HIGH-DIMENSIONAL DATA ANALYSIS PROBLEMS IN INFECTIOUS DISEASE STUDIES

Yutong Liu

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2022

Approved by:

Quefeng Li

Joseph G. Ibrahim

Feng-Chang Lin

Xiaojing Zheng

Fei Zou

ABSTRACT

Yutong Liu: High-dimensional Data Analysis Problems in Infectious Disease Studies
(Under the direction of Quefeng Li)

Recent technological developments give researchers the opportunity to obtain large informative datasets when studying infectious disease. Such datasets are often high-dimensional, which presents challenges for classical multivariate analysis methods. It is critical to develop novel methods that can solve problems arising in infectious disease studies when the data is high-dimensional or has complex structure.

In the first project, we focus on a *Plasmodium vivax* malaria infection study. A standard competing risks set-up requires both time-to-event and cause-of-failure to be fully observable for all subjects. However, in practice, the cause of failure may not be observable, thus impeding the risk assessment. When a recurrent episode of *Plasmodium vivax* malaria happens following treatment, the patient may have suffered a relapse from a previous infection or acquired a new infection from a mosquito bite. In this case, the time to relapse cannot be modeled when a competing risk, a new infection, is present. Therefore, we developed a novel method for classifying the latent cause of failure under a competing risks set-up, which uses not only time to event information but also transition likelihoods between covariates at the baseline and at the time of event occurrence. Our classifier shows superior performance under various scenarios in simulation experiments. The method was applied to *Plasmodium vivax* infection data to classify recurrent infections of malaria.

In the second project, we investigate data collected from a *Chlamydia trachomatis* genital tract infection study, which contains data of mixed types from multiple groups of subjects. To handle mixed type data, we propose a Latent Mixed Gaussian Copula model that can quantify the correlations among binary, categorical, continuous, and truncated variables in a unified framework. We also provide a tool to decompose the variation into the group-specific and the common variation over multiple groups via solving a regularized M-estimation problem. We conduct extensive simulation studies to show the advantage of our proposed method. We also demonstrate that by jointly solving the M-estimation problem over multiple groups, our method is better than decomposing the variation group-by-group. We apply our method to the *Chlamydia trachomatis* genital

tract infection study to demonstrate how it can be used to discover informative biomarkers that differentiate patients.

For data collected from the *Chlamydia trachomatis* genital tract infection study, not all subjects have complete data from all data modalities, resulting in a block-wise missing structure of the mixed type data. To utilize as much data as possible, we propose to impute the missing values by the Latent Mixed Gaussian Copula model in the third project, where we perform imputation for block-wise missing values by the underlying correlations between fully observed and partially observed variables. The method proposed can be applied to multi-modal data with various data types and shows superior performance for imputing the mixed type data in simulation experiments. We applied the method to data from the *Chlamydia trachomatis* genital tract infection study for imputation of missing endometrial infection status, endometrial diagnosis results, and truncated cytokine values.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Infectious diseases have always been an area of focus when it comes to public health. The ongoing COVID-19 pandemic, only emerging at the end of 2019, has posed severe challenges to population health globally. Even before the COVID-19 pandemic, infectious diseases like HIV, tuberculosis, malaria, and hepatits have been one of the leading causes of death and disability (WHO, 2019*a*).

Recent technological developments like quantitative PCR and targeted deep sequencing make it possible for researchers to obtain large scale genetic data from patients when studying infectious disease. However, the nature of genetic data proclaims that, in such datasets, the number of variables will always be larger than the number of patients, and therefore leads to a high-dimensional dataset. For this kind of high-dimensional dataset, classical multivariate analysis methods designed for low-dimensional can not be applied directly. To better study infectious disease, it is critical to develop novel methods that can be used when the data is high-dimensional or of complex structure.

In Chapter 3, we developed a method to classify the cause of *Plasmodium vivax* malaria recurrence from a competing risks perspective. A standard competing risks set-up requires both time to event and cause of failure to be fully observable for all subjects. However, in application, the cause of failure may not always be observable, thus impeding the risk assessment. In some extreme cases, none of the causes of failure is observable. *Plasmodium vivax*, is the most widespread human malaria (Howes et al., 2016). Due to the dormant liver stage of *P. vivax*, *hypnozoites* may reactivate and cause another infection weeks to months after the initial infection (Chu and White, 2016). However, the fact that individuals can also become reinfected due to a new mosquito bite makes it difficult to study the anti-relapse efficacy of treatment. In the case of a recurrent episode of *Plasmodium vivax* malaria following treatment, the patient may have suffered a relapse from a previous infection or acquired a new infection from a mosquito bite. Therefore, the time to relapse cannot be modeled when a competing risk, a new infection, is present. The efficacy of a treatment for preventing relapse from a previous infection may be underestimated when the true cause of infection cannot be classified. To solve this problem, we developed a novel method for classifying the latent cause of failure under a competing risks set-up, which uses not only time to event information but also

transition likelihoods between covariates at the baseline and at the time of event occurrence. Our classifier shows superior performance under various scenarios in simulation experiments. The method was applied to *Plasmodium vivax* infection data to classify recurrent infections of malaria.

Chlamydia is the leading bacterial sexually transmitted infection in the United States and the infection is often asymptomatic. In up to 50% of women, untreated infection can ascend from the cervix to the upper genital tract and potentially lead to severe female reproductive morbidities. In Chapter 4, we look at data collected from a *Chlamydia trachomatis* genital tract infection study. In this study, both gene expression data and SNP data were collected at the same time from two groups of subjects, one group of subjects with cervical infections only, while the other group of subjects have both cervical infections and endometrial infections. Identification of the commonly and differentially expressed genes and their underlying regulatory SNPs between women with and without ascending infection can greatly enhance the understanding of disease. There are many biomedical studies collect data of mixed types of variables from multiple groups of subjects. Some of these studies aim to find the group-specific and the common variation among all these variables. Even though similar problems have been studied by some previous works, their methods mainly rely on the Pearson correlation, which cannot handle mixed data. To address this issue, we propose a Latent Mixed Gaussian Copula model that can quantify the correlations among binary, categorical, continuous, and truncated variables in a unified framework. We also provide a tool to decompose the variation into the group-specific and the common variation over multiple groups via solving a regularized M-estimation problem. We conduct extensive simulation studies to show the advantage of our proposed method over the Pearson correlation-based methods. We also demonstrate that by jointly solving the M-estimation problem over multiple groups, our method is better than decomposing the variation group-by-group. We apply our method to a *Chlamydia trachomatis* genital tract infection study to demonstrate how it can be used to discover informative biomarkers that differentiate patients.

Even though the method proposed in Chapter 4 can handle mixed type data and perform variance decomposition simultaneously, it requires all subjects to have complete data from all data modalities. Subjects with element-wise missing values, or missing values from one or more data modalities, referred to as block-wise missingness, will be excluded. However, by excluding subjects with any missing values, we are also removing valuable information from the dataset. To be able to utilize as much data as possible when the mixed type data has a element-wise or block-wise missing structure, we propose to impute the missing values by a Latent Mixed Gaussian Copula model in Chapter 5. A Winsorized empirical cumulative distribution

function estimator will be used for estimating the transformation functions of the observed variables, which will then be used for imputing the values of the latent layer. By the conditional distribution of the latent variables, we can impute all the element-wise and block-wise missing values. The method proposed can be applied to mixed type data regardless of the missing mechanism.

LITERATURE REVIEW

2.1 Competing Risks with Missing Cause of Failure

It is commonly seen in biomedical research that the occurrence of an event during the follow-up period can be attributed to one of multiple causes. Data of this type is a standard competing risks set-up, where one event occurs per subject, and the failure type is one of many possible causes. Usually, both time to event and the cause of failure are observable. However, in some cases, the cause of failure may be unknown or missing.

The problem of missing cause of failure in competing risks data has been given much attention since Dinse (1982). There are two possible approaches for estimating competing risks data with missing cause of failure when the cause is missing at random (Rubin, 1976): (1) complete-case analysis, utilizing only complete observations, e.g., Effraimidis and Dahl (2014), or, (2) construct a regression model for the missing cause using all observations, including those with missing cause of failure. In the second approach, one can use a global parametric model (Lu and Tsiatis, 2001), a semi-parametric framework (Goetghebeur and Ryan, 1995) or a nonparametric regression method (Gouskova, Lin and Fine, 2017) to estimate the cause-specific hazard functions. A similar problem is also considered in Sun and Gilbert (2012) and Juraska and Gilbert (2016) when considering the competing cause as a mark for the mark-specific hazard function. A doubly robust estimator is proposed in these papers when the mark variable is possibly missing. However, these approaches require at least some of the observations to have complete records. They cannot be applied to studies where the cause of failure is unknown for every subject.

2.2 Variation Decomposition of Mixed Variables

2.2.1 Differential Gene Co-expression Analysis

In gene co-expression network studies, people aim to find differentially expressed genes or pathways across groups of subjects with different phenotypes. A lot of methods have been developed for differential gene co-expression analysis (van Dam et al., 2018), which aims to reveal regulatory genes corresponding to

different phenotypes. Watson (2006) developed CoXpress, where they used hierarchical cluster analysis to identify groups of genes that are differentially co-expressed under different phenotypes. Choi and Kendziorski (2009) introduced GSCA to test whether some given gene sets are differentially coexpressed between different groups. Tesson, Breitling and Jansen (2010) presented DiffCoEx, a method that was built on the Weighted Gene Coexpression Network Analysis (WGCNA, Langfelder and Horvath, 2008) framework for identifying gene modules that are differentially coexpressed between two groups. Amar, Safer and Shamir (2013) developed DICER, where a probabilistic score was used for detecting a group of genes that are coexpressed differently for disease and normal samples. Rahmatallah, Emmert-Streib and Glazko (2014) introduced GSCNA, which assign a weight factor for each gene in the set and provides not only a method to test whether a group of genes are coexpressed differently within different group but also evaluate the importance of genes in gene sets. However, this method also requires predefined gene sets. The methods mentioned above focus on finding the group-specific, or say, phenotype-specific structure across groups of different phenotypes but did not consider the fact that even for groups with different phenotypes, there will be some shared information across groups. To account for the shared information across groups, Alter, Brown and Botstein (2003) developed GSVD, which use generalized singular value decomposition to formulate expression data as a sum of effects of genes that are shared for both datasets and effects that are unique for each dataset. GSVD can only deal with the situation when there are two datasets. Ponnapalli et al. (2011) later defined a higher-order GSVD that can deal with three or more datasets. Ha, Baladandayuthapani and Do (2015) developed DINGO, which model the conditional dependencies among genes through a Gaussian graphical model and decompose them into a global and group-specific components.

