 covariates of mixed data types, which contains 5 categorical variables: "Pathologic stage", "Tumor stage", "Lymph nodes status", "Metastasis status", and "Race"; and 593 continuous variables: "Age" and genes. Note that for computing stability and efficiency, we pre-screened 592 genes using marginal screening method as suggested by [26].

To study the performance of the proposed method and compare it to *GC-Lasso* (equivalently *GC-Scout*$(\cdot, 1)$) as proposed in Chapter 3, we conduct cross-validation

Table 4.5: True positive rates (TPR) and False positive rates (FPR) summary over 100 replicates for each simulation scenario. FPR is in parentheses. In low dimensional setting, (Simulation 1,2,3, and 5) *Scout* methods are doing similar to or even better than *GC-Scout*$(\cdot, 1)$, but worse than it in high-dimensional setting (Simulation 4 and 6). However, both *GC-Scout*$(1, 1)$ and *GC-Scout*$(2, 1)$ outperform other methods consistently in all simulation settings.

Simulation	GC-Lasso	Scout(1,1)	Scout(2,1)	GC-Scout$(1, 1)$	GC-Scout$(2, 1)$
1	0.57 (0.20)	0.81 (0.30)	0.79 (0.26)	0.97 (0.45)	1.00 (0.45)
2	0.54 (0.22)	0.79 (0.34)	0.77 (0.32)	0.95 (0.43)	0.97 (0.37)
3	0.64 (0.02)	0.69 (0.35)	0.69 (0.36)	0.98 (0.41)	1.00 (0.38)
4	0.93 (0.06)	0.21 (0.06)	0.22 (0.06)	0.98 (0.35)	0.99 (0.24)
5	0.49 (0.04)	0.62 (0.30)	0.59 (0.27)	0.75 (0.27)	0.74 (0.16)
6	0.88 (0.02)	0.42 (0.06)	0.41 (0.07)	0.74 (0.09)	0.94 (0.08)

analysis by randomly splitting the entire data 100 times: we put 11 observations in the test set, and leave the rest 100 observations as the training set. After obtaining the coefficients estimator using the training set, we then use it to make predictions for test set. To evaluate the estimators' performances, we consider the following metrics for prediction results: Mean Absolute Percentage Error (MAPE), Root Mean Squared Error(RMSE), and the median of ℓ_2 errors between predicted and observed survival time (in years). The results are summarized in Table 4.6. For both methods, we chose the tuning parameter values via 5-folds cross-validation during training.

Table 4.6: Cross-validation results for the survival time predictions on TCGA data

Metric	*GC-Lasso*	*GC-Scout*$(2, 1)$
RMSE	3.58 (1.17)	1.60 (1.25)
Median L_2 loss	5.51 (3.71)	0.68 (1.24)
MAPE	1.22 (0.58)	0.58 (0.22)
Number of variables selected	11 (3.1)	40 (65.9)

Applying the methods to the entire dataset, we obtained the results shown in Table 4.7. It can be seen that *GC-Scout*$(2, 1)$ achieves much superior prediction results than *GC-Scout*$(\cdot, 1)$. Recall the *GC-Scout*$(2, 1)$ method has variable grouping

effects, it turns out to select a much larger set of variables, as a result of seeking all possible genes/variables that are part of the pathway related to the disease.

Table 4.7: Summary of fitted survival time results in TCGA data. The *GC-Lasso* results are excerpted from Chapter 3.

Metric	*GC-Lasso*	*GC-Scout*$(2,1)$
RMSE	3.323	1.851
Median L_2 loss	3.553	0.383
MAPE	0.957	0.354
Number of variables selected	15	296

4.4.2 Classification case study on Ramaswamy data

This case study is to evaluate the performance of the GC-Scout Procedure for classification. In the original Scout paper, [61] studied the Ramaswamy microarray data set, which was initially described in detail by [49], and is publicly available at `https://www.jstatsoft.org/article/view/v033i01`. For comparison reason, we apply the GC-Scout method to the same data set. The Ramaswamy data consists of a training set of 144 tumor samples and a test set of 54 tumor samples, both contain 16,063 features (measurements for tumor genes expressions, in continuous values). The data span 14 different common tumor types. Our goal is to use the within-class covariance/correlation matrix estimator obtained from the training set to predict/classify the labels in test set. The GC-Scout Procedure for classification is described in Algorithm 5.

The results are summarized in Table 4.8. Three methods are compared here: the support vector machine (SVM) with one-versus-all classification from the original Ramaswamy paper [49], the *Scout*$(2,\cdot)$ from the original Scout paper [61], and the proposed *GC-Scout*$(2,\cdot)$. It can be seen from Table 4.8 that using as little as

20 % of the genes that were involved in training SVM and $Scout(2,\cdot)$, the proposed $GC\text{-}Scout(2,\cdot)$ can achieve the same results on the test set. We also want to point out that the training data for SVM and $Scout(2,\cdot)$ methods have been cube-rooted [61], but the proposed $GC\text{-}Scout(2,\cdot)$ is run on the original data, because the GC-Scout Procedure does not require the normality assumption. The genes are selected in increasing order of their F-statistics [58].

Table 4.8: Classification results for the cancer types in Ramaswamy data. The results in the first two columns are excerpted from [61] where the data has been cube-rooted. But the results for $GC\text{-}Scout(2,\cdot)$ in the last column are obtained from original data, using only 750 genes.

	SVM	$Scout(2,\cdot)$	$GC\text{-}Scout(2,\cdot)$
Test Error	11	7	7
Number of Genes Used	4000	4000	750

4.5 Discussion

In this chapter, we consider the regression problem with high-dimensional mixed data. The main contribution is that we extend the existing method *GC-Lasso* to an entire family of covariance-regularized methods, *GC-Scout*. The shrunken covariance matrix estimate dramatically improves the prediction accuracy, in terms of Mean Squared Error (MSE) and variable selection errors. This improvement is particularly obvious when the normality among observed data is violated, which is often the case for real world data. We also want to emphasize another benefit of the GC-Scout procedure that data transformation will be automatically done by the model with no need to prespecify the transformation function.

One thing to notice is members of the GC-Scout family, for example $GC\text{-}Scout(2,1)$, might suffer from relatively higher False Positive Rates, namely it

might wrongly identify those zero coefficients to be non-zero. This is a result due to the variable grouping effect. Recall that *GC-Scout*$(2, 1)$ aims to seek all features that are on the pathways that are related to the response. In other words, the superior prediction might come at the cost of noisy variable selection. Therefore, depending on the needs of different studies, we recommend researchers to select the most appropriate regression method.

APPENDIX A

CHAPTER 2 OF APEENDIX

A.1 Proof of Lemma 1

Proof. It is equivalent to show $\frac{\partial F_a(r;\Delta_j^1,\Delta_j^2,\Delta_k^1,\Delta_k^2)}{\partial r} > 0$.

Note that

$$\begin{aligned}
&\frac{\partial F(r;\Delta_j^1,\Delta_j^2,\Delta_k^1,\Delta_k^2)}{\partial r} \\
=\ &\frac{\partial[2\Phi_2(\Delta_j^2,\Delta_k^2,r)\Phi_2(-\Delta_j^1,-\Delta_k^1,r)]}{\partial r} \\
&+\frac{\partial\{2[\Phi(\Delta_j^2)-\Phi_2(\Delta_j^2,\Delta_k^1,r)][\Phi_2(\Delta_j^1,\Delta_k^2,r)-\Phi(\Delta_k^2)]\}}{\partial r}
\end{aligned}$$

We look at these two parts individually. Let ∂_{11} denote $\frac{\partial\Phi_2(\Delta_j^1,\Delta_k^1,r)}{\partial r}$, ∂_{12} denote $\frac{\partial\Phi_2(\Delta_j^1,\Delta_k^2,r)}{\partial r}$, and similar notations defined for ∂_{21}, ∂_{22}. Recall that $\frac{\partial\Phi_2(\cdot,\cdot,r)}{\partial r} > 0$ [14],

then we have

$$\begin{aligned}
&\frac{\partial[\Phi_2(\Delta_j^2,\Delta_k^2,r)\Phi_2(-\Delta_j^1,-\Delta_k^1,r)]}{\partial r}\\
&=\frac{\partial\Big[\Phi_2(\Delta_j^2,\Delta_k^2,r)-\Phi_2(\Delta_j^2,\Delta_k^2,r)\Big(\Phi(\Delta_k^1)+\Phi(\Delta_j^1)\Big)+\Phi_2(\Delta_j^1,\Delta_k^1,r)\Phi_2(\Delta_j^2,\Delta_k^2,r)\Big]}{\partial r}\\
&=\partial_{22}-\Phi(\Delta_k^1)\partial_{22}-\Phi(\Delta_j^1)\partial_{22}+\partial_{11}\Phi_2(\Delta_j^2,\Delta_k^2,r)+\partial_{22}\Phi_2(\Delta_j^1,\Delta_k^1,r)\\
&=\big[1-\Phi(\Delta_j^1)-\Phi(\Delta_k^1)+\Phi_2(\Delta_j^1,\Delta_k^1,r)\big]\partial_{22}+\partial_{11}\Phi_2(\Delta_j^2,\Delta_k^2,r)\\
&=\Phi_2(-\Delta_j^1,-\Delta_k^1,r)\partial_{22}+\partial_{11}\Phi_2(\Delta_j^2,\Delta_k^2,r)\\
&>0
\end{aligned}$$

$$\begin{aligned}
&\frac{\partial[\Big(\Phi(\Delta_j^2)-\Phi_2(\Delta_j^2,\Delta_k^1,r)\Big)\Big(\Phi_2(\Delta_j^1,\Delta_k^2,r)-\Phi(\Delta_k^2)\Big)]}{\partial r}\\
&=\Big[\Phi(\Delta_k^2)-\Phi_2(\Delta_j^1,\Delta_k^2,r)\Big]\partial_{21}-\Big[\Phi_2(\Delta_j^2,\Delta_k^1,r)-\Phi(\Delta_j^2)\Big]\partial_{12}\\
&=\mathbb{P}(X_{ij}\geq 1;X_{ik}\leq 1)\partial_{21}+\mathbb{P}(X_{ij}\leq 1;X_{ik}\geq 1)\partial_{12}\\
&>0.
\end{aligned}$$

□

A.2 Proof of Lemma 2

Proof. We know that

$$\begin{aligned}
\operatorname{sign}(X_{ij}-X_{i'j})=&\mathbb{1}\{X_{ij}=2\}-\mathbb{1}\{X_{i'j}=2\}\\
&+\mathbb{1}\{X_{ij}=1,X_{i'j}=0\}-\mathbb{1}\{X_{ij}=0,X_{i'j}=1\}
\end{aligned}$$

thus it is true that

$$
\begin{aligned}
&E[\text{sign}(X_{ij} - X_{i'j})(X_{ik} - X_{i'k})] \\
&= E\Big[\mathbb{1}\{X_{ij} = 2\}\text{sign}(X_{ik} - X_{i'k})\Big] - E\Big[\mathbb{1}\{X_{i'j} = 2\}\text{sign}(X_{ik} - X_{i'k})\Big] \\
&\quad + E\Big[\mathbb{1}\{X_{ij} = 1, X_{i'j} = 0\}\text{sign}(X_{ik} - X_{i'k})\Big] \\
&\quad - E\Big[\mathbb{1}\{X_{ij} = 0, X_{i'j} = 1\}\text{sign}(X_{ik} - X_{i'k})\Big]
\end{aligned}.
$$

We consider the four terms as two parts separately. The first two terms can be further computed as

$$
\begin{aligned}
&E\Big[\mathbb{1}\{X_{ij} = 2\}\text{sign}(X_{ik} - X_{i'k})\Big] - E\Big[\mathbb{1}\{X_{i'j} = 2\}\text{sign}(X_{ik} - X_{i'k})\Big] \\
&= E\Big[\mathbb{1}\{U_{ij} > \Delta_j^2\}\text{sign}(X_{ik} - X_{i'k})\Big] - E\Big[\mathbb{1}\{U_{i'j} > \Delta_j^2\}\text{sign}(X_{ik} - X_{i'k})\Big] \\
&= 2E\Big[\mathbb{1}\{U_{ij} > \Delta_j^2, V_{ik} - V_{i'k} > 0\}\Big] - 2E\Big[\mathbb{1}\{U_{i'j} > \Delta_j^2, V_{ik} - V_{i'k} > 0\}\Big] \\
&= 2\Phi_2(\Delta_j^2, 0, \sigma_{jk}/\sqrt{2}) - 2\Phi_2(\Delta_j^2, 0, -\sigma_{jk}/\sqrt{2}) \\
&= 4\Phi_2(\Delta_j^2, 0, \sigma_{jk}/\sqrt{2}) - 2\Phi(\Delta_j^2).
\end{aligned}
$$