2.2.2 Latent Gaussian Copula Model

2.2.2.1 Continuous variables

Let $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ be a p-dimensional vector. Liu, Lafferty and Wasserman (2009) proposed that if there exists monotone and differentiable functions $\boldsymbol{f} = \{f_j\}_{j=1}^p$ such that $\{f_1(X_1), \ldots, f_p(X_p)\}^T \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a correlation matrix, then $\boldsymbol{X}$ is said to follow a Gaussian copula model, denoted by $\boldsymbol{X} \sim \text{NPN}(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{f})$. To estimate the transformation functions f_j, they also proposed a Winsorized empirical

CDF estimator, which is defined as

$$\widetilde{F}_j(t;\delta_n,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) := T_{\delta_n}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{I}(x_{ij}\leq t)\right) \tag{2.1}$$

for X_j, and

$$T_{\delta_n}(a) := \begin{cases} \delta_n, & \text{if } a < \delta_n, \\ a, & \text{if } \delta_n \leq a \leq 1-\delta_n, \\ 1-\delta_n, & \text{if } a > 1-\delta_n. \end{cases}$$

Define $\widetilde{f}_j(t) = \Phi^{-1}(\widetilde{F}_j(t))$ and use the truncation level $\delta_n = \frac{1}{4n^{1/4}\sqrt{\pi \log n}}$, Han, Zhao and Liu (2013) proved that $\widetilde{f}_j(t)$ converges to f_j uniformly over an expanding interval with high probability.

2.2.2.2 Binary and truncated variables

When the observed data $\boldsymbol{X}$ are not continuous variables, Fan et al. (2017) studies the Latent Gaussian copula model for binary data. That is, the random vector $\boldsymbol{X} \in \mathbb{R}^p$ that takes value 0 or 1 satisfies the binary latent Gaussian copula model, if there exists a p-dimensional random vector $\boldsymbol{Y} = (Y_1,\ldots,Y_p)^T \sim \text{NPN}(0,\boldsymbol{\Sigma},f)$ such that $X_j = I(Y_j > C_j)$ for all $j = 1,\ldots,p$, where $I(\cdot)$ is the indicator function and $\boldsymbol{C} = (C_1,\ldots,C_p)$ is a vector of constants, then we say the random vector $\boldsymbol{X}$ satisfies the latent Gaussian copula model with $\boldsymbol{\Sigma}$ being the latent correlation matrix and denote $\boldsymbol{X} \sim \text{LNPN}(0,\boldsymbol{\Sigma},f,\boldsymbol{C})$.

Yoon, Carroll and Gaynanova (2020) further extended the idea to truncated data. That is, for a random vector $\boldsymbol{X} \in \mathbb{R}^p$, if there exist a p-dimensional random vector $\boldsymbol{Y} = (Y_1,\ldots,Y_p)^T \sim \text{NPN}(0,\boldsymbol{\Sigma},f)$ such that $X_j = I(Y_j > C_j)Y_j$ for $j = 1,\ldots,p$, where $I(\cdot)$ is the indicator function and $\boldsymbol{C} = (C_1,\ldots,C_p)$ is a vector of positive constants, then we say $\boldsymbol{X}$ satisfies the truncated latent Gaussian copula model with $\boldsymbol{\Sigma}$ being the latent correlation matrix and denote $\boldsymbol{X} \sim \text{TLNPN}(0,\boldsymbol{\Sigma},f,\boldsymbol{C})$.

2.2.2.3 Correlation estimation for mixed data via Kendall's τ

Estimating the latent correlation matrix is a critical problem for the latent Gaussian Copula and Mixed Gaussian Copula models. This problem has been studied by Liu, Lafferty and Wasserman (2009), Fan et al. (2017) and Yoon, Carroll and Gaynanova (2020). They proposed to first calculate the Kendall's τ correlations of observed variables and connect them to the correlations of latent variables via some bridge functions.

In particular, let $\{(X_{ij}, X_{ik})\}_{i=1}^{n}$ be the realizations of the observed variables X_j and X_k, the Kendall's τ correlation between X_j and X_k is defined as

$$\widehat{\tau}_{jk} = \frac{2}{n(n-1)} \sum_{1 \le i < i' \le n} \text{sign}(X_{ij} - X_{i'j}) \, \text{sign}(X_{ik} - X_{i'k}). \tag{2.2}$$

Let $\tau_{jk} = \mathbb{E}(\widehat{\tau}_{jk})$ be the population Kendall's τ. Then, the latent correlation R_{jk} between $f_j(Y_j)$ and $f_k(Y_k)$ is $R_{jk} = F_{jk}^{-1}(\tau_{jk})$, where $F_{jk}(\cdot)$ is a bridge function. We summarize the bridge functions for the pairwise correlations among continuous, binary and truncated variables. These formulae were original derived in Liu, Lafferty and Wasserman (2009), Fan et al. (2017) and Yoon, Carroll and Gaynanova (2020).

Theorem 1 *(Liu, Lafferty and Wasserman, 2009; Fan et al., 2017; Yoon, Carroll and Gaynanova, 2020)*

(a) For $j \in \mathscr{C}$ and $k \in \mathscr{C}$, $F_{jk}(R_{jk}) = 2\sin^{-1}(R_{jk})/\pi$.

(b) For $j \in \mathscr{B}$ and $k \in \mathscr{B}$, $F_{jk}(R_{jk}) = 2\Phi_2(\Delta_j, \Delta_k; R_{jk}) - 2\Phi_1(\Delta_j)\Phi_1(\Delta_k)$, where $\Delta_j = f_j(C_j)$ and $\Delta_k = f_k(C_k)$.

(c) For $j \in \mathscr{B}$ and $k \in \mathscr{C}$, $F_{jk}(R_{jk}) = 4\Phi_2(\Delta_j, 0; R_{jk}/\frac{1}{\sqrt{2}}) - 2\Phi_1(\Delta_j)$, where $\Delta_j = f_j(C_j)$.

(d) For $j \in \mathscr{T}$ and $k \in \mathscr{B}$, $F_{jk}(R_{jk}) = 2\big\{1 - \Phi_1(\Delta_j)\big\}\Phi_1(\Delta_k) - 2\Phi_3(-\Delta_j, \Delta_k, 0; \boldsymbol{R}_{3a}) - 2\Phi_3(-\Delta_j, \Delta_k, 0; \boldsymbol{R}_{3b})$, where $\Delta_j = f_j(C_j), \Delta_k = f_k(C_k)$,

$$\boldsymbol{R}_{3a} = \begin{bmatrix} 1 & -R_{jk} & \frac{1}{\sqrt{2}} \\ -R_{jk} & 1 & -\frac{R_{jk}}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{R_{jk}}{\sqrt{2}} & 1 \end{bmatrix} \text{ and } \boldsymbol{R}_{3b} = \begin{bmatrix} 1 & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 1 & -\frac{R_{jk}}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & -\frac{R_{jk}}{\sqrt{2}} & 1 \end{bmatrix}.$$

(e) For $j \in \mathscr{T}$ and $k \in \mathscr{C}$, $F_{jk}(R_{jk}) = -2\Phi_2(-\Delta_j, 0; 1/\sqrt{2}) + 4\Phi_3(-\Delta_j, 0, 0; \boldsymbol{R}_{3c})$, where $\Delta_j = f_j(C_j)$ and

$$\boldsymbol{R}_{3c} = \begin{bmatrix} 1 & \frac{1}{\sqrt{2}} & \frac{R_{jk}}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 1 & R_{jk} \\ \frac{R_{jk}}{\sqrt{2}} & R_{jk} & 1 \end{bmatrix}.$$

(f) For $j \in \mathscr{T}$ and $k \in \mathscr{T}$, $F_{jk}(R_{jk}) = -2\Phi_4(-\Delta_j, -\Delta_k, 0, 0; \boldsymbol{R}_{4a}) + 2\Phi_4(-\Delta_j, -\Delta_k, 0, 0; \boldsymbol{R}_{4b})$, where $\Delta_j = f_j(C_j), \Delta_k = f_k(C_k)$,

$$\boldsymbol{R}_{4a} = \begin{bmatrix} 1 & 0 & \frac{1}{\sqrt{2}} & -\frac{R_{jk}}{\sqrt{2}} \\ 0 & 1 & -\frac{R_{jk}}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{R_{jk}}{\sqrt{2}} & 1 & -R_{jk} \\ -\frac{R_{jk}}{\sqrt{2}} & \frac{1}{\sqrt{2}} & -R_{jk} & 1 \end{bmatrix} \text{ and } \boldsymbol{R}_{4b} = \begin{bmatrix} 1 & R_{jk} & \frac{1}{\sqrt{2}} & \frac{R_{jk}}{\sqrt{2}} \\ R_{jk} & 1 & \frac{R_{jk}}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{R_{jk}}{\sqrt{2}} & 1 & R_{jk} \\ \frac{R_{jk}}{\sqrt{2}} & \frac{1}{\sqrt{2}} & R_{jk} & 1 \end{bmatrix}.$$

It was proved in Liu, Lafferty and Wasserman (2009), Fan et al. (2017) and Yoon, Carroll and Gaynanova (2020) that all these bridge functions are strictly increasing for any $R_{jk} \in (-1, 1)$. Thus, they are invertible. In practice, we can estimate R_{jk} by $\widehat{R}_{jk} = F_{jk}^{-1}(\widehat{\tau}_{jk})$. For a binary or truncated variable, $\Delta_k = f_k(C_k)$ is unknown. To estimate it, Fan et al. (2017) proposed to use the plug-in estimator $\widehat{\Delta}_k = \Phi^{-1}(\sum_{i=1}^n I(X_{ik} \neq 0)/n)$.

2.2.3 Variation Decomposition for Observations from One Population

The problem of variation decomposition can be approached from several different perspectives, including Principal Component Analysis (PCA) (Lock et al., 2013; Zhou et al., 2015; Feng et al., 2018), Canonical Correlation Analysis (CCA) (Shu, Wang and Zhu, 2020), and Partial Least Squares (PLS) (Löfstedt and Trygg, 2011). Lock et al. (2013) introduced the Joint and Individual Variation Explained (JIVE) method that can capture the joint variation across different data types and the individual variation of each data type. Feng et al. (2018) developed AJIVE, where score subspaces were used to ensure an identifiable decomposition. Zhou et al. (2015) proposed COBE for efficient extraction of common and individual features, where they used a low-rank approximation to decompose the data into a shared common subspace and many individual subspaces. Shu, Wang and Zhu (2020) proposed D-CCA, a decomposition-based canonical correlation analysis method. Instead of using the Euclidean space, D-CCA defines the common and unique parts using a more general Hilbert space. Löfstedt and Trygg (2011) derived OnPLS to separate the shared and specific variations. However, these methods are designed only for continuous variable, and cannot be directly applied to other variables, such as binary, categorical or truncated variables. To carry out integrative analysis for data of different types and decompose the data into shared and individual structures, Li, Gaynanova et al. (2018) developed the Generalized Association Study (GAS), which uses the log-likelihood function to integrate

different variables that follow exponential family distributions. Zhu, Li and Lock (2020) generalized the idea of GAS and proposed a Generalized integrative PCA method, which can be used to analyze more than two data sets. It also allows data to have block-wise missing values. However, all these methods focused on finding the similarities and differences among variables collected from one population and thus cannot be used to decompose the variation of two subpopulations.