The last two terms hold the following equivalence:

$$
\begin{aligned}
&E\Big[\mathbb{1}\{X_{ij} = 1, X_{i'j} = 0\}\text{sign}(X_{ik} - X_{i'k})\Big] - E\Big[\mathbb{1}\{X_{ij} = 0, X_{i'j} = 1\}\text{sign}(X_{ik} - X_{i'k})\Big] \\
&= 2E\Big[\mathbb{1}\{U_{ij} \in [\Delta_j^1, \Delta_j^2], U_{i'j} < \Delta_j^1, V_{ik} - V_{i'k} > 0\}\Big] \\
&\quad - 2E\Big[\mathbb{1}\{U_{ij} < \Delta_j^1, U_{i'j} \in [\Delta_j^1, \Delta_j^2], V_{ik} - V_{i'k} > 0\}\Big] \\
&= 2[\Phi_3(\Delta_j^1, \Delta_j^2, 0) - \Phi_3(\Delta_j^2, \Delta_j^1, 0)].
\end{aligned}
$$

Then we further have

$$
\begin{aligned}
&2E\Big[\mathbb{1}\{U_{ij} \in [\Delta_j^1, \Delta_j^2], U_{i'j} < \Delta_j^1, V_{ik} - V_{i'k} > 0\}\Big] \\
&= 2\bigg(\Phi_3(\Delta_j^2, \Delta_j^1, \infty) - \Phi_3(\Delta_j^1, \Delta_j^1, \infty) - \Phi_3(\Delta_j^2, \Delta_j^1, 0) + \Phi_3(\Delta_j^1, \Delta_j^1, 0)\bigg) \\
&= 2\bigg(\Phi(\Delta_j^1)\Big(\Phi(\Delta_j^2) - \Phi(\Delta_j^1)\Big) - \Phi_3(\Delta_j^2, \Delta_j^1, 0) + \Phi_3(\Delta_j^1, \Delta_j^1, 0)\bigg),
\end{aligned}
$$

and likewise

$$
\begin{aligned}
&2E\Big[\mathbb{1}\{U_{ij} < \Delta_j^1, U_{i'j} \in [\Delta_j^1, \Delta_j^2], V_{ik} - V_{i'k} > 0\}\Big] \\
&= 2\bigg(\Phi_3(\Delta_j^1, \Delta_j^2, \infty) - \Phi_3(\Delta_j^1, \Delta_j^1, \infty) - \Phi_3(\Delta_j^1, \Delta_j^2, 0) + \Phi_3(\Delta_j^1, \Delta_j^1, 0)\bigg) \\
&= 2\bigg(\Phi(\Delta_j^1)\Big(\Phi(\Delta_j^2) - \Phi(\Delta_j^1)\Big) - \Phi_3(\Delta_j^1, \Delta_j^2, 0) + \Phi_3(\Delta_j^1, \Delta_j^1, 0)\bigg).
\end{aligned}
$$

Hence the bridge function is given by

$$
\begin{aligned}
&E[\text{sign}(X_{ij} - X_{i'j})(X_{ik} - X_{i'k})] \\
&= 4\Phi_2(\Delta_j^2, 0, \sigma_{jk}/\sqrt{2}) - 2\Phi(\Delta_j^2) + 2[\Phi_3(\Delta_j^1, \Delta_j^2, 0) - \Phi_3(\Delta_j^2, \Delta_j^1, 0)].
\end{aligned}
$$

However, we know the fact that $(U_{ij}, U_{i'j}, \frac{V_{ik}-V_{i'k}}{\sqrt{2}})^T \overset{d}{=} (U_{i'j}, U_{ij}, -\frac{V_{ik}-V_{i'k}}{\sqrt{2}})^T$, therefore

$$
\Phi_3(\Delta_j^2, \Delta_j^1, 0) = \mathbb{P}(U_{ij} < \Delta_j^1, U_{i'j} < \Delta_j^2, \frac{V_{ik} - V_{i'k}}{\sqrt{2}} > 0)
$$

hence $\Phi_3(\Delta_j^1, \Delta_j^2, 0) - \Phi_3(\Delta_j^2, \Delta_j^1, 0) = \mathbb{P}(U_{ij} < \Delta_j^1, U_{i'j} < \Delta_j^2) = \Phi(\Delta_j^1)\Phi(\Delta_j^2)$. And the bridge function for ternary-continuous mixed data is found to be

$$
F_a(\sigma_{jk}; \Delta_j^1, \Delta_j^2) = 4\Phi_2(\Delta_j^2, 0, \sigma_{jk}/\sqrt{2}) - 2\Phi(\Delta_j^2) + 4\Phi_3(\Delta_j^1, \Delta_j^2, 0) - 2\Phi(\Delta_j^1)\Phi(\Delta_j^2).
$$

□

A.3 Proof of Lemma 3

Proof. We need to theoretically show the monotonicity of the bridge function for ternary-continuous data which boils down to show the following it monotonically increasing in r:

$$
4\Phi_3(\Delta_j^l, \Delta_j^{l+1}, 0) - 2\Phi(\Delta_j^l)\Phi(\Delta_j^{l+1})
$$

hence it suffices to show that for all l, $\dfrac{\partial \Phi_3(\Delta_j^{l-1}, \Delta_j^l, 0)}{\partial r} > 0$. Recall that the Φ_3 is the cumulative distribution function for random variables $(U_{ij}, U_{i'j}, \frac{V_{ik}-V_{i'k}}{\sqrt{2}})^T$ as defined in Section 3.2, where $(U_{ij}, U_{i'j}, \frac{V_{ik}-V_{i'k}}{\sqrt{2}})^T \sim N_3\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & r/\sqrt{2} \\ 0 & 1 & -r/\sqrt{2} \\ r/\sqrt{2} & -r/\sqrt{2} & 1 \end{bmatrix} \right)$.

For easy notation, we denote $(U_{ij}, U_{i'j}, \frac{V_{ik}-V_{i'k}}{\sqrt{2}})^T$ as $\mathbf{x} = (x_1, x_2, x_3)^T$, and

$$\Sigma = \begin{bmatrix} 1 & 0 & r/\sqrt{2} \\ 0 & 1 & -r/\sqrt{2} \\ r/\sqrt{2} & -r/\sqrt{2} & 1 \end{bmatrix}$$

for the rest of this proof.

Note that we can rewrite the normal density function $\phi_3(\mathbf{x}, \Sigma)$ as the transform of its characteristic function [12]:

$$\phi_3(\mathbf{x}, \Sigma) = (2\pi)^{-3} \iiint \exp(-i\mathbf{t}^T\mathbf{x} - \frac{1}{2}\mathbf{t}^T\mathbf{\Sigma}\mathbf{t}) d\mathbf{t} \tag{A.1}$$

A result of this is

$$\frac{\partial \phi_3(\mathbf{x})}{\partial r} = \Big(\frac{\partial^2 \phi_3}{\partial x_1 \partial x_3} - \frac{\partial^2 \phi_3}{\partial x_2 \partial x_3} \Big) \cdot (1/\sqrt{2})$$

which can be seen after interchanging the order of differentiation and integration in equation A.1.

So we now have

$$\begin{aligned}
&\frac{\partial \Phi_3(\Delta_j^l, \Delta_j^{l+1}, 0)}{\partial r/\sqrt{2}} \\
&= \int_{-\infty}^{\Delta_j^l} \int_{-\infty}^{\Delta_j^{l-1}} \int_{-\infty}^{0} \left[\frac{\partial^2 \phi_3(x_1, x_2, x_3)}{\partial x_1 \partial x_3} - \frac{\partial^2 \phi_3(x_1, x_2, x_3)}{\partial x_2 \partial x_3} \right] dx_1 dx_2 dx_3 \\
&= \int_{-\infty}^{\Delta_j^l} \int_{-\infty}^{\Delta_j^{l-1}} \left[\frac{\partial^2 \phi_3(x_1, x_2, 0)}{\partial x_1} - \frac{\partial^2 \phi_3(x_1, x_2, 0)}{\partial x_2} \right] dx_1 dx_2 \\
&= \int_{-\infty}^{\Delta_j^l} \phi_3(\Delta_j^{l-1}, x_2, 0) dx_2 - \int_{-\infty}^{\Delta_j^{l-1}} \phi_3(x_1, \Delta_j^l, 0) dx_1.
\end{aligned}$$

Recall that $(x_1, x_2, x_3) \stackrel{d}{=} (x_2, x_1, -x_3)$, we then have

$$\begin{aligned}
&= \int_{-\infty}^{\Delta_j^l} \phi_3(\Delta_j^{l-1}, x, 0) dx - \int_{-\infty}^{\Delta_j^{l-1}} \phi_3(\Delta_j^l, x, 0) dx \\
&= \Phi(\Delta_j^l) \phi_2(\Delta_j^{l-1}, 0, r/\sqrt{2}) - \Phi(\Delta_j^{l-1}) \phi_2(\Delta_j^l, 0, r/\sqrt{2})
\end{aligned}$$

where the last step arises from the fact that $X_2 | X_1 = \Delta_j^{l-1}, X_3 = 0 \sim N(0, 1)$.

Since $\Phi(\cdot) > 0$, $\phi_2(\cdot, \cdot, r/\sqrt{2}) > 0$, so in order to show $\partial \Phi_3(\Delta_j^l, \Delta_j^{l+1}, 0)/\partial r > 0$, we only need to show

$$\frac{\Phi(\Delta_j^l)}{\phi_2(\Delta_j^l, 0, r/\sqrt{2})} > \frac{\Phi(\Delta_j^{l-1})}{\phi_2(\Delta_j^{l-1}, 0, r/\sqrt{2})},$$

which is equivalent to show $\Phi(y)/\phi_2(y, 0, r/\sqrt{2})$ is increasing in y. Now we also notice that

$$\phi_2(y, 0, r/\sqrt{2}) = \phi\Big(\frac{y}{\sqrt{1 - \frac{r^2}{2}}}\Big)\phi(0)$$

due to the conditional distribution property of bivariate normal variables. Let $c = \sqrt{(1 - r^2/2)}$, then it is equivalent to show $\Phi(y)/\phi(y/c)$ increasing in x. However, we know that $\Phi(y)/\phi(y/c) = \Phi(-y)/\phi(-y/c)$. Let $\lambda(y) = \Phi(y)/\phi(y/c)$, then since $\phi'(x) = \phi(x) \cdot (-x)$ we have

$$\begin{aligned}
\lambda'(y) &= \frac{\phi(y)\phi(y/c) - \Phi(y)\phi(y/c)(-y/c^2)}{\phi^2(y/c)} \\
&= \frac{\phi(y)}{\phi(y/c)} - \frac{\Phi(y)}{\phi(y/c)} \cdot (-\frac{y}{c^2}) \\
&= \lambda(y) \left[\frac{\phi(y)}{\Phi(y)} + \frac{y}{c^2} \right]
\end{aligned}$$

We know that $\lambda(y) > 0$ for all y, so it reduces to show $\phi(y)/\Phi(y) + y/c^2 > 0$. When $y \geq 0$, it is true that $\phi(y)/\Phi(y) + y/c^2 \geq 0$. It remains to show $\phi(-y)/\Phi(-y) + (-y/c^2) > 0$ for $y > 0$. However, this is a well-known property of Mill's ratio (see Fact 7.5.6 in [57]), which states that the lower bound of $\frac{\phi(-y)}{\Phi(-y)}$ is y/c^2 (recall that $y \sim N(0, c^2)$). We thus complete the proof.