2.3 High-dimensional Block-wise and Element-wise Missing Data Imputation

2.3.1 Matrix Completion

The matrix completion problem (Laurent, 2001) aims to recover an unknown large matrix based on a small number of its known entries. One of the most famous application of matrix completion methods is the *Netflix Prize* problem, where a training set of people's existed ratings, usually only a few number of movies from each person, are given, and the goal is to predict people's ratings for all other movies that they never rated before. Such a problem is impossible to solve without additional information or assumptions. However, in many cases, it is reasonable to assume that the matrix to be recovered is low-rank or approximately low-rank, and the problem will become solvable (Hastie, Tibshirani and Wainwright, 2019, Section 7). This assumption makes sense as it is implying that the decision is only driven by a few factors. Assume the matrix we wish to recover is a square $n \times n$ matrix $\boldsymbol{M}$ with rank r. Even though there are n^2 entries in $\boldsymbol{M}$, its degrees of freedom is only $(2n - r)r$ (Candès and Recht, 2009), which is notably much smaller than n^2 when the rank r is small. Following Candès and Tao (2010), let Ω be the set of locations corresponding to the observed entries in $\boldsymbol{M}$, and define $\boldsymbol{Y} = \mathcal{P}_\Omega(\boldsymbol{X})$ as

$$Y_{ij} = \begin{cases} X_{ij}, & (i,j) \in \Omega \\ 0, & \text{otherwise} \end{cases}. \tag{2.3}$$

Intuitively, one would consider recovering matrix by solving the optimization problem

$$\begin{aligned} &\text{minimize} \quad \text{rank}(\boldsymbol{X}) \\ &\text{subject to} \quad \mathcal{P}_\Omega(\boldsymbol{X}) = \mathcal{P}_\Omega(\boldsymbol{M}) \end{aligned} \tag{2.4}$$

where $\boldsymbol{X}$ is the decision variable. However, this problem is NP-hard. Instead, Candès and Recht (2009) considered minimizing the nuclear norm of $\boldsymbol{X}$, defined as $\|\boldsymbol{X}\|_* = \sum_{k=1}^{n} \lambda_k(\boldsymbol{X})$, where $\lambda_k(\boldsymbol{X})$ denote the kth largest singular value of $\boldsymbol{X}$ and the nuclear norm is calculated as the sum of the singular values over the constraint set. The optimization problem then becomes

$$\begin{aligned} &\text{minimize} \quad \|\boldsymbol{X}\|_* \\ &\text{subject to} \quad \mathcal{P}_\Omega(\boldsymbol{X}) = \mathcal{P}_\Omega(\boldsymbol{M}) \end{aligned} \tag{2.5}$$

The nuclear norm is a convex relaxation of the rank of a matrix, and therefore, (2.5) is a convex problem. Candès and Recht (2009) proved that if $\boldsymbol{M}$ has row and column spaces that are incoherent with the standard basis, then $\boldsymbol{M}$ can be recovered from a random sampling of a small number of entries using nuclear norm minimization if Ω is random and the number of entries in $\boldsymbol{M}$ observed is greater than $Cn^{6/5} r \log n$. Candès and Tao (2010) then quantitatively improved condition of observed entries to greater than $C\mu^4 n(\log n)^2$, given that $\boldsymbol{M}$ obeys the strong incoherence property with parameter $\mu = \mathcal{O}(\sqrt{\log n})$. Gross (2011) later showed that the exact recovery can be achieved by nuclear norm optimization if the number of observed entries is greater than $Crn \log n$. To efficiently solve (2.5), Cai, Candès and Shen (2010) developed the singular value thresholding algorithm, which iteratively perform soft-thresholding on the singular value of the matrix obtained from the previous step and can handle very large scale problems. Mazumder, Hastie and Tibshirani (2010) rewrite (2.5) in Lagrange form as

$$\min_{\boldsymbol{X}} \frac{1}{2} \|\mathcal{P}_\Omega(\boldsymbol{M}) - \mathcal{P}_\Omega(\boldsymbol{X})\|_F^2 + \lambda \|\boldsymbol{X}\|_* \tag{2.6}$$

where λ is a non-negative regularization parameter. To solve the nuclear norm regularized least-square problem (2.6), Mazumder, Hastie and Tibshirani (2010) proposed SOFT-IMPUTE, an algorithm that iteratively solve the (2.6) via SVD and soft-thresholding, where the non-sparse matrix can be written as the sum of a low rank matrix plus a sparse matrix at each iteration.

Other than using nuclean norm minimization, Rennie and Srebro (2005) proposed to use the maximum margin matrix factorization (MMMF) method, where a factor model was used, to recover $\boldsymbol{M}$. Rennie and Srebro (2005) proved that for a $n_1 \times n_2$ rank r matrix $\boldsymbol{M}$ that can be factorized as $\boldsymbol{M} = \boldsymbol{A}\boldsymbol{B}'$, where

$\boldsymbol{A} \in \mathbb{R}^{n_1 \times r}, \boldsymbol{B} \in \mathbb{R}^{r \times n_2}$, the nuclear norm of $\boldsymbol{M}$ is equal to

$$\min_{\boldsymbol{M}=\boldsymbol{A}\boldsymbol{B}'} \frac{1}{2}\big(\|\boldsymbol{A}\|_F^2 + \|\boldsymbol{B}\|_F^2\big). \tag{2.7}$$

They considered the following problem

$$\min_{\boldsymbol{A},\boldsymbol{B}} \frac{1}{2}\|\mathcal{P}_\Omega(\boldsymbol{M} - \boldsymbol{A}\boldsymbol{B}')\|_F^2 + \frac{\lambda}{2}\big(\|\boldsymbol{A}\|_F^2 + \|\boldsymbol{B}\|_F^2\big), \tag{2.8}$$

which is biconvex and the solution agrees with that given by using nuclear norm minimization. Besides MMMF, Keshavan, Montanari and Oh (2010) proposed to solve the following problem

$$\min_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{S}} \|\mathcal{P}_\Omega(\boldsymbol{M}) - \mathcal{P}_\Omega(\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}')\|_F^2 + \lambda\|\boldsymbol{S}\|_F^2, \tag{2.9}$$

where $\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}_r$ and $\boldsymbol{S}$ is an $r \times r$ matrix. They also introduced a gradient descent algorithm for solving this problem, which is a 3-step algorithm that utilize both singular value decomposition and Grassmann manifolds optimization. Hastie et al. (2015) combined the ideas from SOFT-IMPUTE and MMMF and propose to solve for

$$\min_{\boldsymbol{A},\boldsymbol{B}} \frac{1}{2}\|(\widehat{\boldsymbol{M}} - \boldsymbol{A}\boldsymbol{B}')\|_F^2 + \frac{\lambda}{2}\big(\|\boldsymbol{A}\|_F^2 + \|\boldsymbol{B}\|_F^2\big), \tag{2.10}$$

where $\widehat{\boldsymbol{M}} = \big(\mathcal{P}_\Omega(\boldsymbol{M}) - \mathcal{P}_\Omega(\widehat{\boldsymbol{X}})\big) + \widehat{\boldsymbol{X}}$, and $\widehat{\boldsymbol{X}}$ is obtained by soft-thresholding SVD of $\widehat{\boldsymbol{M}}$ from the last iteration. They presented the softImpute-ALS algorithm for solving (2.10) and showed that the solution converges given sufficiently large r.

2.3.2 High-dimensional block-wise missing data imputation

In biomedical research, the block-wise missing structure is very common for high-dimensional multi-modality datasets. There are several popular approaches for solving high-dimensional data problems when a block-wise missing structure is presented. The most straightforward way is to only use data with complete observations and remove those with any missing values. However, in lots of biomedical research, only a small number of observations have data from all modalities. By simply removing observations with any missing values, we would lose a lot of valuable information. Moreover, the estimator obtained will be biased

if the missing mechanism is not missing completely at random, which is usually not the case for block-wise missing data.

Another approach for handling block-wise missing data is to use all data available, without any deletion or imputation. Yuan et al. (2012) developed the method iMSF, where a multi-task sparse learning framework is used, and observations with data from at least one modality can all be included. Xiang et al. (2014) extended the idea of iMSF and proposed a bi-level model, where they performed covariate-level and modality-level analysis at the same time. Yu et al. (2020) introduced DISCOM, where coefficients in the optimal linear prediction were estimated using an extended Lasso-type estimator, based on estimates for covariance matrices among covariates and between the response and covariates. However, DISCOM require the missing mechanism to be missing completely at random.

Finally, there are also works aim to impute this kind of block-wise missing data. Cai, Cai and Zhang (2016) proposed a structured matrix completion (SMC) method based on SVD and the use of Schur complement. They showed that the data matrix can be recovered if the matrix is exactly or approximately low rank. However, SMC requires there to be some complete rows and columns in the data matrix, and that the missing need to be completely at random. Also, the method can only be applied on data with Gaussian distribution, and can not handle more than two data modalities easily. Zhang, Tang and Qu (2020) considered a factor model approach for imputing the missing blocks. The method does not rely on any specific missing mechanism and remains efficient when there are a lot of block-wise missings. However, this approach can only be used for continuous data. To solve the problem of imputation for mixed-type data. Xue and Qu (2020) proposed a multiple block-wise imputation (MBI) approach that can handle mixed type covariate. Zhu, Li and Lock (2020) developed GIPCA, a low rank approach that can impute block-wise missing data of mixed types. For the methods proposed in Xue and Qu (2020) and Zhu, Li and Lock (2020), since they used parametric methods, the distribution of the covariates has to fall in the exponential family.