□

A.4 Proof of Lemma 3.4

Proof. Suppose X_{ij} is ternary and X_{ik} is continuous, then the sign expectation can break down as follows:

$$
\begin{aligned}
&\mathbb{E}[\text{sign}(X_{ij} - X_{i'j})(X_{ik} - X_{i'k})] \\
&= 2\Big(\mathbb{E}[I(X_{ij} = p-1, X_{ik} - X_{i'k} > 0)] - \mathbb{E}[I(X_{i'j} = p-1, X_{ik} - X_{i'k} > 0)]\Big) \\
&\quad + 2\Big(\mathbb{E}\Big[I(X_{ij} = p-2, X_{i'j} \leq p-3, X_{ik} - X_{i'k} > 0)\Big] - \mathbb{E}\Big[I(X_{ij} \leq p-3, X_{i'j} = p-2, X_{ik} - X_{i'} \\
&\quad + 2\Big(\mathbb{E}\Big[I(X_{ij} = p-3, X_{i'j} \leq p-4, X_{ik} - X_{i'k} > 0)\Big] - \mathbb{E}\Big[I(X_{ij} \leq p-4, X_{i'j} = p-3, X_{ik} - X_{i'} \\
&\quad ... \\
&\quad + 2\Big(\mathbb{E}\Big[I(X_{ij} \leq 1, X_{i'j} = 0, X_{ik} - X_{i'k} > 0)\Big] - \mathbb{E}\Big[I(X_{ij} = 0, X_{i'j} \leq 1, X_{ik} - X_{i'k} > 0)\Big]\Big).
\end{aligned}
$$

However, it is a fact that

$$
\begin{aligned}
&\mathbb{E}[I(X_{ij} = p-1, X_{ik} - X_{i'k} > 0)] - \mathbb{E}[I(X_{i'j} = p-1, X_{ik} - X_{i'k} > 0)] \\
&= \mathbb{P}(X_{i'j} = p-1, X_{ik} - X_{i'k} < 0) - \mathbb{P}(X_{ij} = p-1, X_{ik} - X_{i'k} < 0) \\
&= \Big[1 - \mathbb{P}(X_{i'j} \leq p-2, X_{ik} - X_{i'k} < 0)\Big] - \Big[1 - \mathbb{P}(X_{ij} \leq p-2, X_{ik} - X_{i'k} < 0)\Big] \\
&= \mathbb{P}(X_{ij} \leq p-2, X_{ik} - X_{i'k} < 0) - \mathbb{P}(X_{i'j} \leq p-2, X_{ik} - X_{i'k} < 0) \\
&= \Phi_2(\Delta_j^{p-1}, 0, \sigma_{jk}/\sqrt{2}) - \Phi_2(\Delta_j^{p-1}, 0, -\sigma_{jk}/\sqrt{2}) \\
&= 2\Phi_2(\Delta_j^{p-1}, 0, \sigma_{jk}/\sqrt{2}) - \Phi(\Delta_j^{p-1})
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}\Big[I(X_{ij} = p-2, X_{i'j} \leq p-3, X_{ik} - X_{i'k} > 0)\Big] \\
&\quad - \mathbb{E}\Big[I(X_{ij} \leq p-3, X_{i'j} = p-2, X_{ik} - X_{i'k} > 0)\Big] \\
&= \mathbb{P}\Big(X_{ij} \leq p-3, X_{i'j} = p-2, X_{ik} - X_{i'k} < 0\Big) - \mathbb{P}\Big(X_{ij} = p-2, X_{i'j} \leq p-3, X_{ik} - X_{i'k} < 0\Big) \\
&= \Bigg[\mathbb{P}\Big(X_{ij} \leq p-3, X_{i'j} \leq p-2, X_{ik} - X_{i'k} < 0\Big) \\
&\qquad - \mathbb{P}\Big(X_{ij} \leq p-3, X_{i'j} \leq p-3, X_{ik} - X_{i'k} < 0\Big)\Bigg] \\
&\quad - \Bigg[\mathbb{P}\Big(X_{ij} \leq p-2, X_{i'j} \leq p-3, X_{ik} - X_{i'k} < 0\Big) \\
&\qquad - \mathbb{P}\Big(X_{ij} \leq p-3, X_{i'j} \leq p-3, X_{ik} - X_{i'k} < 0\Big)\Bigg] \\
&= \mathbb{P}\Big(X_{ij} \leq p-3, X_{i'j} \leq p-2, X_{ik} - X_{i'k} < 0\Big) - \mathbb{P}\Big(X_{ij} \leq p-2, X_{i'j} \leq p-3, X_{ik} - X_{i'k} < 0\Big) \\
&= \Phi_3(\Delta_j^{p-2}, \Delta_j^{p-1}, 0) - \Phi_3(\Delta_j^{p-1}, \Delta_j^{p-2}, 0)
\end{aligned}
$$

and the other pairs of terms will follow the similar fashion.

Also notice that $(U_1, U_2, \frac{V_1-V_2}{\sqrt{2}}) \stackrel{d}{=} (U_2, U_1, -\frac{V_1-V_2}{\sqrt{2}})$, so

$$\begin{aligned}
&\Phi_3(\Delta_j^{p-2}, \Delta_j^{p-1}, 0) + \Phi_3(\Delta_j^{p-1}, \Delta_j^{p-2}, 0) \\
&= \mathbb{P}(U_1 < \Delta_j^{p-2}, U_2 < \Delta_j^{p-1}, \frac{V_1 - V_2}{\sqrt{2}} < 0) + \mathbb{P}(U_2 < \Delta_j^{p-1}, U_1 < \Delta_j^{p-2}, \frac{V_1 - V_2}{\sqrt{2}} > 0) \\
&= \mathbb{P}(U_1 < \Delta_j^{p-2}, U_2 < \Delta_j^{p-1}) \\
&= \Phi(\Delta_j^{p-2})\Phi(\Delta_j^{p-1}).
\end{aligned}$$

Therefore

$$\Phi_3(\Delta_j^{p-2}, \Delta_j^{p-1}, 0) - \Phi_3(\Delta_j^{p-1}, \Delta_j^{p-2}, 0) = 2\Phi_3(\Delta_j^{p-2}, \Delta_j^{p-1}, 0) - \Phi(\Delta_j^{p-2})\Phi(\Delta_j^{p-1}).$$

In addition, recall that $\hat{\Delta}_j^p = \Phi^{-1}(\frac{I(X_{ij} \leq p-1)}{n}) = \Phi^{-1}(1) = \infty$, so it holds that

$$\Phi_2(\hat{\Delta}_j^p, 0, \sigma_{jk}/\sqrt{2}) = \Phi_3(\hat{\Delta}_j^{p-1}, \hat{\Delta}_j^p, 0),$$

so now we can alternatively express the bridge function as

$$F(\sigma_{jk}; \boldsymbol{\Delta}_j) = \sum_{l=1}^{p-1} 4\Phi_3(\Delta_j^l, \Delta_j^{l+1}, 0) - 2\Phi(\Delta_j^l)\Phi(\Delta_j^{l+1}).$$

□

A.4.1 Proof of Theorem 3.1

Proof. We begin the proof by showing the Lipschitz continuity of the bridge function. Recall that

$$\begin{aligned}
F_a(\sigma_{jk}; \boldsymbol{\Delta}_j) &= \sum_{l=1}^{p-1} 4\Phi_3(\Delta_j^l, \Delta_j^{l+1}, 0) - 2\Phi(\Delta_j^l)\Phi(\Delta_j^{l+1}) \\
&= 4\Phi_2(\Delta_j^{p-1}, 0, \sigma_{jk}/\sqrt{2}) - 2\Phi(\Delta_j^{p-1}) + \sum_{l=1}^{p-2} 4\Phi_3(\Delta_j^l, \Delta_j^{l+1}, 0) - 2\Phi(\Delta_j^l)\Phi(\Delta_j^{l+1})
\end{aligned}$$

Also it holds that

$$\frac{\Phi_2(\Delta_j^{p-1}, 0, \sigma_{jk}/\sqrt{2})}{\sigma_{jk}} > \frac{1}{L_2}$$

[14], and from Lemma 3.3 we have

$$\frac{\Phi_3(\Delta_j^{l}, \Delta_j^{l+1}, 0)}{\sigma_{jk}} \geq 0$$

for all l, then $F_a(\sigma_{jk}; \mathbf{\Delta}_j)$ is Lipschitz continuous with constant $L2$.

Consequently, for $\hat{\mathbf{\Delta}}_\mathbf{j} \in A_j$, the Lipschitz continuity of $F^{-1}(\tau^a; \mathbf{\Delta}_j)$ gives rise to

$$|F_a^{-1}(\hat{\tau}^a; \hat{\Delta}_j^1, \hat{\Delta}_j^2) - F^{-1}(F_a(\sigma_{jk}; \hat{\Delta}_j^1, \hat{\Delta}_j^2); \hat{\mathbf{\Delta}}_j)| \leq L_2 |\hat{\tau}^a - F(r; \hat{\mathbf{\Delta}}_j)|.$$

Now recall that $\Phi^{-1}(\cdot)$ is Lipschitz continuous in $[\Phi(-2M), \Phi(2M)]$, we have a Lipschitz constant L_1 such that

$$\begin{aligned} |\hat{\Delta}_j^1 - \Delta_j^1| &= \left| \Phi^{-1}\left(\frac{\sum_{i=1}^n I(X_{ij} = 0)}{n} \right) - \Phi^{-1}(\Phi(\Delta_j^1)) \right| \\ &\leq L_1 \left| \frac{\sum_{i=1}^n I(X_{ij} = 0)}{n} - \Phi(\Delta_j^1) \right|. \end{aligned}$$

The exception probability is controlled by

$$\begin{aligned} P(A_{j,1}^c) &= P(|\hat{\Delta}_j^1| > 2M) \\ &\leq P(|\hat{\Delta}_j^1| - |\Delta_j^1| > M) \\ &\leq P(|\hat{\Delta}_j^1 - \Delta_j^1| > M) \\ &\leq P\left(\left| \frac{\sum_{i=1}^n I(X_{ij} = 0)}{n} - \Phi(\Delta_j^1) \right| > \frac{M}{L_1} \right) \\ &\leq 2 \exp\left(- \frac{2M^2 n}{L_1^2} \right) \qquad \text{(by Hoeffding's inequality).} \end{aligned}$$

Likewise, under $A_{j,l} = \{|\hat{\Delta}_j^l| \leq 2M\}$, we have

$$|\hat{\Delta}_j^l - \Delta_j^l| \leq L_1 \left| \frac{\sum_{i=1}^n I(X_{ij} \leq l-1)}{n} - \Phi(\Delta_j^l) \right|;$$

and

$$P(A_{j,l}^c) \leq 2\exp\Big(-\frac{2M^2n}{L_1^2}\Big).$$

Now we define the event $A_j = \bigcap_{l=1}^{2} A_{j,l}$, as a result we have

$$\begin{aligned} P(A_j^c) &= P(\bigcup_{l=1}^{p-1} A_{j,l}^c) \\ &\leq \sum_{l=1}^{p-1} P(A_{j,l}^c) \\ &\leq 2(p-1)\exp\Big(-\frac{2M^2n}{L_1^2}\Big). \end{aligned}$$

For any $t > 0$, we have

$$\begin{aligned} &P(|F_a^{-1}(\hat{\tau}^a; \hat{\mathbf{\Delta}}_j) - \sigma_{jk}| \geq t) \\ &= P(\Big\{|F_a^{-1}(\hat{\tau}^a; \hat{\mathbf{\Delta}}_j) - \sigma_{jk}| \geq t\Big\} \cap A_j) + P(\Big\{|F_a^{-1}(\hat{\tau}^a; \hat{\mathbf{\Delta}}_j) - \sigma_{jk}| \geq t\Big\} \cap A_j^c) \\ &\leq P(\Big\{|F_a^{-1}(\hat{\tau}^a; \hat{\mathbf{\Delta}}_j) - \sigma_{jk}| \geq t\Big\} \cap A_j) + P(A_j^c). \end{aligned}$$

Recall that $F^{-1}(\tau^a; \mathbf{\Delta})$ is Lipschitz continuous on $[-1, 1]$ with Lipschitz constant L, we then have

$$\begin{aligned} &P(\Big\{|F_a^{-1}(\hat{\tau}^a; \hat{\mathbf{\Delta}}_j) - \sigma_{jk}| \geq t\Big\} \cap A_j) \\ &= P(\Big\{|F_a^{-1}(\hat{\tau}^a; \hat{\mathbf{\Delta}}_j) - F^{-1}(F_a(\sigma_{jk}; \hat{\mathbf{\Delta}}_j); \hat{\mathbf{\Delta}}_j)| \geq t\Big\} \cap A_j) \\ &\leq P(\{L|\hat{\tau}^a - F(r; \hat{\mathbf{\Delta}}_j)| > t\} \cap A_j) \\ &\leq P(\{L|\hat{\tau}^a - F_a(\sigma_{jk}; \mathbf{\Delta}_j)| + L|F_a(\sigma_{jk}; \mathbf{\Delta}_j)) - F(r; \hat{\mathbf{\Delta}}_j)| > t\} \cap A_j) \\ &\leq P(\{L|\hat{\tau}^a - F_a(\sigma_{jk}; \mathbf{\Delta}_j)| > \frac{t}{2}\} \cap A_j) + P(\{L|F_a(\sigma_{jk}; \mathbf{\Delta}_j)) - F(r; \hat{\mathbf{\Delta}}_j)| > \frac{t}{2}\} \cap A_j) \\ &\leq P(L|\hat{\tau}^a - F_a(\sigma_{jk}; \mathbf{\Delta}_j)| > \frac{t}{2}) + P(\{L|F_a(\sigma_{jk}; \mathbf{\Delta}_j)) - F(r; \hat{\mathbf{\Delta}}_j)| > \frac{t}{2}\} \cap A_j) \\ &\equiv I_1 + I_2. \end{aligned}$$

Since $\hat{\tau}^a$ is a U-statistic with bounded kernel, it is immediate by Hoeffding's inequality that

$$I_1 = P(L_2|\hat{\tau}^a - F_a(\sigma_{jk}; \mathbf{\Delta}_j)| > \frac{t}{2}) \leq 2\exp\Big(-\frac{nt^2}{2L_2^2}\Big).$$

Let $\Phi_{21}(x,y,t) = \frac{\partial \Phi_2(x,y,t)}{\partial x}$, $\Phi_{31}(x,y,z) = \frac{\partial \Phi_3(x,y,t)}{\partial x}$, and $\Phi_{32}(x,y,z) = \frac{\partial \Phi_3(x,y,t)}{\partial y}$. For I_2, we have

$$
\begin{aligned}
&|F(\sigma_{jk};\mathbf{\Delta}_j) - F(\sigma_{jk};\hat{\mathbf{\Delta}}_j)| \\
&= \left| \sum_{l=1}^{p-1} 4\Phi_3(\Delta_j^l, \Delta_j^{l+1}, 0) - 2\Phi(\Delta_j^l)\Phi(\Delta_j^{l+1}) - 4\Phi_3(\hat{\Delta}_j^l, \hat{\Delta}_j^{l+1}, 0) + 2\Phi(\hat{\Delta}_j^l)\Phi(\hat{\Delta}_j^{l+1}) \right| \\
&\leq \sum_{l=1}^{p-1} 4\left|\Phi_3(\Delta_j^l, \Delta_j^{l+1}, 0) - \Phi_3(\hat{\Delta}_j^l, \hat{\Delta}_j^{l+1}, 0)\right| + 2\left|\Phi(\Delta_j^l)\Phi(\Delta_j^{l+1}) - \Phi(\hat{\Delta}_j^l)\Phi(\hat{\Delta}_j^{l+1})\right| \\
&\leq 4\sum_{l=1}^{p-1} \left(\Phi_{31}(\zeta_{1,l})|\Delta_j^l - \hat{\Delta}_j^l| + \Phi_{32}(\zeta_{2,l})|\Delta_j^{l+1} - \hat{\Delta}_j^{l+1}| \right) \\
&\qquad + 2\sum_{l=1}^{p-1} \left(\Phi(\hat{\Delta}_j^l)\phi(\eta_{1,l})|\Delta_j^{l+1} - \hat{\Delta}_j^{l+1}| + \Phi(\hat{\Delta}_j^{l+1})\phi(\eta_{2,l})|\Delta_j^l - \hat{\Delta}_j^l| \right) \\
&\leq 4\sum_{l=1}^{p-1} \frac{1}{\sqrt{2\pi}}|\Delta_j^l - \hat{\Delta}_j^l| + \frac{1}{\sqrt{2\pi}}|\Delta_j^{l+1} - \hat{\Delta}_j^{l+1}| \\
&\qquad + 2\sum_{l=1}^{p-1} \frac{1}{\sqrt{2\pi}}|\Delta_j^l - \hat{\Delta}_j^l| + \frac{1}{\sqrt{2\pi}}|\Delta_j^{l+1} - \hat{\Delta}_j^{l+1}| \\
&= 6\sum_{l=2}^{p-2} \frac{\sqrt{2}}{\sqrt{\pi}}|\Delta_j^l - \hat{\Delta}_j^l| + \frac{6}{\sqrt{2\pi}}|\Delta_j^1 - \hat{\Delta}_j^1| + \frac{6}{\sqrt{2\pi}}|\Delta_j^{p-1} - \hat{\Delta}_j^{p-1}|.
\end{aligned}
$$

We now can establish the bound for I_2:

$$
\begin{aligned}
I_2 &\leq P(\{6\sum_{l=2}^{p-2} \frac{\sqrt{2}}{\sqrt{\pi}}|\Delta_j^l - \hat{\Delta}_j^l| + \frac{6}{\sqrt{2\pi}}|\Delta_j^1 - \hat{\Delta}_j^1| + \frac{6}{\sqrt{2\pi}}|\Delta_j^{p-1} - \hat{\Delta}_j^{p-1}| > \frac{t}{2L_2}\} \cap A_j) \\
&\leq P(\left|\frac{\sum_{i=1}^n I(X_{ij}=0)}{n} - \Phi(\Delta_j^1)\right| > \frac{t\sqrt{2\pi}}{12L_1L_2(p-1)}) \\
&\quad + P(\left|\frac{\sum_{i=1}^n I(X_{ij}\leq p-1)}{n} - \Phi(\Delta_j^{p-1})\right| > \frac{t\sqrt{2\pi}}{12L_1L_2p}) \\
&\quad + \sum_{l=2}^{p-1} P(\left|\frac{\sum_{i=1}^n I(X_{ij}\leq l-1)}{n} - \Phi(\Delta_j^l)\right| > \frac{t\sqrt{2\pi}}{24L_1L_2p}) \\
&\leq 2\exp(-\frac{4nt^2\pi}{12^2L_1^2L_2^2p^2}) + 2\exp(-\frac{4nt^2\pi}{12^2L_1^2L_2^2p^2}) + 2(p-2)\exp(-\frac{4nt^2\pi}{24^2L_1^2L_2^2p^2}) \\
&\leq 2(p-1)\exp(-\frac{4nt^2\pi}{24^2L_1^2L_2^2p^2}).
\end{aligned}
$$

So putting everything together we have

$$P(||F_a^{-1}(\hat{\tau}^a;\boldsymbol{\Delta}_j)-\sigma_{jk}||>t)\leq 2(p-1)\exp\Big(-\frac{2M^2n}{L_1^2}\Big)+2\exp\big(-\frac{nt^2}{2L_2^2}\big)$$
$$+2(p-1)\exp(-\frac{4nt^2\pi}{24^2L_1^2L_2^2p^2})$$

implying that

$$P(\sup||\Sigma-R||>t)\leq\sum_{j,k}P(||F_a^{-1}(\hat{\tau}^a;\boldsymbol{\Delta}_j)-\sigma_{jk}||>t)$$
$$\leq 2d^2p\exp\Big(-\frac{2M^2n}{L_1^2}\Big)+2d^2\exp\big(-\frac{nt^2}{2L_2^2}\big)$$
$$+2d^2p\exp(-\frac{4nt^2\pi}{24^2L_1^2L_2^2p^2}).$$

Therefore at fixed p, taking $t=C\sqrt{\frac{\log d}{n}}$ we have

$$P(\sup||\Sigma-R||<C\sqrt{\frac{\log d}{n}})>1-d^{-1}.$$

□

A.5 Proof of Corollary 3.1

Proof. By Lipschitz continuity of $\Phi^{-1}(\cdot)$ in $[\Phi(-2M),\Phi(2M)]$, we know that under the event $A_{j,1}=\{|\hat{\Delta}_j^1|\leq 2M\}$, there exists a Lipschitz constant L_1 such that

$$|\hat{\Delta}_j^1-\Delta_j^1|=\left|\Phi^{-1}\left(\frac{\sum_{i=1}^n I(X_{ij}=0)}{n}\right)-\Phi^{-1}(\Phi(\Delta_j^1))\right|$$
$$\leq L_1\left|\frac{\sum_{i=1}^n I(X_{ij}=0)}{n}-\Phi(\Delta_j^1)\right|.$$

The exception probability is controlled by

$$
\begin{aligned}
P(A_{j,1}^c) &= P(|\hat{\Delta}_j^1| > 2M) \\
&\le P(|\hat{\Delta}_j^1| - |\Delta_j^1| > M) \\
&\le P(|\hat{\Delta}_j^1 - \Delta_j^1| > M) \\
&\le P\left(\left|\frac{\sum_{i=1}^n I(X_{ij} = 0)}{n} - \Phi(\Delta_j^1)\right| > \frac{M}{L_1}\right) \\
&\le 2\exp\left(-\frac{2M^2 n}{L_1^2}\right) \qquad \text{(by Hoeffding's inequality).}
\end{aligned}
$$

Likewise, under $A_{j,2} = \{|\hat{\Delta}_j^2| \le 2M\}$, we have

$$
|\hat{\Delta}_j^2 - \Delta_j^2| \le L_1 \left|\frac{\sum_{i=1}^n I(X_{ij} \le 1)}{n} - \Phi(\Delta_j^2)\right|;
$$

and

$$
P(A_{j,2}^c) \le 2\exp\left(-\frac{2M^2 n}{L_1^2}\right).
$$

Now we define the event $A_j = \bigcap_{l=1}^{2} A_{j,l}$, as a result we have

$$
\begin{aligned}
P(A_j^c) &= P(\bigcup_{l=1}^{2} A_{j,l}^c) \\
&\le \sum_{l=1}^{2} P(A_{j,l}^c) \\
&\le 4\exp\left(-\frac{2M^2 n}{L_1^2}\right).
\end{aligned}
$$

For any $t > 0$, we have

$$
\begin{aligned}
&P(|F_a^{-1}(\hat{\tau}^a; \hat{\Delta}_j^1, \hat{\Delta}_j^2) - \sigma_{jk}| \ge t) \\
&= P(\left\{|F_a^{-1}(\hat{\tau}^a; \hat{\Delta}_j^1, \hat{\Delta}_j^2) - \sigma_{jk}| \ge t\right\} \cap A_j) + P(\left\{|F_a^{-1}(\hat{\tau}^a; \hat{\Delta}_j^1, \hat{\Delta}_j^2) - \sigma_{jk}| \ge t\right\} \cap A_j^c) \\
&\le P(\left\{|F_a^{-1}(\hat{\tau}^a; \hat{\Delta}_j^1, \hat{\Delta}_j^2) - \sigma_{jk}| \ge t\right\} \cap A_j) + P(A_j^c).
\end{aligned}
$$

Recall that $F_a(\sigma_{jk}; \Delta_j^1, \Delta_j^2) = 4\Phi_2(\Delta_j^2, 0, \sigma_{jk}/\sqrt{2}) - 2\Phi(\Delta_j^2) + 4\Phi_3(\Delta_j^1, \Delta_j^2, 0) - 2\Phi(\Delta_j^1)\Phi(\Delta_j^2)$ and

$$
\frac{\partial}{\partial r} 4\Phi_2(\Delta_j^2, 0, r/\sqrt{2}) - 2\Phi(\Delta_j^2) > \frac{1}{L_2}
$$

from [14], also we have $\frac{\partial}{\partial r}\Phi_3(\Delta_j^{l-1}, \Delta_j^l, 0) \geq 0$ from Lemma 3.3, so

$$\frac{\partial}{\partial r} F_a(\sigma_{jk}; \Delta_j^1, \Delta_j^2) > \frac{1}{L_2},$$

implying $F_a(\sigma_{jk}; \cdot)$ is Lipschitz-continuous in $[-1, 1]$ with Lipschitz constant L_2.