CHAPTER 3: DYNAMIC CLASSIFICATION OF *PLASMODIUM VIVAX* MALARIA RECURRENCE: AN APPLICATION OF CLASSIFYING UNKNOWN CAUSE OF FAILURE IN COMPETING RISKS

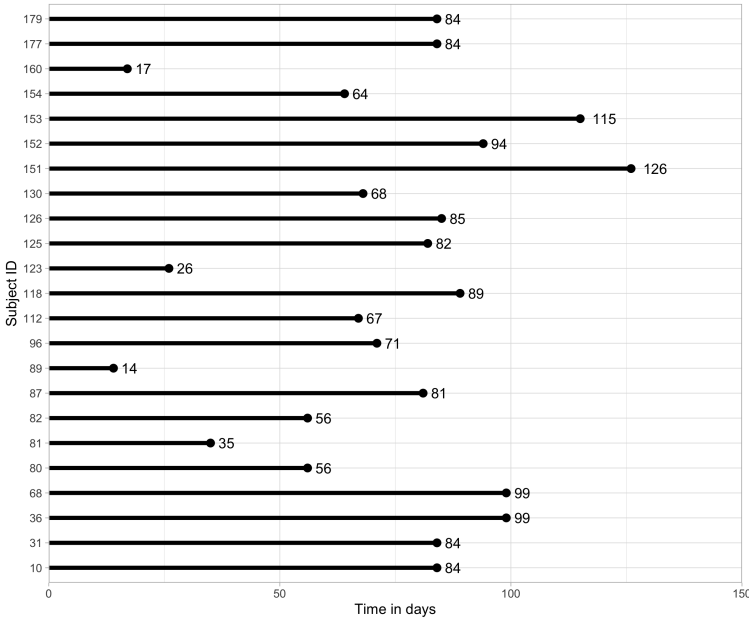
3.1 Introduction

3.1.1 *Plasmodium vivax* Malaria Infection

Plasmodium vivax, in short, *P. vivax*, is the most widespread human malaria (Howes et al., 2016). According to the 2019 World Malaria Report released by World Health Organization (WHO), 53% of the global *P. vivax* burden is in the South-East Asia Region, and 75% of malaria cases in the Region of the Americas are resulted from *P. vivax*. Due to the dormant liver stage of *P. vivax*, *hypnozoites* may reactivate and cause another infection weeks to months after the initial infection (Chu and White, 2016). Relapse due to inadequately treated blood stages is less common and is referred to as treatment failure or recrudescence. Therefore, when first-line antimalarials are used, relapse is usually attributed to *hypnozoite*-induced relapse. *P. vivax* relapses are an important source of morbidity and contribute to malaria mortality (Dini et al. 2020, Robinson et al. 2015, Baird 2013). However, the fact that individuals can also become reinfected due to a new mosquito bite makes it difficult to study the anti-relapse efficacy of treatment. Previous studies have concluded that even when the level of transmission is relatively low, there is a high genetic diversity in *P. vivax* parasites within patient populations in Cambodia (Lin et al., 2013, Friedrich et al., 2016). Such genetic diversity, often resulting in multiple parasites haplotypes present in a single infection, provides an opportunity for researchers to distinguish relapse from a recurrent infection by examining the overlap of haplotypes between infections and the appearance of haplotypes associated with relapse.

Lin et al. (2015) applied targeted deep sequencing to 108 isolates collected from 78 Cambodian volunteers with *P. vivax* infection (Lon et al., 2014). Subjects in the study were treated initially with dihydroartemisinin-piperaquine (DP), an effective drug to treat the blood stages of *P. vivax*, all but precluding treatment failure due to recrudescence. To detect recurrent infection, blood smears of study subjects were taken firstly at baseline, then weekly for six weeks following treatment, then monthly thereafter. At the end of the study,

Figure 3.1: Time to First Recurrence Infection for 23 subjects with recurrence infections.



23 of the 78 subjects experienced recurrent infections, with a median of 68 days in the time to recurrence. Subjects' participation in the study ranged from 2 to 6 months, with a median of 4 months follow-up. Since treatment failure with DP is unlikely, these recurrences most likely represent relapse or reinfection. In fact, of the 23 subjects with recurrent infection, five subjects had a second recurrent infection, and one subject had a third recurrent infection. To simplify the analysis, we only consider the first recurrent infection among those 23 subjects. Panel (a) in Figure 3.2 shows the Kaplan-Meier curve for the first recurrent infection along with the risk table showing number of subjects at risk over ten-day intervals. The horizontal axis in the plot indicates days from baseline, and the vertical axis is the estimated survival probability. The solid line is the step function and shaded area is associated 95% point-wise confidence interval of the step function. The longest follow-up time is 180 days, and 70% (55 subjects) were disease-free at the end of the follow-up period. Figure 3.1 is a subject-by-subject time to first infection plot which shows the 23 subjects' time to first recurrence infection.

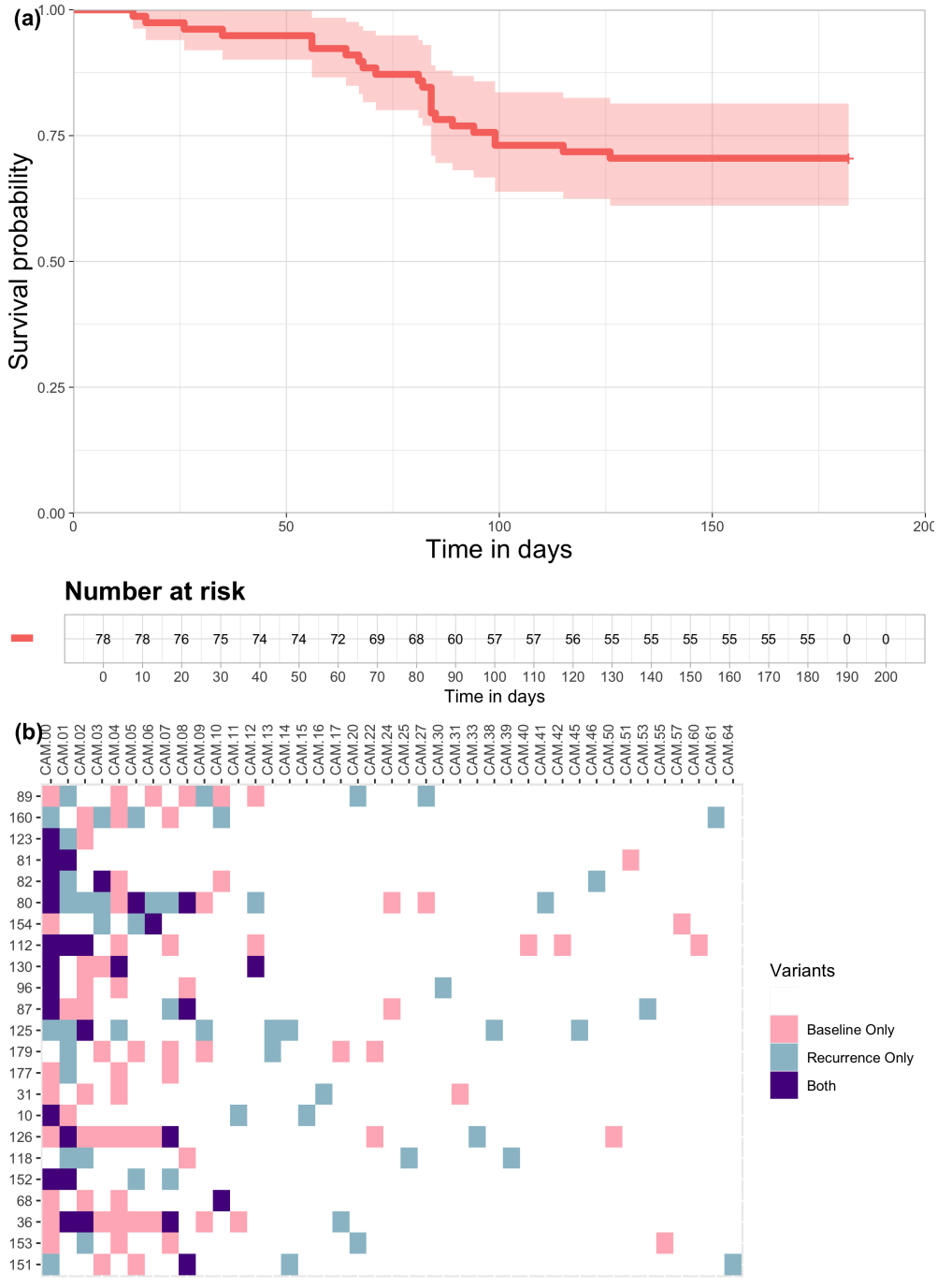
P. vivax exhibits great genetic diversity, surpassing that seen in *P. falciparum* (Neafsey et al., 2012). Parobek et al. (2014) identified a highly variable 117-base pair (bp) segment of the *P. vivax* merozoite surface protein 1 gene (*pvmsp1*) within the 33-kDa subunit of the 42-kDa region, which exhibits great nucleotide diversity. After extracting DNA from filter paper blood spots, Lin et al. (2015) applied deep sequencing to

this region and used a bioinformatics pipeline called *SeekDeep* (Hathaway et al., 2018) to determine different haplotypes of *pvmsp1* defined by at least a single nucleotide difference between haplotypes. They identified 67 unique *pvmsp1* haplotypes across 108 isolates from either initial infection or recurrent infections, with each patient isolate harboring, on average, three different haplotypes. They found nine haplotypes that are common and appeared in at least 10% of individuals. 46 rare haplotypes appeared in only one isolate, with some later attributed to sequencing error. Only 41 unique haplotypes were identified in those subjects with recurrent infection. Panel (b) in Figure 3.2 shows a heatmap that indicates the presence/absence of these 41 haplotypes (genetic variants) in the initial and recurrent infections from those 23 subjects. Each column represents one unique haplotype, and each row represents one subject with an identification number. The subjects were sorted based on their time to the first recurrent infection, with the shortest time at the top and the longest time at the bottom. Pink cells indicate the presence of the haplotype in the initial infection but absence in the recurrent infection. Blue cells show the absence of the haplotype in the initial infection but presence in the recurrent infection. Purple cells show haplotypes that were present in both infections. Interestingly, only 16 subjects had overlapping haplotypes between initial and recurrent infections. Two subjects with the shortest time to recurrent infection did not have any shared haplotypes.

3.1.2 Competing Risks with Unknown Cause of Failure

In *P. vivax* malaria research, subjects who live in endemic areas suffer recurrent infections which can arise from (1) mosquito bites representing new infection, (2) relapse from latent infection in the liver, or (3) recrudescence due to treatment failure. The cause of recurrent infection is unknown or indeterminable in this case, thus impeding the efficacy assessment of anti-relapse treatment. Developing a reliable method to distinguish new infections from relapse is critical. When analyzing the causes of *P. vivax* malaria recurrence from a competing risks perspective, it is natural to assume that the time to recurrent infection is associated with baseline covariates (e.g., genetic variants or haplotypes) collected at the initial infection. We assume that each cause has a distinct cause-specific hazard function conditional on the baseline covariates, enabling us to build an initial cause classifier that can distinguish the cause based on the time to recurrence information. Subsequently, by observing changes in the values of genetic variants between initial and recurrent infections, one can build another classifier that can distinguish the cause of failure, as the changes are driven by the

Figure 3.2: Survival curve and heatmap for presence/absence of haplotypes



latent cause. Thus, one can update the initial classifier by utilizing the information contained in the transition of covariates between initial infection and recurrent infection. To study the transition mechanism, Lin, Li and Lin (2020) proposed an approach that estimates the transition likelihoods using both shared and non-shared genetic variants to improve classification accuracy when the cause of recurrent infection is unknown. Bureau, Shiboski and Hughes (2003) utilized a continuous-time hidden Markov chain to obtain the true transition probabilities between states when the disease status is possibly misclassified. However, Lin, Li and Lin (2020) did not consider the time to recurrent infection, and Bureau, Shiboski and Hughes (2003) required the disease status to be fully observed but subject to misclassification. Neither of these two approaches is ideal for our malaria data, and can not be applied to the classification problem when dealing with competing risks data with missing cause of failure.