Then by Lipschitz continuity we have

$$\begin{aligned}
&P(\Big\{|F_a^{-1}(\hat{\tau}^a; \hat{\Delta}_j^1, \hat{\Delta}_j^2) - \sigma_{jk}| \geq t\Big\} \cap A_j) \\
&= P(\Big\{|F_a^{-1}(\hat{\tau}^a; \hat{\Delta}_j^1, \hat{\Delta}_j^2) - F_a^{-1}(F_a(\sigma_{jk}; \hat{\Delta}_j^1, \hat{\Delta}_j^2); \hat{\Delta}_j^1, \hat{\Delta}_j^2)| \geq t\Big\} \cap A_j) \\
&\leq P(\{L_2|\hat{\tau}^a - F(r; \hat{\Delta}_j^1, \hat{\Delta}_j^2)| > t\} \cap A_j) \\
&\leq P(\{L_2|\hat{\tau}^a - F_a(\sigma_{jk}; \Delta_j^1, \Delta_j^2)| + L_2|F_a(\sigma_{jk}; \Delta_j^1, \Delta_j^2)) - F_a(\sigma_{jk}; \hat{\Delta}_j^1, \hat{\Delta}_j^2)| > t\} \cap A_j) \\
&\leq P(\{L_2|\hat{\tau}^a - F_a(\sigma_{jk}; \Delta_j^1, \Delta_j^2)| > \frac{t}{2}\} \cap A_j) \\
&\qquad + P(\{L_2|F_a(\sigma_{jk}; \Delta_j^1, \Delta_j^2)) - F_a(\sigma_{jk}; \hat{\Delta}_j^1, \hat{\Delta}_j^2)| > \frac{t}{2}\} \cap A_j) \\
&\leq P(L_2|\hat{\tau}^a - F_a(\sigma_{jk}; \Delta_j^1, \Delta_j^2)| > \frac{t}{2}) \\
&\qquad + P(\{L_2|F_a(\sigma_{jk}; \Delta_j^1, \Delta_j^2)) - F_a(\sigma_{jk}; \hat{\Delta}_j^1, \hat{\Delta}_j^2)| > \frac{t}{2}\} \cap A_j) \\
&\equiv I_1 + I_2.
\end{aligned}$$

Since $\hat{\tau}^a$ is a U-statistic with bounded kernel, it is immediate by Hoeffding's inequality that

$$I_1 = P(L_2|\hat{\tau}^a - F_a(\sigma_{jk}; \Delta_j^1, \Delta_j^2)| > \frac{t}{2}) \leq 2\exp\Big(-\frac{nt^2}{2L_2^2}\Big).$$

Let $\Phi_{21}(x, y, t) = \frac{\partial \Phi_2(x,y,t)}{\partial x}$, $\Phi_{31}(x, y, z) = \frac{\partial \Phi_3(x,y,t)}{\partial x}$, and $\Phi_{32}(x, y, z) = \frac{\partial \Phi_3(x,y,t)}{\partial y}$. For I_2, we have

$$
\begin{aligned}
&|F_a(\sigma_{jk};\Delta_j^1,\Delta_j^2)) - F_a(\sigma_{jk};\hat{\Delta}_j^1,\hat{\Delta}_j^2)| \\
&\le 4|\Phi_2(\Delta_j^2,0,\sigma_{jk}/\sqrt{2}) - \Phi_2(\hat{\Delta}_j^2,0,\sigma_{jk}/\sqrt{2})| + 2|\Phi(\Delta_j^2) - \Phi(\hat{\Delta}_j^2)| \\
&\quad + 4|\Phi_3(\Delta_j^1,\Delta_j^2,0) - \Phi_3(\hat{\Delta}_j^1,\hat{\Delta}_j^2,0)| + 2|\Phi(\Delta_j^1)\Phi(\Delta_j^2) - \Phi(\hat{\Delta}_j^1)\Phi(\hat{\Delta}_j^2)| \\
&\le 4\Phi_{21}(\zeta_1)|\Delta_j^2 - \hat{\Delta}_j^2| + 2\phi(\zeta_2)|\Delta_j^2 - \hat{\Delta}_j^2| + 4\Phi_{31}(\zeta_3)|\Delta_j^1 - \hat{\Delta}_j^1| + 4\Phi_{32}(\zeta_4)|\Delta_j^2 - \hat{\Delta}_j^2| \\
&\quad 2\Phi(\hat{\Delta}_j^1)\phi(\zeta_5)|\Delta_j^2 - \hat{\Delta}_j^2| + 2\Phi(\hat{\Delta}_j^2)\phi(\zeta_6)|\Delta_j^1 - \hat{\Delta}_j^1|.
\end{aligned}
$$

It has been shown that $\Phi_{21}(x,y,t) \le \frac{1}{\sqrt{2\pi}}$ from [14] . For the upper bound of $\Phi_{31}(x,y,z)$, we know that the conditional distribution of (Y,Z) given X is bivariate normal:

$$
Y,Z|X = x \sim N\left(\begin{bmatrix} 0 \\ \frac{x\sigma_{jk}}{\sqrt{2}} \end{bmatrix}, \begin{bmatrix} 1 & -\sigma_{jk}/\sqrt{2} \\ -\sigma_{jk}/\sqrt{2} & 1 \end{bmatrix} \right).
$$

Let $\phi_2(y,z|x)$ denote the density function for the conditional distribution, and $\Phi_2(y,z|x)$ denote the distribution function. Therefore

$$
\Phi_3(\Delta_j^1,\Delta_j^2,0) = \int_{-\infty}^{\Delta_j^1}\int_{-\infty}^{\Delta_j^2}\int_{-\infty}^{0} \phi_2(y,z|x)\phi(x)dzdydx = \int_{-\infty}^{\Delta_j^1} \Phi_2(\Delta_j^2,0|x)\phi(x)dx
$$

hence

$$
\Phi_{31} = \frac{\partial\Phi_3(\Delta_j^1,\Delta_j^2,0)}{\partial\Delta_j^1} = \frac{\partial}{\partial\Delta_j^1}\int_{-\infty}^{\Delta_j^1} \Phi_2(\Delta_j^2,0|x)\phi(x)dx = \Phi_2(\Delta_j^2,0|\Delta_j^1)\phi(\Delta_j^1) \le \frac{1}{\sqrt{2\pi}}
$$

and

$$
|F_a(\sigma_{jk};\Delta_j^1,\Delta_j^2)) - F_a(\sigma_{jk};\hat{\Delta}_j^1,\hat{\Delta}_j^2)| \le \frac{12}{\sqrt{2\pi}}|\Delta_j^2 - \hat{\Delta}_j^2| + \frac{6}{\sqrt{2\pi}}|\Delta_j^1 - \hat{\Delta}_j^1|.
$$

As a result, the upper bound for I_2 is established:

$$
\begin{aligned}
I_2 &\le P(\{\frac{12}{\sqrt{2\pi}}L_2|\Delta_j^2 - \hat{\Delta}_j^2| + \frac{6}{\sqrt{2\pi}}L_2|\Delta_j^1 - \hat{\Delta}_j^1| > \frac{t}{2}\} \cap A_j) \\
&\le P(\left|\frac{\sum_{i=1}^n I(X_{ij} \le 1)}{n} - \Phi(\Delta_j^2)\right| > \frac{t\sqrt{2\pi}}{48L_1L_2}) + P(\left|\frac{\sum_{i=1}^n I(X_{ij} = 0)}{n} - \Phi(\Delta_j^1)\right| > \frac{t\sqrt{2\pi}}{24L_1L_2}) \\
&\le 2\exp(-\frac{nt^2\pi}{48^2L_1^2L_2^2}) + 2\exp(-\frac{nt^2\pi}{24^2L_1^2L_2^2}).
\end{aligned}
$$

So putting together we have

$$P\Big(\Big|F_a^{-1}(\hat{\tau}_a;\hat{\Delta}_j^1,\hat{\Delta}_j^2)-\sigma_{jk}\Big|>t\Big)\leq 4\exp\bigg(-\frac{2M^2n}{L_1^2}\bigg)+2\exp\Big(-\frac{nt^2}{2L_2^2}\Big)$$
$$+2\exp(-\frac{nt^2\pi}{48^2L_1^2L_2^2})+2\exp(-\frac{nt^2\pi}{24^2L_1^2L_2^2}).$$

□

A.6 Proof of Lemma 4.1

Proof. The 1st-order Taylor expansion gives rise to

$$\begin{aligned}\mathbb{E}(\hat{\tau}_{jk}^b)&=\mathbb{E}\bigg[\frac{C-D}{\sqrt{\Big[\binom{n}{2}-t_{X_j}\Big]\Big[\binom{n}{2}-t_{X_k}\Big]}}\bigg]\\&\approx\frac{\mathbb{E}(C-D)}{\mathbb{E}\Big(\sqrt{\Big[\binom{n}{2}-t_{X_j}\Big]\Big[\binom{n}{2}-t_{X_k}\Big]}\Big)}\\&=\frac{2\Big[\Phi_2(\Delta_j,\Delta_k,\sigma_{jk})-\Phi(\Delta_j)\Phi(\Delta_k)\Big]}{\sqrt{1-p_j}\,\sqrt{1-p_k}}\end{aligned}$$

where p_j is the probability of getting a tied pair at X_j, and likewise for p_k.

We know that

$$\begin{aligned}1-p_j&=P([(1,x_{ik})(0,x_{i'k})])+P([(0,x_{ik})(1,x_{i'k})])\\&=2P([(1,x_{ik})(0,x_{i'k})])\\&=2\bigg[\Phi(\Delta_j)\Big(1-\Phi(\Delta_j)\Big)\bigg]\end{aligned}$$

and likewise $1-p_k = 2\Big[\Phi(\Delta_k)\big(1-\Phi(\Delta_k)\big)\Big]$. Combining these results, we have

$$F_b(\sigma_{jk};\Delta_j,\Delta_k) = \frac{\Phi_2(\Delta_j,\Delta_k,\sigma_{jk}) - \Phi(\Delta_j)\Phi(\Delta_k)}{\sqrt{(\Phi(\Delta_j)-\Phi(\Delta_j)^2)(\Phi(\Delta_k)-\Phi(\Delta_k)^2)}}.$$

□

A.7 Proof of Lemma 4.2

Proof. Since $\mathbf{X}_k$ is continuous, we do not need to consider tieing at $\mathbf{X}_k$. Therefore, the bridge function is easily derived as

$$\begin{aligned}
\mathbb{E}(\hat{\tau}^b_{jk}) &= \mathbb{E}\left[\frac{C-D}{\sqrt{\left[\binom{n}{2}-t_{X_j}\right]\left[\binom{n}{2}\right]}}\right] \\
&\approx \frac{\mathbb{E}(C-D)}{\mathbb{E}\Big(\sqrt{\left[\binom{n}{2}-t_{X_j}\right]\left[\binom{n}{2}\right]}\Big)} \\
&= \frac{4\Phi_2(\Delta_j,0,\sigma_{jk}/\sqrt{2}) - 2\Phi(\Delta_j)}{\sqrt{1-p_j}} \\
&= \frac{4\Phi_2(\Delta_j,0,\sigma_{jk}/\sqrt{2}) - 2\Phi(\Delta_j)}{\sqrt{2(\Phi(\Delta_j)) - 2(\Phi(\Delta_j))^2}}
\end{aligned}$$

where in the second last step we adopt the result from Kendall's τ^a version bridge function in [14] and the last step uses the result derived in A7.

□

A.8 Proof of Lemma 4.3

Proof. 2nd order Taylor expansion gives:

$$\begin{aligned} E(Y/X) &\approx \frac{\mu_Y}{\mu_X} + \sigma_X^2 \frac{\mu_Y}{\mu_X^3} - \frac{\sigma_{XY}}{\mu_X^2} \\ &= \frac{\mu_Y}{\mu_X} + \frac{1}{\mu_X^2}\left(\sigma_X^2 \frac{\mu_Y}{\mu_X} - \rho\sigma_X\sigma_Y\right). \end{aligned}$$

Therefore we have

$$\begin{aligned} E(\hat{\tau}_{jk}^b) &= E\Big(\frac{\sum\limits_{1\le i<i'\le n}(X_{ij}-X_{i'j})\mathrm{sign}(X_{ik}-X_{i'k})}{\sqrt{\binom{n}{2}-\sum_i\binom{n_{i+}}{2}}\sqrt{\binom{n}{2}}}\Big) \\ &= E\Big(\frac{\sqrt{\binom{n}{2}}\hat{\tau}_{jk}^a}{\sqrt{\binom{n}{2}-T}}\Big) \\ &\approx \frac{\sqrt{\binom{n}{2}}E[\hat{\tau}_{jk}^a]}{E[\sqrt{\binom{n}{2}-T}]} \\ &\quad + \frac{1}{\Big[E[\sqrt{\binom{n}{2}-T}]\Big]^2}\Bigg[\binom{n}{2}\mathrm{var}\Big(\sqrt{\binom{n}{2}-T}\Big)\frac{\sqrt{\binom{n}{2}}E[\hat{\tau}_{jk}^a]}{E[\sqrt{\binom{n}{2}-T}]} \\ &\qquad - \mathrm{cov}\Big(\sqrt{\binom{n}{2}}\hat{\tau}_{jk}^a,\ \sqrt{\binom{n}{2}-T}\Big)\Bigg]. \end{aligned}$$

We compute each part separately. First, note that $E[\hat{\tau}_{jk}^a]$ can be directly adopted from [14], namely

$$\begin{aligned} E[\hat{\tau}_{jk}^a] &= E\Big[\sum_{1\le i<i'\le n}(X_{ij}-X_{i'j})\mathrm{sign}(X_{ik}-X_{i'k})\Big] \\ &= 4\Phi_2(\Delta_j, 0, \sigma_{jk}/\sqrt{2}) - 2\Phi(\Delta_j). \end{aligned}$$

For $E\left[\sqrt{\binom{n}{2}-t_{X_j}}\right]$, we know that the number of ties are $\binom{n_0}{2}+\binom{n_1}{2}$ where

n_0 is the number of $X_{ij} = 0$ for $i = 1, \ldots, n$ and n_1 is the number of $X_{ij} = 1$ for $i = 1, \ldots, n$. Also recall that $\mathbb{P}(X_{ij} = 0) = \Phi(\Delta_j)$, therefore we have

$$E\left[\sqrt{\binom{n}{2} - T}\right] = E\left[\sqrt{\binom{n}{2} - \binom{n_0}{2} - \binom{n_1}{2}}\right]$$
$$= \sum_{n_0=0}^{n} \left[\sqrt{\binom{n}{2} - \binom{n_0}{2} - \binom{n-n_0}{2}}\right] \binom{n}{n_0} \big(\Phi(\Delta_j)\big)^{n_0} \big(1 - \Phi(\Delta_j)\big)^{n-n_0}$$

and consequently

$$\text{var}\left[\sqrt{\binom{n}{2} - T}\right] = E\left[\binom{n}{2} - T\right] - (E\left[\sqrt{\binom{n}{2} - T}\right])^2$$
$$= \binom{n}{2}\big(2\Phi(\Delta_j) - 2[\Phi(\Delta_j)]^2\big) - (E\left[\sqrt{\binom{n}{2} - T}\right])^2.$$

As for $\text{cov}\left(\sqrt{\binom{n}{2}}\tau^a_{jk}, \sqrt{\binom{n}{2} - T}\right)$, we know that $\tau^a_{jk} = \frac{C-D}{\binom{n}{2}}$, and $\binom{n}{2} - T = C + D$, so we can instead compute

$$\text{cov}\Big((C - D), \sqrt{C + D}\Big) = E[(C - D)\sqrt{C + D}] - E(C - D)E[\sqrt{C + D}]$$
$$= E[(C - D)\sqrt{C + D}] - E(C - D)E[\sqrt{C + D}]$$

where we can compute $E(\sqrt{(C - D)(C^2 - D^2)})$ from the fact that (C, D) follows a multinomial distribution with parameters

$$p_C = 2\mathbb{P}[(X_{ij} = 0, X_{i'j} = 1, (X_{ik} - X_{i'k})/\sqrt{2} < 0)]$$
$$= 2[\Phi_2(\Delta_j, 0, \sigma_{jk}/\sqrt{2}) - \Phi_3(\Delta_j, \Delta_j, 0)]$$

and

$$p_D = 2\mathbb{P}[(X_{ij} = 1, X_{i'j} = 0, (X_{ik} - X_{i'k})/\sqrt{2} < 0)]$$
$$= 2[\Phi_2(\Delta_j, 0, -\sigma_{jk}/\sqrt{2}) - \Phi_3(\Delta_j, \Delta_j, 0)]$$

so

$$E[(C-D)\sqrt{C+D}] = \sum_{(C,D)\in S} (C-D)\sqrt{C+D}\frac{\binom{n}{2}}{C!D!\left(\binom{n}{2}-C-D\right)!}p_C^C p_D^D(1-p_C-p_D)^{\binom{n}{2}-C-D}$$

with the sample space of (C, D) being $S = \{(C, D) : C \in \mathbb{Z}^+, D \in \mathbb{Z}^+, C+D \leq n\}$.

Putting these together, we have

$$\begin{aligned}
&\operatorname{cov}\left(\sqrt{\binom{n}{2}}\tau_{jk}^a, \sqrt{\binom{n}{2}-t_{X_j}}\right) \\
&= \sum_{(C,D)\in S}\left\{(C-D)\sqrt{(C+D)}\frac{\sqrt{\binom{n}{2}}}{C!D!\left(\binom{n}{2}-C-D\right)!}\cdot\right. \\
&\qquad \left. p_C^C p_D^D(1-p_C-p_D)^{\binom{n}{2}-C-D}\right\} - \sqrt{\binom{n}{2}}\mathbb{E}(\hat{\tau}_{jk}^a)\mathbb{E}\Big[\sqrt{\binom{n}{2}-t_{X_j}}\Big].
\end{aligned}$$

□

APPENDIX B

CHAPTER 3 OF APPENDIX

B.1 Proof of Theorem 2

Proof. Before we prove the theorem, we adapted the following lemmas from [10] regarding the theoretical properties of the Kendall's tau based estimator, $\hat{\Sigma}$, of the correlation/covariance matrix.

Lemma 8 (Adapted results from [22], [59],[46])**.** *The Kendall's tau based correlation/covariance matrix estimator $\hat{\Sigma}$ in Equation (3.6) has the following concentration rates under different norms:*

1. $$P(\sup_{1\leq j<k\leq d}|\hat{\Sigma}_{jk}-\Sigma_{jk}|\lesssim\sqrt{\frac{\log d}{n}})\geq 1-p^{-1};$$

2. *If the covariance matrix has a bounded condition number, namely $\kappa(\Sigma)<M$ for some $M>0$, then*
$$P(||\hat{\Sigma}-\Sigma||_2\lesssim\max\{\sqrt{\frac{d+t}{n}},\frac{d+t}{n}\})\geq 1-e^{-t};$$

3. *If certain sub-matrices of Σ has a bounded condition number that $\kappa(\Sigma_S)\leq M_s$ where $S\subset\{1,\ldots,n\}$ such that S has cardinality s (the sparsity level of β), then*
$$P(||\hat{\Sigma}-\Sigma||_{2,s}\lesssim\sqrt{\frac{s\log d}{n}})\geq 1-p^{-s}$$

In addition, we consider the following lemma ([35], [10]) about the convergence of optimal solution to convex programs:

Lemma 9. *If the loss function*

$$L(\boldsymbol{\beta}) = \boldsymbol{\beta}^T \hat{\Sigma}_{\tilde{X}\tilde{X}} \boldsymbol{\beta} - 2\hat{\Sigma}_{\tilde{X}\tilde{Y}}^T \boldsymbol{\beta} + 1$$

satisfies RSC,

$$\delta L(\Delta, \boldsymbol{\beta}) := L(\boldsymbol{\beta} + \Delta) - L(\boldsymbol{\beta}) - \Delta^T (\bigtriangledown(\boldsymbol{\beta}) \geq \kappa_L ||\Delta||_2^2 \tag{B.1}$$

for some $\kappa_L > 0$ *and* $\Delta \in \{\Delta \in \mathbb{R}^p : ||\Delta_{S^c}||_1 \leq \alpha ||\Delta_S||_1, |S| \leq s\}$*, then for* $\lambda \geq ||\bigtriangledown L(\boldsymbol{\beta})||_\infty$*, any optimal solution* $\hat{\boldsymbol{\beta}}(\lambda)$ *to the convex program in (3.9) satisfies*

$$||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_2 \lesssim \sqrt{s}\lambda, \quad ||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_1 \lesssim s\lambda.$$

Therefore, to prove Theorem 2, it is sufficient to verify (B.1). However, by definition of $\delta L(\Delta, \boldsymbol{\beta})$,

$$\begin{aligned}
\delta L(\Delta, \boldsymbol{\beta}) &= L(\boldsymbol{\beta} + \Delta) - L(\boldsymbol{\beta}) - \Delta^T (\bigtriangledown L(\boldsymbol{\beta})) \\
&= (\boldsymbol{\beta} + \Delta)^T \hat{\Sigma}_{\tilde{X}\tilde{X}} (\boldsymbol{\beta} + \Delta) - 2\hat{\Sigma}_{\tilde{X}\tilde{Y}}^T (\boldsymbol{\beta} + \Delta) - \boldsymbol{\beta}^T \hat{\Sigma}_{\tilde{X}\tilde{X}} \boldsymbol{\beta} + 2\hat{\Sigma}_{\tilde{X}\tilde{Y}}^T \boldsymbol{\beta} \\
&\quad - \Delta^T (2\hat{\Sigma}_{\tilde{X}\tilde{X}} \boldsymbol{\beta} - 2\hat{\Sigma}_{\tilde{X}\tilde{Y}}^T) \\
&= \Delta^T \hat{\Sigma}_{\tilde{X}\tilde{X}} \Delta
\end{aligned}$$

hence it boils down to prove $\Delta^T \hat{\Sigma}_{\tilde{X}\tilde{X}} \Delta \geq \kappa_L ||\Delta||_2^2$ for some $\lambda \geq || \bigtriangledown L(\boldsymbol{\beta})||_\infty$. We first consider another lemma that further simplifies the proof, followed by providing a satisfactory lower bound of λ.

Lemma 10 ([51], [10])**.** *Let* $\delta \in (0, \frac{1}{5})$ *and* $k_0 = 3$*. Then there exists a constant* C_0 *independent with* n, p, s *such that* $\tilde{s} = C_0 s$ *and let* $E(\tilde{s}) = \{\omega \in \mathbb{R}^p : ||\omega||_0 = \tilde{s}\}$ *for* $\tilde{s} < d$ *and* $E = \mathbb{R}^p$ *otherwise. If* $\hat{\Sigma}_{\tilde{X}\tilde{X}}$ *satisfies*

$$(1 - \delta)||\omega||_2^2 \leq \omega^T \hat{\Sigma}_{\tilde{X}\tilde{X}} \omega \leq (1 + \delta)||\omega||_2^2$$

for all $\omega \in E(\tilde{s})$*, then for any* $\omega \in \{\theta \in \mathbb{R}^p : ||\theta S^c||_1 \leq \alpha ||\theta_S||_1, |S| \leq s\}$*,*

$$(1 - 5\delta)||\omega||_2^2 \leq \omega^T \hat{\Sigma}_{\tilde{X}\tilde{X}} \omega \leq (1 + \delta)||\omega||_2^2.$$

Therefore it is sufficient to show $\Delta^T\hat{\Sigma}_{\tilde{X}\tilde{X}}\Delta \geq (1-\delta)||\Delta||_2^2$ for $\Delta \in E(\tilde{s})$ and some $\delta \in (0, \frac{1}{5})$. Now recall lemma 8 about the convergence rate of $\hat{\Sigma}$ under s-restricted spectral norm under the assumption that $\kappa(\Sigma_S) \leq M$, with probability at least $1 - p^{-2}$, we have

$$\begin{aligned}
\Delta^T\hat{\Sigma}_{\tilde{X}\tilde{X}}\Delta &= |\Delta^T\Sigma_{\tilde{X}\tilde{X}}\Delta + \Delta^T(\hat{\Sigma}_{\tilde{X}\tilde{X}} - \Sigma_{\tilde{X}\tilde{X}})\Delta| \\
&\geq |\Delta^T\Sigma_{\tilde{X}\tilde{X}}\Delta| - |\Delta^T(\hat{\Sigma}_{\tilde{X}\tilde{X}} - \Sigma_{\tilde{X}\tilde{X}})\Delta| \\
&\geq |\Delta^T\Sigma_{\tilde{X}\tilde{X}}\Delta| - ||\hat{\Sigma}_{\tilde{X}\tilde{X}} - \Sigma_{\tilde{X}\tilde{X}}||_{2,\tilde{s}}||\Delta||_2^2 \\
&\geq |\Delta^T\Sigma_{\tilde{X}\tilde{X}}\Delta| - \sqrt{\frac{C_0 s\log d}{n}}||\Delta||_2^2 \text{lemma 8} \\
&\geq \gamma_1||\Delta||_2^2 - \sqrt{\frac{C_0 s\log d}{n}}||\Delta||_2^2.
\end{aligned}$$

where the second inequality arises from the fact that the spectral norm of a sub-matrix is bounded by the spectral norm of the whole matrix, and the last inequality is obtained under RSC assumption on Σ (Definition 4). Therefore, as $s\log d/n \to 0$, (B.1) holds. We now give the lower bound of λ such that $\lambda \geq ||\bigtriangledown L(\boldsymbol{\beta})||_\infty$. Note that

$$\begin{aligned}
\frac{1}{2}||\bigtriangledown L(\boldsymbol{\beta})||_\infty &= ||\hat{\Sigma}_{\tilde{X}\tilde{X}}\boldsymbol{\beta} - \hat{\Sigma}_{\tilde{X}\tilde{Y}}||_\infty = ||\hat{\Sigma}_{\tilde{X}\tilde{X}}\Sigma_{\tilde{X}\tilde{X}}^{-1}\Sigma_{\tilde{X}\tilde{Y}} - \hat{\Sigma}_{\tilde{X}\tilde{Y}}||_\infty \\
&= ||(\hat{\Sigma}_{\tilde{X}\tilde{X}} - \Sigma_{\tilde{X}\tilde{X}})\Sigma_{\tilde{X}\tilde{X}}^{-1}\Sigma_{\tilde{X}\tilde{Y}} + \Sigma_{\tilde{X}\tilde{Y}} - \hat{\Sigma}_{\tilde{X}\tilde{Y}}||_\infty \\
&= ||(\hat{\Sigma}_{\tilde{X}\tilde{X}} - \Sigma_{\tilde{X}\tilde{X}})\boldsymbol{\beta} + \Sigma_{\tilde{X}\tilde{Y}} - \hat{\Sigma}_{\tilde{X}\tilde{Y}}||_\infty \\
&\leq ||(\hat{\Sigma} - \Sigma)(1, -\boldsymbol{\beta}^T)^T||_\infty \\
&\leq \sup|\hat{\Sigma} - \Sigma|||(1, -\boldsymbol{\beta}^T)^T||_1 \\
&\leq \sqrt{\frac{\log p}{n}}(1 + ||\boldsymbol{\beta}||_1) \leq \sqrt{\frac{\log p}{n}}(1 + \sqrt{s}||\boldsymbol{\beta}||_2) \\
&= \sqrt{\frac{\log p}{n}}(1 + \sqrt{s}||\Sigma_{\tilde{X}\tilde{X}}^{-1}\Sigma_{\tilde{X}\tilde{Y}}||_2) \\
&\leq \sqrt{\frac{\log p}{n}}(1 + \sqrt{s}||\Sigma_{\tilde{X}\tilde{X}}^{-1}||_2||\Sigma_{\tilde{X}\tilde{Y}}||_2) \\
&\leq \sqrt{\frac{s\log p}{n}}M.
\end{aligned}$$

Therefore, with $\lambda \geq 2M\sqrt{\frac{s\log p}{n}}$, we have with probability at least $1-p^{-1}$ that

$$||\hat{\boldsymbol{\beta}}(\lambda)-\boldsymbol{\beta}||_2 \lesssim \sqrt{s}\lambda \lesssim \sqrt{\frac{s\log p}{n}} \quad \text{and} \quad ||\hat{\boldsymbol{\beta}}(\lambda)-\boldsymbol{\beta}||_1 \lesssim s\lambda \lesssim s\sqrt{\frac{\log p}{n}}$$

as $s\log p/n \to 0$. □

B.2 Proof of Theorem 3

Proof. We start with adapting the following lemma from [38] that establishes the concentration rate of $\hat{f_Y}$.