In the classification problem with unknown cause of malaria recurrence, Taylor et al. (2019) proposed a Bayesian approach that models the time to recurrent infection for prior classification probability and then computes the posterior probability based on an assumed genetic model with a strong prior assumption. Ferreira et al. (2020) treated relapse (combined with recrudescence) and new infection as competing risks assuming an exponential distribution with a time-constant hazard for both causes. In contrast, we analyze the time to event data under a competing risks set-up without specifying any temporal pattern of the hazard function. We generalize the idea in Lin, Li and Lin (2020) to incorporate the transition likelihoods between covariates to classify the unknown cause of infection. By considering the time to event information and transition likelihoods at the same time, we utilize more information from the data and thus lead to a more accurate classifier. Our method allows the causes of failure to be completely missing and can be applied to *P. vivax* malaria data (Lin et al., 2015). The classification procedure includes two main steps. First, we utilize the time to event and baseline covariates information to obtain an initial classifier. Then, we update the classification probability obtained in the first step using transition likelihoods between covariates to obtain the second classifier, whose performance is better than the first one. The challenges of building these classifiers are that the covariates are high-dimensional, and they can be of different kinds of variables. To resolve the first challenge, we propose a penalized maximum partial likelihood estimator and use an efficient proximal gradient descent algorithm to obtain the estimator. To resolve the second challenge, we propose a general transition likelihood that can incorporate different kinds of variables.

The rest of this chapter is organized as follows. In Section 3.2, we describe the method of modeling competing risk data under a proportional hazards model with baseline covariates. In Section 3.3, we

introduce general formulae for the two classifiers. An algorithm for the computation of parameters needed for constructing the classifiers is laid out in Section 3.4. We carry out comprehensive simulation experiments under various scenarios to evaluate the performance of the proposed classifiers in Section 3.5. Finally, we apply the developed method to the *P. vivax* malaria data and show the classification result in Section 3.6. We summarize our current approach and discuss its extensions in Section 3.7.

3.2 Model and Estimation

In a general setting of competing risks, let T_i^* be the failure time and $\epsilon_i \in \{1, 2\}$ be the cause of failure for subject i. We consider only two causes of failure since this is the most general setting of competing risks application. If there are more than two causes, one may combine causes other than the primary interest into one category and format the model with two causes of failure. To model the time to failure when competing risks are presented, we consider a cause-specific hazard function for cause $k, (k = 1, 2)$, defined by: $\lambda_{ik}(t) = \lim_{dt \to 0} P(t \leq T_i^* < t + dt, \epsilon_i = k | T_i^* \geq t)/dt$. With $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iJ})'$ being the J-dimensional vector of covariates at the baseline, we consider a proportional hazards model for the cause-specific hazard function, defined by $\lambda_{ik}(t; \boldsymbol{\beta}) = \lambda_{0k}(t) \exp(\boldsymbol{\beta_k}' \boldsymbol{X}_i)$, where $\lambda_{0k}(t)$ is the baseline hazard function for cause k, $\boldsymbol{\beta_k} = (\beta_{k1}, \ldots, \beta_{kJ})'$ is the vector of regression coefficients, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ (Kalbfleisch and Prentice, 2002, section 8.2).

When the causes of failure are fully observed and time to failure is right-censored, one observes $T_i = \min(T_i^*, C_i)$, $\delta_i = I(T_i \leq C_i)$, and failure type ϵ_i when $\delta_i = 1$, where $I(\cdot)$ is the indicator function. Assume $\{T_i, \delta_i, \delta_i \epsilon_i, \boldsymbol{X}_i\}$ are i.i.d. for $i = 1, \ldots, n$. Under the fully observed data, we estimate $\boldsymbol{\beta}$ using the partial likelihood function

$$\prod_{i=1}^{n} \prod_{k=1}^{2} \left\{ \frac{\exp(\boldsymbol{\beta_k}' \boldsymbol{X}_i)}{\sum_{l \in R_i} \exp(\boldsymbol{\beta_k}' \boldsymbol{X}_l)} \right\}^{\delta_{ik}}, \tag{3.1}$$

where $\delta_{ik} = \delta_i I(\epsilon_i = k)$ indicates whether the failure of cause k occurs, and $R_i \equiv \{l : T_l \geq T_i\}$ is a set of subjects who are at risk at T_i. However, in our case, *neither* cause was observed. Thus, the partial likelihood function above is not feasible since δ_{ik} is not observable. When neither cause is observed, the available data is $\{T_i, \delta_i, \boldsymbol{X}_i\}$ for $i = 1, \ldots, n$, which is identical to the conventional right-censoring time to event data.

Partial likelihood function for $\boldsymbol{\beta}$ is

$$\prod_{i=1}^{n} \left\{ \frac{\lambda_i(T_i)}{\sum_{\ell \in R_i} \lambda_\ell(T_i)} \right\}^{\delta_i}, \tag{3.2}$$

where $\lambda_i(t)$ is the overall hazard function. Assuming only one event can occur at time $t + dt$, one write the overall hazard function as $\lambda_i(t) = \sum_{k=1}^{2} \lambda_{ik}(t)$ since $P(t \leq T_i^* < t + dt | T_i^* \geq t) = \sum_{k=1}^{2} P(t \leq T_i^* < t + dt, \epsilon_i = k | T_i^* \geq t)$. Hence, (3.2) becomes

$$\prod_{i=1}^{n} \left\{ \frac{\sum_{k=1}^{2} \lambda_{0k}(T_i) \exp(\boldsymbol{\beta}_k{}' \boldsymbol{X}_i)}{\sum_{\ell \in R_i} \sum_{k=1}^{2} \lambda_{0k}(T_i) \exp(\boldsymbol{\beta}_k{}' \boldsymbol{X}_\ell)} \right\}^{\delta_i},$$

where the baseline hazard function $\lambda_{0k}(t)$ cannot be completely unspecified for $k = 1, 2$, unlike the partial likelihood function in (3.1).

The primary interest of the competing risks model in our application is written as

$$\lambda_{i1}(t) = \lambda_0(t) \exp(\alpha), \tag{3.3}$$

$$\lambda_{i2}(t) = \lambda_0(t) \exp(\boldsymbol{\beta}' \boldsymbol{X}_i). \tag{3.4}$$

This model fits naturally with the *P. vivax* malaria data we intend to analyze. Reinfection is considered as the first cause of failure ($\epsilon_i = 1$) that randomly occurs from the environment following a time-to-event distribution with no association with the baseline covariates $\boldsymbol{X}_i$. We assume its hazard $\lambda_{i1}(t)$ can be written as the baseline hazard $\lambda_0(t)$ attenuated by a constant factor $\exp(\alpha)$ as shown in model (3.3). The hazard function $\lambda_{i1}(t)$ is considered as the background hazard. For the *P. vivax* malaria study, $\lambda_{i1}(t)$ represents a random mosquito bite from the living or working environment. Relapse is considered the second cause of failure ($\epsilon_i = 2$) that is associated with the baseline covariates $\boldsymbol{X}_i$ in model (3.4), which follows a proportional hazards model. These two causes of failure compete to occur, and only one of the causes, either relapse or reinfection, would occur if the event time is not censored. Under models (3.3) and (3.4), both hazard functions share the same baseline hazard $\lambda_0(t)$. The ratio of $\lambda_{i1}(t)$ and $\lambda_{i2}(t)$ only depends on baseline covariates $\boldsymbol{X}_i$, and can be considered as a semiparametric two-sample density ratio model promoted by Qin (1998). The baseline hazard $\lambda_0(t)$ here needs no specification, and can be any function of time. It can also be a function of covariates, under the condition that covariates included in $\lambda_0(t)$ are independent of those in $\boldsymbol{X}_i$.

Without any specification of $\lambda_0(t)$, one can use the partial likelihood function

$$\boldsymbol{P}\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\frac{\exp(\alpha) + \exp(\boldsymbol{\beta}'\boldsymbol{X}_i)}{\sum_{\ell \in R_i} \{ \exp(\alpha) + \exp(\boldsymbol{\beta}'\boldsymbol{X}_\ell) \}} \right]^{\delta_i}, \tag{3.5}$$

to estimate $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}')'$, where α and $\boldsymbol{\beta}$ are unknown parameters of interest. However, the dimensionality of $\boldsymbol{\theta}$ is a concern in our case since genetic sequencing produces a large number of haplotypes that are considered as covariates in our model. In Section 3.4, we introduce a penalized maximum partial likelihood method to estimate the high-dimensional $\boldsymbol{\theta}$.

In addition, we discuss an approach to verify the specification of models (3.3) and (3.4) for the *P. vivax* malaria data. The model diagnosis can be explored by martingale residuals defined by $\widehat{M}_i = \delta_i - \widehat{\Lambda}_i(T_i)$ for subjects $i = 1, \ldots, n$, where $\widehat{\Lambda}_i(t)$ is the estimated cumulative hazard function for $\Lambda_i(t) = \Lambda_0(t)\{\exp(\alpha) + \exp(\boldsymbol{\beta}'\boldsymbol{X}_i)\}$. The estimation involves not only parameter estimates for $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}')'$, but also baseline hazard estimate for $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$. One can use a Breslow-type estimator $\widehat{\Lambda}_0(t) = \sum_{i=1}^n I(T_i \leq t)\delta_i / \sum_{j \in R_i}\{\exp(\widehat{\alpha}) + \exp(\widehat{\boldsymbol{\beta}}'\boldsymbol{X}_j)\}$ for $\Lambda_0(t)$ and calculate a test statistic $T(x) = \sum_{i=1}^n I(\widehat{\boldsymbol{\beta}}'\boldsymbol{X}_i \leq x)\widehat{M}_i$ for a lack of fit test over the follow-up time. One can construct a confidence band for $T(x)$ via Monte-Carlo simulation, as proposed in Lin, Wei and Ying (1993). Model diagnosis results for the *P. vivax* malaria data are given in Section 3.6.

3.3 Classification

We propose two classifiers to classify the event to one of the two causes. The first classifier uses the baseline information and partial likelihood function (3.5) to obtain the initial estimate of the probability that the event is of cause k. The second classifier updates the first classifier using transition likelihoods under different causes. We expect that the second classifier will perform better when the transition of covariates is informative since more information is involved. If the transition of covariates is not informative of the cause of failure, the second classifier improves little from the first classifier.

3.3.1 Based on baseline information

Let $N_i^*(t)$ be the number of events up to time t, and $dN_i^*(t) = N_i^*(t+dt) - N_i^*(t)$ be the event indicator in the next instantaneous time dt after t. The observed counting process is $N_i(t) = Y_i(t)N_i^*(t)$, where

$Y_i(t) = I(T_i \geq t)$ indicates whether subject i is at risk at time t. Let $\xi_{ik}^{(0)}(t) = P(\epsilon_i = k | dN_i(t) = 1, \boldsymbol{X}_i = \boldsymbol{x}_i)$ be the probability of cause k, given that an event occurs in $[t, t + dt)$ and the realization of baseline covariate is $\boldsymbol{X}_i = \boldsymbol{x}_i$. We have: $\xi_{ik}^{(0)}(t) = P(\epsilon_i = k | dN_i(t) = 1, \boldsymbol{X}_i = \boldsymbol{x}_i) = \lambda_{ik}(t; \boldsymbol{\theta}) / \lambda_i(t; \boldsymbol{\theta})$. If an event occurs at $T_i = t_i$ for subject i, $\xi_{ik}^{(0)}(t_i)$ can be estimated by

$$\widehat{\xi}_{i1}^{(0)}(t_i) = \frac{\lambda_{i1}(t_i; \widehat{\boldsymbol{\theta}})}{\lambda_i(t_i; \widehat{\boldsymbol{\theta}})} = \frac{\lambda_0(t_i) \exp(\widehat{\alpha})}{\lambda_0(t_i)\{\exp(\widehat{\alpha}) + \exp(\widehat{\boldsymbol{\beta}}' \boldsymbol{x}_i)\}} = \frac{\exp(\widehat{\alpha})}{\exp(\widehat{\alpha}) + \exp(\widehat{\boldsymbol{\beta}}' \boldsymbol{x}_i)}, \tag{3.6}$$

$$\widehat{\xi}_{i2}^{(0)}(t_i) = \frac{\lambda_{i2}(T_i; \widehat{\boldsymbol{\theta}})}{\lambda_i(T_i; \widehat{\boldsymbol{\theta}})} = \frac{\lambda_0(T_i) \exp(\widehat{\boldsymbol{\beta}}' \boldsymbol{x}_i)}{\lambda_0(T_i)\{\exp(\widehat{\alpha}) + \exp(\widehat{\boldsymbol{\beta}}' \boldsymbol{x}_i)\}} = \frac{\exp(\widehat{\boldsymbol{\beta}}' \boldsymbol{x}_i)}{\exp(\widehat{\alpha}) + \exp(\widehat{\boldsymbol{\beta}}' \boldsymbol{x}_i)}, \tag{3.7}$$

where $\widehat{\boldsymbol{\theta}}$ is the maximum partial likelihood estimator of $\boldsymbol{\theta}$ in (3.5). Since formulae (3.6) and (3.7) are independent of t_i, we write $\widehat{\xi}_{i1}^{(0)}$ and $\widehat{\xi}_{i2}^{(0)}$ in short for $\widehat{\xi}_{i1}^{(0)}(t_i)$ and $\widehat{\xi}_{i2}^{(0)}(t_i)$, respectively.

We classify an event to be of cause 2 if $\widehat{\xi}_{i2}^{(0)} > \widehat{\xi}_{i1}^{(0)}$ and to be of cause 1 otherwise.

3.3.2 Based on both baseline and event information

When an event occurs for subject i, we assume that $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{iJ})'$ is collected at the event time, which is the same set of covariates as baseline covariates $\boldsymbol{X}_i$. We propose to utilize the transitions from $\boldsymbol{X}_i$ to $\boldsymbol{Z}_i$ to aid the cause classification. Let $\xi_{ik}^{(1)}(t) = P(\epsilon_i = k | dN_i(t) = 1, \boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{Z}_i = \boldsymbol{z}_i)$ be the probability of cause k given realizations of both $\boldsymbol{X}_i = \boldsymbol{x}_i$ and $\boldsymbol{Z}_i = \boldsymbol{z}_i$. One can show that

$$\begin{aligned}
\xi_{ik}^{(1)}(t) &= \frac{f(\boldsymbol{z}_i | \epsilon_i = k, dN_i(t) = 1, \boldsymbol{X}_i = \boldsymbol{x}_i) P(\epsilon_i = k | dN_i(t) = 1, \boldsymbol{X}_i = \boldsymbol{x}_i)}{\sum_{k=1}^{2} f(\boldsymbol{z}_i | \epsilon_i = k, dN_i(t) = 1, \boldsymbol{X}_i = \boldsymbol{x}_i) P(\epsilon_i = k | dN_i(t) = 1, \boldsymbol{X}_i = \boldsymbol{x}_i)} \\
&= \frac{\phi_i(k) \xi_{ik}^{(0)}(t)}{\sum_{k=1}^{2} \phi_i(k) \xi_{ik}^{(0)}(t)},
\end{aligned}$$

where $\phi_i(k) = f(\boldsymbol{z}_i | \epsilon_i = k, dN_i(t) = 1, \boldsymbol{X}_i = \boldsymbol{x}_i)$ is the conditional density function of $\boldsymbol{Z}_i$ given $\boldsymbol{X}_i$ under cause k. We call $\phi_i(k)$ the conditional *transition likelihood* of cause k. One can treat the classification probability $\xi_{ik}^{(1)}(t)$ as an updated version of $\xi_{ik}^{(0)}(t)$ by the ratio of transition likelihoods between possible causes since $\frac{\xi_{ik}^{(1)}(t)}{\xi_{i\ell}^{(1)}(t)} = \frac{\phi_i(k)}{\phi_i(\ell)} \frac{\xi_{ik}^{(0)}(t)}{\xi_{i\ell}^{(0)}(t)}$ for $\ell = 1, 2$ and $\ell \neq k$. Note that if the transition likelihoods are informative, $\phi_i(1)$ and $\phi_i(2)$ will be very different from each other and thus lead to more accurate classification of $\xi_{ik}^{(1)}(t)$.

We assume that the transition likelihood $\phi_i(k)$ follows a parametric model $\phi_i(k, \boldsymbol{\gamma}_k)$, where $\boldsymbol{\gamma}_k$ is the vector of parameters to be estimated. More details of this parametric model $\phi_i(k)$ follow in Section 3.3.3.

The distribution of $\boldsymbol{Z}_i$ is a mixture of transition likelihoods from two latent causes:

$$
\begin{aligned}
f(\boldsymbol{z}_i|dN_i(t)=1,\boldsymbol{X}_i=\boldsymbol{x}_i) &= \sum_{k=1}^{2} f(\boldsymbol{z}_i,\epsilon_i=k|dN_i(t)=1,\boldsymbol{X}_i=\boldsymbol{x}_i) \\
&= \sum_{k=1}^{2} f(\boldsymbol{z}_i|\epsilon_i=k,dN_i(t)=1,\boldsymbol{X}_i=\boldsymbol{x}_i)P(\epsilon_i=k|dN_i(t)=1,\boldsymbol{X}_i=\boldsymbol{x}_i) \\
&= \sum_{k=1}^{2} \phi_i(k,\boldsymbol{\gamma}_k)\xi_{ik}^{(0)}(t).
\end{aligned}
$$

With $\xi_{ik}^{(0)}(t)$ being estimated by $\widehat{\xi}_{ik}^{(0)}$, and let $m=\sum_{i=1}^{n}\delta_i$ be the number of subjects having recurrent infections. We estimate $\boldsymbol{\gamma}_k$ by maximizing a pseudo log-likelihood function:

$$
\ell(\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2)=\sum_{i=1}^{m}\log\Big\{\sum_{k=1}^{2}\phi_i(k,\boldsymbol{\gamma}_k)\widehat{\xi}_{ik}^{(0)}\Big\}. \tag{3.8}
$$

Let $(\widehat{\boldsymbol{\gamma}}_1',\widehat{\boldsymbol{\gamma}}_2')'=\mathrm{argmax}_{\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2}\,\ell(\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2)$ and write $\widehat{\xi}_{ik}^{(1)}$ in short for $\widehat{\xi}_{ik}^{(1)}(t_i)$. We estimate $\xi_{ik}^{(1)}$ by

$$
\widehat{\xi}_{ik}^{(1)}=\frac{\phi_i(k,\widehat{\boldsymbol{\gamma}}_k)\widehat{\xi}_{ik}^{(0)}}{\sum_{k=1}^{2}\phi_i(k,\widehat{\boldsymbol{\gamma}}_k)\widehat{\xi}_{ik}^{(0)}}. \tag{3.9}
$$

We classify the event to be of cause 2 if and only if $\widehat{\xi}_{i2}^{(1)}>\widehat{\xi}_{i1}^{(1)}$.

3.3.3 Transition likelihood

The transition likelihood plays a critical role in classification. In this section, we discuss a generalized linear model to model the transition likelihood function $\phi_i(k,\boldsymbol{\gamma}_k)$. Suppose the density of Z_{ij} conditioning on X_{ij} and $\epsilon_i=k$ has the form of

$$
f(z;\vartheta_{ijk},\psi_{jk})=\exp\big\{(z\vartheta_{ijk}-b(\vartheta_{ijk}))/a(\psi_{jk})+c(z,\psi_{jk})\big\},
$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions, ϑ_{ijk} is the natural parameter, and ψ_{jk} is the dispersion parameter (McCullagh and Nelder, 1989). Let $g(\mu_{ijk})=\vartheta_{ijk}$ be the cause-specific canonical link function,

where $\mu_{ijk} = \mathrm{E}(Z_{ij}|\epsilon_i = k, dN_i(t) = 1, X_{ij} = x_{ij})$. We define $\phi_i(k, \boldsymbol{\gamma}_k)$ as:

$$\begin{aligned}\phi_i(k, \boldsymbol{\gamma}_k) &= f(\boldsymbol{z}_i|\epsilon_i = k, dN_i(t) = 1, \boldsymbol{X}_i = \boldsymbol{x}_i) \\ &= \prod_{j=1}^{J} \exp\left[\{z_{ij} g(\mu_{ijk}) - b(g(\mu_{ijk}))\}/a(\psi_{jk}) + c(z_{ij}, \psi_{jk})\right],\end{aligned}$$

where $g(\mu_{ijk}) = q_{jk0} + x_{ij} q_{jk1}$, q_{jk0} is the intercept term and q_{jk1} is the coefficient of x_{ij}.