Lemma 11 (Adapted results from [38])**.** *For any $\gamma \in (0,1)$, $t \in I_n$ where*

$$I_n := [f_Y^{-1}(-\sqrt{2\gamma}\log n), f_Y^{-1}(\sqrt{2\gamma}\log n)],$$

it holds for $\hat{f_Y}$ that

$$P(\sup_{t\in I_n}|\hat{f}_i(t)-f_i(t)| \geq \epsilon) \leq 2\exp(-\frac{n^{1-\gamma}}{32\pi^2\gamma\log n}\epsilon^2)+\exp(-\frac{n^{1-\gamma}}{16\pi\gamma\log n}).$$

By Lemma 11 and Boole's inequality, we have

$$P(\max_{i\in\{0,\dots,d\}}|\hat{f}_i(t)-f_i(t)| \geq \epsilon) \leq 2\exp(\log p-\frac{n^{1-\gamma}}{32\pi^2\gamma\log n}\epsilon^2)+\exp(\log p-\frac{n^{1-\gamma}}{16\pi\gamma\log n})$$

for $t \in I_n$, for any $\gamma \in (0,1)$.

By taking $\epsilon = \sqrt{\frac{64\pi^2\gamma\log n\log p}{n^{1-\gamma}}}$, then with probability at least $1-p^{-1}$,

$$\max_{i\in\{0,\dots,p\}}|\hat{f}_i(t)-f_i(t)| \lesssim \sqrt{\frac{\gamma\log n\log p}{n^{1-\gamma}}}$$

Recall that $\max_{i\in\{1,\dots,p\}}F_i(x_i^*) \in (\delta^\star, 1-\delta^\star)$, therefore there exists some constant $M_\star > 0$ such that $\max_{i\in\{0,\dots,p\}}f_i(x_i^\star) = \max_{i\in\{0,\dots,d\}}\Phi^{-1}(F_i(x_i^\star)) < M_\star$. Since

$f_Y(\mu^\star) < M$, by letting $\gamma = \frac{M_\star^2}{\log n}$, we have with probability at least $1 - p^{-1}$,

$$|\hat{f}_Y(f_Y^{-1}(\sum_{i=1}^d \hat{f}_i(x_i^\star)\hat{\beta}(\lambda)_i)) - f_Y(f_Y^{-1}(\sum_{i=1}^d \hat{f}_i(x_i^\star)\hat{\beta}(\lambda)_i))| \lesssim \sqrt{\frac{\log p}{n}} \tag{B.2}$$

Furthermore, we have the following holds

$$\begin{aligned}
&|\mu^\star - \hat{\mu}^\star| \\
&= |\hat{f_Y}^{-1}(\sum_{i=1}^d \hat{f}_i(x_i^\star)\hat{\beta}(\lambda)_i) - f_Y^{-1}(\sum_{i=1}^d f_i(x_i^\star)\beta(\lambda)_i)| \\
&\leq |\hat{f_Y}^{-1}(\sum_{i=1}^d \hat{f}_i(x_i^\star)\hat{\beta}(\lambda)_i) - f_Y{}^{-1}(\sum_{i=1}^d \hat{f}_i(x_i^\star)\hat{\beta}(\lambda)_i)| + |f_Y{}^{-1}(\sum_{i=1}^d \hat{f}_i(x_i^\star)\hat{\beta}(\lambda)_i) - f_Y^{-1}(\sum_{i=1}^d f_i(x_i^\star)\beta(\lambda)_i) \\
&\leq |\hat{f_Y}^{-1}(\sum_{i=1}^d \hat{f}_i(x_i^\star)\hat{\beta}(\lambda)_i) - f_Y{}^{-1}(\sum_{i=1}^d \hat{f}_i(x_i^\star)\hat{\beta}(\lambda)_i)| + \frac{1}{c}|\sum_{i=1}^d \hat{f}_i(x_i^\star)\hat{\beta}(\lambda)_i - \sum_{i=1}^d f_i(x_i^\star)\beta(\lambda)_i| \\
&:= L_1 + L_2
\end{aligned}$$

where the last inequality is due to Lipschitz continuity of f_Y with constant c.

We next look at the two parts L_1 and L_2 respectively.

$$\begin{aligned}
L_2 &= |\sum_{i=1}^d \hat{f}_i(x_i^\star)\hat{\beta}(\lambda)_i - \sum_{i=1}^d f_i(x_i^\star)\beta(\lambda)_i| \\
&\leq |\sum_{i=1}^d \hat{f}_i(x_i^\star)\hat{\beta}(\lambda)_i - \sum_{i=1}^d f_i(x_i^\star)\hat{\beta}(\lambda)_i| + |\sum_{i=1}^d f_i(x_i^\star)\hat{\beta}(\lambda)_i - \sum_{i=1}^d f_i(x_i^\star)\beta(\lambda)_i| \\
&\leq \sum_{i=1}^d |\hat{\beta}(\lambda)_i(\hat{f}_i(x_i^*) - f_i(x_i^*))| + \sum_{i=1}^d |f_i(x_i^\star)| \cdot |\hat{\beta}(\lambda)_i - \beta(\lambda)_i| \\
&\leq \sum_{i=1}^d |\hat{\beta}(\lambda)_i| \cdot \max_{i\in\{0,\ldots,p\}} |\hat{f}_i(t) - f_i(t)| + \sum_{i=1}^d |f_i(x_i^\star)| \cdot ||\hat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}||_1 \\
&= ||\hat{\boldsymbol{\beta}}(\lambda)||_1 \max_{i\in\{0,\ldots,p\}} |\hat{f}_i(t) - f_i(t)| + \sum_{i=1}^d |f_i(x_i^\star)| \cdot ||\hat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}||_1
\end{aligned}$$

(by results of Theorem 2 and the assumption that $\max_{i\in\{0,\ldots,p\}} f_i(x_i^\star) < M_\star$)

$$\begin{aligned}
&\lesssim (||\boldsymbol{\beta}||_1 + s\sqrt{\frac{\log p}{n}}) \max_{i\in\{1,\ldots,p\}} |\hat{f}_i(t) - f_i(t)| + ||\hat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}||_1 \\
&\leq (s||\boldsymbol{\beta}||_2 + s\sqrt{\frac{\log p}{n}}) \max_{i\in\{1,\ldots,p\}} |\hat{f}_i(t) - f_i(t)| + ||\hat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}||_1 \\
&\lesssim s\sqrt{\frac{\log p}{n}}
\end{aligned}$$

Before looking at L_1, we make the following claim to help with analyzing L_1.

Claim 1 (Adapted from [10]). *For two monotonically increasing functions g_1 and g_2, if $|g_1(g_1^{-1}(t)) - g_2(g_1^{-1}(t))| < c_1$ for some $t \in \mathbb{R}$ and $c_1 > 0$, and if g_2 is Lipschitz continuous with constant $c_2 > 0$, namely $|g_2(v_1) - g_2(v_2)| \geq c_2|v_1 - v_2|$, then*

$$|g_1^{-1}(t) - g_2^{-1}(t)| \leq \frac{c_1}{c_2}.$$

This can be proved by contradiction: if $|g_1^{-1}(t) - g_2^{-1}(t)| > \frac{c_1}{c_2}$ then

$$\begin{aligned}
|g_1(g_1^{-1}(t)) - g_2(g_1^{-1}(t))| &= |g_1(g_1^{-1}(t)) - g_2(g_2^{-1}(t)) + g_2(g_2^{-1}(t)) - g_2(g_1^{-1}(t))| \\
&\geq |g_1(g_1^{-1}(t)) - g_2(g_2^{-1}(t))| + |g_2(g_2^{-1}(t)) - g_2(g_1^{-1}(t))| \\
&> c_2 \cdot \frac{c_1}{c_2} + 0 \\
&= c_1
\end{aligned}$$

where contradiction arises.

Using the claim, we can see that

$$\begin{aligned}
L_1 &= |\hat{f_Y}^{-1}(\sum_{i=1}^{d} \hat{f}_i(x_i^\star)\hat{\beta}(\lambda)_i) - {f_Y}^{-1}(\sum_{i=1}^{d} \hat{f}_i(x_i^\star)\hat{\beta}(\lambda)_i)| \\
&\leq |\hat{f_Y}(f_Y^{-1}(\sum_{i=1}^{d} \hat{f}_i(x_i^\star)\hat{\beta}(\lambda)_i)) - f_Y(f_Y^{-1}(\sum_{i=1}^{d} \hat{f}_i(x_i^\star)\hat{\beta}(\lambda)_i))| \\
&\lesssim \sqrt{\frac{\log d}{n}}
\end{aligned}$$

Combining the results, we have

$$L_1 + L_2 \lesssim s\sqrt{\frac{\log d}{n}},$$

hence we complete the proof. □

APPENDIX C

CHAPTER 4 OF APPENDIX

The following proof is largely inspired by the original Scout paper [61].

C.1 Proof to Theorem 4

Proof. Since $q_1 = 2$ and $\lambda_1 > 0$, we want to find $\hat{\mathbf{\Theta}}_{\tilde{X}\tilde{X}}$ that maximizes

$$\log(\det \mathbf{\Theta}_{\tilde{X}\tilde{X}}) - \mathrm{tr}(\hat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}}\mathbf{\Theta}_{\tilde{X}\tilde{X}}) - \|\mathbf{\Theta}_{\tilde{X}\tilde{X}}\|^2. \tag{C.1}$$

Recall that $\frac{d \log\det \mathbf{A}}{d\mathbf{A}} = \mathbf{A}^{-1}$ for $\mathbf{A} \succ 0$ and $\frac{d \,\mathrm{tr}\,\mathbf{AB}}{d\mathbf{B}} = \mathbf{A}$, then assuming $\hat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}} \succ 0$ the derivative of the objective function (C.1) is given by

$$\mathbf{\Theta}^{-1}_{\tilde{X}\tilde{X}} - \hat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}} - 2\lambda_1 \mathbf{\Theta}_{\tilde{X}\tilde{X}}. \tag{C.2}$$

Therefore the solution to (C.1) solves

$$\mathbf{\Theta}^{-1}_{\tilde{X}\tilde{X}} - 2\lambda_1 \mathbf{\Theta}_{\tilde{X}\tilde{X}} = \hat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}}. \tag{C.3}$$