To improve the classification performance, we want the transition likelihoods to be as informative as possible. When some external variables contain information about the transition, we also would like to incorporate them into the transition likelihoods. Let $\boldsymbol{W}_{ij} = (W_{ij1}, W_{ij2}, \ldots, W_{ijL})'$ be the L-dimensional vector of these external variables and $\boldsymbol{W}_i = (\boldsymbol{W}'_{i1}, \boldsymbol{W}'_{i2}, \ldots, \boldsymbol{W}'_{iJ})'$. Then, we have

$$\begin{aligned}\phi_i(k, \boldsymbol{\gamma}_k) &= f(\boldsymbol{z}_i|\epsilon_i = k, dN_i(t) = 1, \boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{W}_i = \mathbf{w}_i) \\ &= \prod_{j=1}^{J} \exp\left[\{z_{ij} g(\mu_{ijk}) - b(g(\mu_{ijk}))\}/a(\psi_{jk}) + c(z_{ij}, \psi_{jk})\right],\end{aligned}$$

where $g(\mu_{ijk}) = q_{jk0} + x_{ij} q_{jk1} + \mathbf{w}'_{ij} \boldsymbol{q}^*_{jk}$, $\mathbf{w}_{ij}$ is a realization of $\boldsymbol{W}_{ij}$ with the corresponding coefficients $\boldsymbol{q}^*_{jk} = (q^*_{jk1}, q^*_{jk2}, \ldots, q^*_{jkL})'$. Let $\boldsymbol{q}_{k0} = (q_{1k0}, \ldots, q_{Jk0})'$, $\boldsymbol{q}_{k1} = (q_{1k1}, \ldots, q_{Jk1})'$, $\boldsymbol{q}^*_k = (\boldsymbol{q}^{*\prime}_{1k}, \ldots, \boldsymbol{q}^{*\prime}_{Jk})'$ and $\boldsymbol{\psi}_k = (\psi_{1k}, \ldots, \psi_{Jk})'$. Then, we let $\boldsymbol{\gamma}_k = (\boldsymbol{q}'_{k0}, \boldsymbol{q}'_{k1}, \boldsymbol{q}^{*\prime}_k, \boldsymbol{\psi}'_k)'$ represent all the parameters in $\phi_i(k, \boldsymbol{\gamma}_k)$.

Our proposed transition likelihood model manifests differently according to the type of covariates. We give three examples showing how to construct $\phi_i(k, \boldsymbol{\gamma}_k)$ when the covariates are binary, normal, or Poisson.

Example 1 (Binary Covariates) *When X_{ij} and Z_{ij} are binary covariates, we have*

$$g(\mu_{ijk}) = \log\left(\frac{\mu_{ijk}}{1 - \mu_{ijk}}\right) = q_{jk0} + x_{ij} q_{jk1} + \mathbf{w}'_{ij} \boldsymbol{q}^*_{jk},$$

where the link function g is a logit function. The transitional likelihood for cause k becomes

$$\phi_i(k, \boldsymbol{\gamma}_k) = \prod_{j=1}^{J} \mu_{ijk}^{z_{ij}} \big(1 - \mu_{ijk}\big)^{1 - z_{ij}}, \tag{3.10}$$

where $\mu_{ijk} = \exp(\vartheta_{ijk})/\{1 + \exp(\vartheta_{ijk})\}$, $\vartheta_{ijk} = q_{jk0} + x_{ij} q_{jk1} + \mathbf{w}'_{ij} \boldsymbol{q}^*_{jk}$, $\boldsymbol{\gamma}_k = (\boldsymbol{q}'_{k0}, \boldsymbol{q}'_{k1}, \boldsymbol{q}^{*\prime}_k)'$ and $\boldsymbol{\psi}_k = (1, \ldots, 1)'$.

Example 2 (Normal Covariates) *When X_{ij} and Z_{ij} are normally distributed covariates, we have*

$$g(\mu_{ijk}) = \mu_{ijk} = q_{jk0} + x_{ij}q_{jk1} + \mathbf{w}'_{ij}\boldsymbol{q}^*_{jk},$$

$$\psi_{jk} = Var(Z_{ij}|\epsilon_i = k, dN_i(t) = 1, X_{ij} = x_{ij}, \boldsymbol{W}_{ij} = \mathbf{w}_{ij}),$$

where the link function g is an identity function. The transitional likelihood for cause k becomes

$$\phi_i(k, \boldsymbol{\gamma}_k) = \prod_{j=1}^{J} \frac{1}{\sqrt{2\pi\psi_{jk}}} \exp\Big\{ - \frac{(z_{ij} - \mu_{ijk})^2}{2\psi_{jk}} \Big\}, \tag{3.11}$$

where $\boldsymbol{\gamma}_k = (\boldsymbol{q}'_{k0}, \boldsymbol{q}'_{k1}, \boldsymbol{q}^{*'}_k, \boldsymbol{\psi}'_k)'$ and $\boldsymbol{\psi}_k = (\psi_{1k}, \ldots, \psi_{Jk})'$.

Example 3 (Poisson Covariates) *When X_{ij} and Z_{ij} are Poisson covariates, we have*

$$g(\mu_{ijk}) = \log(\mu_{ijk}) = q_{jk0} + x_{ij}q_{jk1} + \mathbf{w}'_{ij}\boldsymbol{q}^*_{jk},$$

where the link function g is a log function. The transitional likelihood for cause k becomes

$$\phi_i(k, \boldsymbol{\gamma}_k) = \prod_{j=1}^{J} \frac{\mu_{ijk}^{z_{ij}} \exp(-\mu_{ijk})}{z_{ij}!}, \tag{3.12}$$

where $\mu_{ijk} = \exp(\vartheta_{ijk})$, $\vartheta_{ijk} = q_{jk0} + x_{ij}q_{jk1} + \mathbf{w}'_{ij}\boldsymbol{q}^*_{jk}$, $\boldsymbol{\gamma}_k = (\boldsymbol{q}'_{k0}, \boldsymbol{q}'_{k1}, \boldsymbol{q}^{*'}_k)'$ and $\boldsymbol{\psi}_k = (1, \ldots, 1)'$.

3.4 Computation

3.4.1 Estimation of parameters

Define the negative partial log-likelihood function as

$$\ell(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \delta_i \left[\log\Big\{ \exp(\alpha) + \exp(\boldsymbol{\beta}'\boldsymbol{X}_i) \Big\} - \log\Big\{ \sum_{l \in R_i} \Big\{ \exp(\alpha) + \exp(\boldsymbol{\beta}'\boldsymbol{X}_l) \Big\} \Big\} \right]. \tag{3.13}$$

To estimate $\boldsymbol{\theta}$ in (3.5), we propose to solve a penalized likelihood problem

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \big\{ \ell(\boldsymbol{\theta}) + \nu p(\boldsymbol{\beta}) \big\}, \tag{3.14}$$

where ν is a positive tuning parameter and $p(\boldsymbol{\beta})$ is a penalty function. When sample size n is larger than the number of covariates J, (4.1) is a low-dimensional problem, in which case we set $\nu = 0$. When n is smaller than J, (4.1) is a high-dimensional problem, in which case we choose the optimal ν by minimizing the Bayesian Information Criterion (BIC, Schwarz et al., 1978), which is given by $\text{BIC} = 2\ell(\widehat{\boldsymbol{\theta}}) + c \cdot \log(n)$, where c is the number of covariates selected in the model. Popular choices of $p(\boldsymbol{\beta})$ include the L_1-penalty (Tibshirani, 1996), the elastic net penalty (Zou and Hastie, 2005), or some folded concave penalty (Fan and Lv, 2011). In this paper, we choose the L_1-penalty.

To solve (4.1), we use a proximal gradient algorithm (Parikh and Boyd, 2014). First, we find a quadratic approximation to $\ell(\boldsymbol{\theta})$ centered at $\boldsymbol{\theta}^{(h)}$, the estimate of $\boldsymbol{\theta}$ at the hth iteration of the algorithm, that majorizes $\ell(\boldsymbol{\theta})$. That is

$$\ell(\boldsymbol{\theta}) \leq \ell(\boldsymbol{\theta}^{(h)}) + (\boldsymbol{\theta} - \boldsymbol{\theta}^{(h)})' \nabla \ell(\boldsymbol{\theta}^{(h)}) + \frac{1}{2d} ||\boldsymbol{\theta} - \boldsymbol{\theta}^{(h)}||_2^2, \tag{3.15}$$

where d is a scalar that plays the role as a step size, $\boldsymbol{\theta}^{(h)} = (\alpha^{(h)}, \boldsymbol{\beta}^{(h)'})'$ and the gradient vector $\nabla \ell(\boldsymbol{\theta}^{(h)})$ is given by $\nabla \ell(\boldsymbol{\theta}^{(h)}) = (\nabla_\alpha \ell(\boldsymbol{\theta}^{(h)}), \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\theta}^{(h)})')'$, where

$$\nabla_\alpha \ell(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \delta_i \left[\frac{\exp(\alpha)}{\exp(\alpha) + \exp(\boldsymbol{\beta}' \boldsymbol{X}_i)} - \frac{\sum_{l \in R_i} \exp(\alpha)}{\sum_{l \in R_i} \left\{ \exp(\alpha) + \exp(\boldsymbol{\beta}' \boldsymbol{X}_l) \right\}} \right], \tag{3.16}$$

$$\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \delta_i \left[\frac{\boldsymbol{X}_i \exp(\boldsymbol{\beta}' \boldsymbol{X}_i)}{\exp(\alpha) + \exp(\boldsymbol{\beta}' \boldsymbol{X}_i)} - \frac{\sum_{l \in R_i} \left\{ \mathbf{X}_l \exp(\boldsymbol{\beta}' \boldsymbol{X}_l) \right\}}{\sum_{l \in R_i} \left\{ \exp(\alpha) + \exp(\boldsymbol{\beta}' \boldsymbol{X}_l) \right\}} \right]. \tag{3.17}$$

Denote the right-hand side of (3.15) by $Q_d(\boldsymbol{\theta}; \boldsymbol{\theta}^{(h)})$ and let $g(\boldsymbol{\beta}) = \nu p(\boldsymbol{\beta})$. Then we minimize $Q_d(\boldsymbol{\theta}, \boldsymbol{\theta}^{(h)}) + g(\boldsymbol{\beta})$, which gives the proximal problem

$$\alpha^{(h+1)} = \underset{\alpha}{\text{argmin}} \ \frac{1}{2} ||\alpha - [\alpha^{(h)} - d \nabla_\alpha \ell(\boldsymbol{\theta}^{(h)})]||_2^2, \tag{3.18}$$

$$\boldsymbol{\beta}^{(h+1)} = \underset{\boldsymbol{\beta}}{\text{argmin}} \ \frac{1}{2} ||\boldsymbol{\beta} - [\boldsymbol{\beta}^{(h)} - d \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\theta}^{(h)})]||_2^2 + d g(\boldsymbol{\beta}). \tag{3.19}$$

The solution of (3.18) is given by $\alpha^{(h+1)} = \alpha^{(h)} - d \nabla_\alpha \ell(\boldsymbol{\theta}^{(h)})$. The solution of (3.19) is given by a proximal operator $\boldsymbol{\beta}^{(h+1)} = \text{prox}_{dg}(\boldsymbol{\beta}^{(h)} - d \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\theta}^{(h)}))$. Depending on the choice of penalty function, such an operator has a closed-form expression. For example, if we use an L_1-penalty: $p(\boldsymbol{\beta}) = ||\boldsymbol{\beta}||_1$, then $\text{prox}_{dg}(\boldsymbol{\beta}^{(h)} - d \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\theta}^{(h)})) = s(\boldsymbol{\beta}^{(h)} - d \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\theta}^{(h)}), \nu d)$, where $s(\boldsymbol{x}, \pi)$ is the elementwise soft-thresholding operator, whose jth element is defined as $s(\boldsymbol{x}, \pi)_j = \text{sgn}(x_j)(|x_j| - \pi)_+$. As for the step size, we follow

Parikh & Boyd (2014, section 4.2) and perform a backtracking line search; namely, we iteratively decrease step size until the majorization holds, i.e, the inequality (3.15) holds. This strategy is commonly used in the proximal gradient method.