It implies that $\mathbf{\Theta}_{\tilde{X}\tilde{X}}\hat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}} = \hat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}}\mathbf{\Theta}_{\tilde{X}\tilde{X}}$ hence $\hat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}}$ and $\mathbf{\Theta}_{\tilde{X}\tilde{X}}$ share the same eigen-vectors. Then we could write $\mathbf{\Theta}_{\tilde{X}\tilde{X}} = \mathbf{V}\mathbf{D}_{\mathbf{\Theta}}\mathbf{V}^T$ and $\hat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}} = \mathbf{V}\mathbf{D}_{\mathbf{\Sigma}}\mathbf{V}^T$ where $\mathbf{V} \in \mathbb{R}^p$ is the matrix of eigen-vectors and $\mathbf{D}_{\mathbf{\Theta}}, \mathbf{D}_{\mathbf{\Sigma}}$ are $p \times p$ diagonal matrices of the corresponding eigen-values of $\mathbf{\Theta}_{\tilde{X}\tilde{X}}$ and $\hat{\mathbf{\Sigma}}_{\tilde{X}\tilde{X}}$, respectively. Hence (C.3) is equivalent to

$$\frac{1}{(\mathbf{D}_{\mathbf{\Theta}})_{ii}} - 2\lambda_1 (\mathbf{D}_{\mathbf{\Theta}})_{ii} = (\mathbf{D}_{\mathbf{\Sigma}})_{ii} \tag{C.4}$$

which can be exactly solved by $(\mathbf{D}_{\mathbf{\Theta}})_{ii} = \Big(-(\mathbf{D}_{\mathbf{\Sigma}})_{ii} + \sqrt{(\mathbf{D}_{\mathbf{\Sigma}})_{ii}^2 + 8\lambda_1}\Big)/4\lambda_1$ which can be rearranged to

$$(\mathbf{D}_{\mathbf{\Theta}})_{ii} = \frac{2}{(\mathbf{D}_{\mathbf{\Sigma}})_{ii} + \sqrt{(\mathbf{D}_{\mathbf{\Sigma}})_{ii}^2 + 8\lambda_1}}. \tag{C.5}$$

And the inverse of $\mathbf{\Theta}_{\tilde{X}\tilde{X}}$ can then be obtained after simply taking the inverse of $(\mathbf{D_\Theta})_{ii}$:

$$\mathbf{\Theta}^{-1}_{\tilde{X}\tilde{X}} = \mathbf{V}\mathbf{D}^{-\mathbf{1}}_{\mathbf{\Theta}}\mathbf{V}^T,$$

where D^{-1}_{Θ} is a $p \times p$ diagonal matrix with the i-th diagonal entry as $\frac{1}{2}\Bigg((\mathbf{D_\Sigma})_{ii} + \sqrt{(\mathbf{D_\Sigma})^2_{ii} + 8\lambda_1}\Bigg)$, hence we complete the proof.

□

BIBLIOGRAPHY

[1] Martin C. Abba, Martin C. Abba, Jeffrey A. Drake, Jeffrey A. Drake, Kathleen A. Hawkins, Kathleen A. Hawkins, Yuhui Hu, Yuhui Hu, Hongxia Sun, Hongxia Sun, Cintia Notcovich, Cintia Notcovich, Sally Gaddis, Sally Gaddis, Aysegul Sahin, Aysegul Sahin, Keith Baggerly, Keith Baggerly, C. M. Aldaz, and C. M. Aldaz. Transcriptomic changes in human breast cancer progression as determined by serial analysis of gene expression. *Breast Cancer Research : BCR*, 6(5):R499–513;R513;, 2004.

[2] Philip Adler, Casey Falk, Sorelle A Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. Auditing black-box models for indirect influence. *Knowledge and Information Systems*, 54(1):95–122, 2018.

[3] A. Agresti. *Analysis of Ordinal Categorical Data*. John Wiley & Sons, Inc., 2010.

[4] D. F. Andrews. *Data : A Collection of Problems from Many Fields for the Student and Research Worker*. Springer, New York, 1985.

[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 2016.

[6] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9(Mar):485–516, 2008.

[7] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: the state of the art. *arXiv preprint arXiv:1703.09207*, 2017.

[8] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009;2008;.

[9] D. P. Byar and S. B. Green. The choice of treatment for cancer patients based on covariate information: Application to prostate cancer. *Bulletin du Cancer*, 67(4):477–490, 1980.

[10] Tony T. Cai and Linjun Zhang. High-dimensional gaussian copula regression: Adaptive estimation and statistical inference. *Statistica Sinica*, 28(2):963–993, 2018.

[11] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.

[12] Harald Cramer. *Mathematical Methods of Statistics*. Princeton University Press, 1946.

[13] William Dieterich, Christina Mendoza, and Tim Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpoint Inc*, 2016.

[14] Jianqing Fan, Han Liu, Yang Ning, and Hui Zou. High dimensional semiparametric latent graphical model for mixed data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):405–421, 2017.

[15] Jianqing Fan, Lingzhou Xue, and Hui Zou. Multitask quantile regression under the transnormal model. *Journal of the American Statistical Association*, 111(516):1726–1735, 2016.

[16] Wei Fang, Ziying Wang, Quanxin Li, Xiaojie Wang, Yan Zhang, Yu Sun, Wei Tang, Chunhong Ma, Jinpeng Sun, Ningjun Li, et al. Gpr97 exacerbates aki by mediating sema3a signaling. *Journal of the American Society of Nephrology*, 29(5):1475–1489, 2018.

[17] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[18] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[19] Bert Gold, Tomas Kirchhoff, Stefan Stefanov, James Lautenberger, Agnes Viale, Judy Garber, Eitan Friedman, Steven Narod, Adam B Olshen, Peter Gregersen, et al. Genome-wide association study provides evidence for a breast cancer risk locus at 6q22. 33. *Proceedings of the National Academy of Sciences*, 105(11):4340–4345, 2008.

[20] Leo A. Goodman and William H. Kruskal. Measures of association for cross classifications. *Journal of the American Statistical Association*, 49(268):732–764, 1954.

[21] Yaqian Guo, Trevor Hastie, and Robert Tibshirani. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, 8(1):86–100, 2006.

[22] Fang Han and Han Liu. Statistical analysis of latent generalized correlation matrix estimation in transelliptical distribution. *Bernoulli*, 23(1):23–57, 2017;2013;.

[23] Christian Hennig and Tim F Liao. How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *Journal*

of the Royal Statistical Society: Series C (Applied Statistics), 62(3):309–369, 2013.

[24] Manuela Hummel, Dominic Edelmann, and Annette Kopp-Schneider. Clustering of samples and variables with mixed-type data. *PloS One*, 12(11):e0188274, 2017.

[25] Lynette Hunt and Murray Jorgensen. Mixture model clustering using the multimix program. *Australian & New Zealand Journal of Statistics*, 41(2):154–171, 1999.

[26] Bochao Jia and Faming Liang. Learning gene regulatory networks with high-dimensional heterogeneous data. *arXiv preprint arXiv:1805.02547*, 2018.

[27] Lei Jin, Wei-Ren Liu, Meng-Xin Tian, Xi-Fei Jiang, Han Wang, Pei-Yun Zhou, Zhen-Bin Ding, Yuan-Fei Peng, Zhi Dai, Shuang-Jian Qiu, et al. Ccl24 contributes to hcc malignancy via rhob-vegfa-vegfr2 angiogenesis pathway and indicates poor prognosis. *Oncotarget*, 8(3):5135, 2017.

[28] James E Johndrow and Kristian Lum. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *arXiv preprint arXiv:1703.04957*, 2017.

[29] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81, 1938.

[30] M. G. Kendall. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251, 1945.

[31] M.G. Kendall. *Rank Correlation Methods*. C. Griffin, 1948.

[32] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[33] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm. *ProPublica (2016)*, 9, 2016.

[34] Gunhee Lee and Minho Lee. Classification of genes based on age-related differential expression in breast cancer. *Genomics & Informatics*, 15(4):156, 2017.

[35] Jason D. Lee, Yuekai Sun, and Jonathan E. Taylor. On model selection consistency of regularized m-estimators. *Electronic Journal of Statistics*, 9(1):608–642, 2015.

[36] Han Liu, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman. High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, 2012.

[37] Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10:2295–2328, 2009.

[38] Qing Mai and Hui Zou. Sparse semiparametric discriminant analysis. *Journal of Multivariate Analysis*, 135:175–188, 2015.

[39] J. T Kent Mardia, K. V and J. M Bibby. *Multivariate Analysis*. Academic Press, London, 1979.

[40] Rahul Mazumder and Trevor Hastie. The graphical lasso: New insights and alternatives. *Electronic Journal of Statistics*, 6:2125, 2012.

[41] Damien McParland and Isobel C. Gormley. Model based clustering for mixed data: clustmd. *Advances in Data Analysis and Classification*, 10(2):155–169, 2016.

[42] Karin Milde-Langosch, Holger Kappes, Sabine Riethdorf, Thomas Löning, and Ana-Maria Bamberger. Fosb is highly expressed in normal mammary epithelia, but down-regulated in poorly differentiated breast carcinomas. *Breast Cancer Research and Treatment*, 77(3):265–275, 2003.

[43] Tomoyuki Mukai, Shunichi Fujita, and Yoshitaka Morita. Tankyrase (parp5) inhibition induces bone loss through accumulation of its substrate sh3bp2. *Cells*, 8(2), 2019.

[44] Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012;2010;.

[45] Aleix Prat, Cristina Cruz, Katherine A. Hoadley, Orland Díez, Charles M. Perou, and Judith Balmaña. Molecular features of the basal-like breast cancer subtype based on brca1 mutation status. *Breast Cancer Research and Treatment*, 147(1):185–191, Aug 2014.

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

[47] Sophia Rabe-Hesketh and Anders Skrondal. Latent variable modelling: A survey. *Scandinavian Journal of Statistics*, 34(4):712–745, 2007.

[48] Peter Radchenko. High dimensional single index models. *Journal of Multivariate Analysis*, 139:266–282, 2015.

[49] Sridhar Ramaswamy, Pablo Tamayo, Ryan Rifkin, Sayan Mukherjee, Chen-Hsiang Yeang, Michael Angelo, Christine Ladd, Michael Reich, Eva Latulippe, Jill P Mesirov, et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences*, 98(26):15149–15154, 2001.

[50] Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.

[51] M. Rudelson and Shuheng Zhou. Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory*, 59(6):3434–3447, 2013.

[52] Ayesha N Shajahan-Haq, Simina M Boca, Lu Jin, Krithika Bhuvaneshwar, Yuriy Gusev, Amrita K Cheema, Diane D Demas, Kristopher S Raghavan, Ryan Michalek, Subha Madhavan, and Robert Clarke. Egr1 regulates cellular metabolism and survival in endocrine resistant breast cancer. *Oncotarget*, 8(57):96865–96884, 2017.

[53] Robert H. Somers. A new asymmetric measure of association for ordinal variables. *American Sociological Review*, 27(6):799–811, 1962.

[54] Daniel E. Stange, Bernhard Radlwimmer, Falk Schubert, Frank Traub, Andreas Pich, Grischa Toedt, Frank Mendrzyk, Ulrich Lehmann, Roland Eils, Hans Kreipe, and Peter Lichter. High-resolution genomic profiling reveals association of chromosomal aberrations on 1q and 16p with histologic and genetic subgroups of invasive breast cancer. *Clinical Cancer Research*, 12(2):345–352, 2006.

[55] Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. Detecting bias in black-box models using transparent model distillation. *arXiv preprint arXiv:1710.06169*, 2017.

[56] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[57] YL Tong. *The Multivariate Normal Distribution*. Springer, New York, 1990.

[58] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

[59] Marten Wegkamp and Yue Zhao. Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. *Bernoulli*, 22(2):1184–1226, 2016;2013;.

[60] Daniela M Witten, Jerome H Friedman, and Noah Simon. New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, 20(4):892–900, 2011.

[61] Daniela M Witten and Robert Tibshirani. Covariance-regularized regression and classification for high dimensional problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):615–636, 2009.

[62] Jie Wu, Maolan Li, and Yijian Zhang. Long noncoding rna hoxa-as2 regulates the expression of scn3a by sponging mir-106a in breast cancer. *Journal of Cellular Biochemistry*, 120(9):14465–14475, 2019.

[63] Yilin Xie, Yaqing Liu, Xiaoyue Fan, Lan Zhang, Qing Li, Shenglei Li, Honglei Wang, and Yi Xiao. Microrna-21 promotes progression of breast cancer via

inhibition of mitogen-activated protein kinase10 (mapk10). *Bioscience reports*, 2019.

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

[65] Grace Yoon, Raymond J. Carroll, and Irina Gaynanova. Sparse semiparametric canonical correlation analysis for data of mixed types, 2018.

[66] Fangfang Zhou, Yvette bsch, Tim J. A. Dekker, Amaya G. de Vinuesa, Yihao Li, Lukas J. A. C Hawinkels, Kelly-Ann Sheppard, Marie-José Goumans, Rodney B. Luwor, Carlie J. de Vries, Wilma E. Mesker, Rob A. E. M Tollenaar, Peter Devilee, Chris X. Lu, Hongjian Zhu, Long Zhang, and Peter T. Dijke. Nuclear receptor nr4a1 promotes breast cancer invasion and metastasis by activating tgf- signalling. *Nature Communications*, 5(1):3388–13 p., 2014.

[67] Yichen Zhou, Zhengze Zhou, and Giles Hooker. Approximation trees: Statistical stability in model distillation. *arXiv preprint arXiv:1808.07573*, 2018.

[68] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.