We stop iterating the algorithm when the change in the objective function between two consecutive iterations is less than $\zeta\%$ of the objective function's value at the former iteration, where $\zeta \in (0, 100)$ is a user-defined stopping threshold, which is chosen by us to be 10. A detailed algorithm is summarized as follows:

Data: $\boldsymbol{X}_i, T_i, \delta_i; i = 1, \ldots, n$.

Result: Estimates for $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}')'$.

Initialize d at $d^{(0)} \in \mathbb{R}^+$, $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}')'$ at $\boldsymbol{\theta}^{(0)} = (\alpha^{(0)}, \boldsymbol{\beta}^{(0)'})'$, where $\alpha^{(0)} \in \mathbb{R}^1$, $\boldsymbol{\beta}^{(0)} \in \mathbb{R}^p$;

At the hth iteration, let $d = d^{(h-1)}$,

repeat

Let $\alpha = \alpha^{(h-1)} - d\nabla_\alpha \ell(\boldsymbol{\theta}^{(h-1)})$ and $\boldsymbol{\beta} = \text{prox}_{dg}(\boldsymbol{\beta}^{(h-1)} - d\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\theta}^{(h-1)}))$,

if $\ell(\boldsymbol{\theta}) \leq Q_d\{\boldsymbol{\theta}; \boldsymbol{\theta}^{(h-1)}\}$ **then**

let $d^{(h)} = d$, $\alpha^{(h)} = \alpha$, $\boldsymbol{\beta}^{(h)} = \boldsymbol{\beta}$; break;

else

let $d = 0.8d$.

end if

until $\left| \frac{\{\ell(\boldsymbol{\theta}^{(h)}) + g(\boldsymbol{\beta}^{(h)})\} - \{\ell(\boldsymbol{\theta}^{(h-1)}) + g(\boldsymbol{\beta}^{(h-1)})\}}{\ell(\boldsymbol{\theta}^{(h-1)}) + g(\boldsymbol{\beta}^{(h-1)})} \right| \leq \zeta\%$.

Algorithm 1: The Proximal Gradient Algorithm

3.4.2 Classification Algorithm

We give a complete algorithm for classifying the causes of an event by using time to event information T_i and δ_i, baseline covariates $\boldsymbol{X}_i$, covariates collected when the event occurs $\boldsymbol{Z}_i$, and external informative covariates $\boldsymbol{W}_i$ in this section.

Firstly, Given $\boldsymbol{X}_i, T_i$, and δ_i, estimate $\boldsymbol{\theta}$ using partial likelihood (3.5) and Algorithm 1. Secondly, estimate $\xi_{ik}^{(0)}$ by (3.6) and (3.7). Thirdly, based on the type of covariates $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$, estimate the *transition likelihood* $\phi_i(k, \boldsymbol{\gamma}_k)$ of cause k by maximizing the pseudo likelihood function in (3.8). Next, estimate $\xi_{ik}^{(1)}$ as in (3.9). Finally, if $\widehat{\xi}_{i2}^{(1)} > \widehat{\xi}_{i1}^{(1)}$ then classify the event to be of cause 2, otherwise classify the event to be of cause 1.

3.5 Simulation Experiments

To study the improvement in classification using transition likelihoods compared with using baseline information alone, we carry out comprehensive simulation experiments to evaluate the performance of two classifiers based on $\widehat{\xi}_i^{(0)}$ and $\widehat{\xi}_i^{(1)}$ respectively. We evaluate the performance of the proposed classifiers by comparing their sensitivity, specificity, and overall accuracy in classifying the causes of events. We mimic the data observed in the *P. vivax* malaria infection study (Lin et al., 2015) and assume that the cause could be either reinfection ($\epsilon_i = 1$) or relapse ($\epsilon_i = 2$). Sensitivity is defined as the number of subjects correctly classified as relapse divided by the number of relapse subjects; specificity is defined as the number of subjects correctly classified as reinfection divided by the number of reinfection subjects and overall accuracy is defined as the number of correctly classified subjects divided by the number of subjects.

Following the proposed model in Section 3.2, we assume that the baseline hazard is a homogeneous Poisson process with hazard function $\lambda_0(t)$, which is a constant for $t > 0$ and the same for all subjects. By using the partial likelihood function (3.5), we do not need to specify $\lambda_0(t)$ and expect the classification performance to be similar under different baseline hazard functions. We carry out simulations with three different baseline hazard functions $\lambda_0(t) = \exp(\tau)$, where $\tau = -0.5, 0, 0.5$.

The reinfection process was assumed to be the same for all subjects with hazard function $\lambda_{i1}(t) = \lambda_0(t)\exp(\alpha)$. The relapse process was assumed to have a proportional hazard function $\lambda_{i2}(t) = \lambda_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{X}_i)$ for subject i. The first classifier classifies a recurrent infection as relapse if $\widehat{\xi}_{i2}^{(0)} > 0.5$, and the second classifier classifies a recurrent infection as relapse if $\widehat{\xi}_{i2}^{(1)} > 0.5$.

We consider two situations where $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are binary and normally distributed variables. We allow dimensions of $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ to be either low or high. Under the low-dimensional settings, we set two combinations for n and J, with $(n, J) = (400, 10)$ and $(n, J) = (800, 20)$. For the high-dimensional settings, we focus on the classification performance of the classifiers, as well as the variable selection performance. We consider $(n, J) = (100, 200)$ and $(n, J) = (200, 400)$, where the former is closer to the real *P. vivax* malaria infection study. When evaluating the variable selection performance, we focus on the sensitivity, specificity, and overall accuracy of selecting covariates with non-zero regression coefficients.

Remark that the improvement of the second classifier is mainly attributed to including the transition likelihoods from the baseline covariates $\boldsymbol{X}_i$ to the covariates at recurrence infection $\boldsymbol{Z}_i$. If $\boldsymbol{Z}_i$ associates with $\boldsymbol{X}_i$, the transition likelihood is informative, and the second classifier would have a better classification

performance. However, when $\boldsymbol{Z}_i$ is not associated with $\boldsymbol{X}_i$, then little information would be contained in the transition likelihood. Thus, the second classifier would have a similar performance to the first classifier. We consider two scenarios where the association between $\boldsymbol{Z}_i$ and $\boldsymbol{X}_i$ is either strong or weak. For simplicity, we assume that for each pair of X_{ij} and Z_{ij}, there exists only one external covariate W_{ij} that is associated with the transition.

3.5.1 Binary Covariates

For the low-dimensional setting, we set α to be 0, the first 3 components of $\boldsymbol{\beta}$ to be $\log(1.5)$, and the rest of the components to be 0. We generated $\boldsymbol{X}_i$ from the Bernoulli distribution with probability $P(X_{ij} = 1) = 0.5 \exp\{-0.1(j-1)\}$ for $j = 1, \ldots, 10$. Such a choice of $\boldsymbol{X}_i$ and $\boldsymbol{\beta}$ indicates that the three most prevalent variants are associated with the relapse. We generated failure time T_i^* based on the all-cause hazard function $\lambda_i(t) = \lambda_{i1}(t) + \lambda_{i2}(t)$ and then determined whether the infection is a relapse or reinfection by a Bernoulli random variable with success probability equals to $\exp(\boldsymbol{\beta}'\boldsymbol{X}_i)/\{\exp(\alpha) + \exp(\boldsymbol{\beta}'\boldsymbol{X}_i)\}$. The right censoring time C_i was generated following a uniform distribution between 0 and c, where c is a constant controlling for 20% censoring. The observed time T_i is the minimum between T_i^* and C_i. We assume that for any $j \leq J$, there is one external covariate W_{ij} affecting the transition from X_{ij} to Z_{ij}. For each i and j, we independently generate W_{ij} from a uniform distribution between 0 and 1, which is also independent of X_{ij}.

If the event is reinfection, $\boldsymbol{Z}_i$ was generated independently from the same distribution as $\boldsymbol{X}_i$. If the event is a relapse, we generated $\boldsymbol{Z}_i$ following the transition model (3.10). We let $q_{j21} = q_{j21}^* = 0.9$ in the first scenario when $\boldsymbol{Z}_i$ strongly associates with $\boldsymbol{X}_i$, and $q_{j21} = q_{j21}^* = 0.001$ in the second scenario when $\boldsymbol{Z}_i$ weakly associates with $\boldsymbol{X}_i$. The intercept q_{j20} was set to be 0.3 for both scenarios. We repeat the simulation 500 times for each combination of n and J under both scenarios. The operating characteristics of the two classifiers are reported in Table 3.1. Reported values are means and standard deviations over 500 simulations.

Table 3.1 shows that performance of the first classifier $I(\widehat{\xi}_i^{(0)} > 0.5)$ is similar under both scenarios in terms of sensitivity, specificity, and overall accuracy. This result is reasonable since we only included baseline covariates and time to event information when constructing the first classifier, and these information were generated using the same mechanisms under both scenarios. The second classifier $I(\widehat{\xi}_{i2}^{(1)} > 0.5)$ has a better performance than the first classifier $I(\widehat{\xi}_{i2}^{(0)} > 0.5)$ in scenario 1, where sensitivity, specificity, and overall accuracy are all in favor of the second classifier. The classification accuracy gets better when sample

Table 3.1: Classification of proposed classifiers with low-dimensional binary covariates.

			$I(\widehat{\xi}_i^{(0)} > 0.5)$			$I(\widehat{\xi}_i^{(1)} > 0.5)$		
Scenario	τ	$(n,\ J)$	Sensitivity	Specificity	Overall	Sensitivity	Specificity	Overall
1	-0.5	(400, 10)	50.3 (20.2)	59.0 (19.2)	53.6 (5.1)	90.7 (4.2)	94.3 (3.2)	92.1 (2.2)
		(800, 20)	50.1 (19.9)	59.6 (19.2)	54.0 (4.5)	97.8 (0.9)	98.7 (0.8)	98.0 (0.5)
	0	(400, 10)	49.1 (18.4)	60.1 (17.6)	53.6 (4.6)	89.3 (10.8)	93.2 (10.8)	90.9 (10.4)
		(800, 20)	49.6 (18.2)	59.7 (17.8)	53.8 (3.9)	97.9 (0.9)	98.2 (0.8)	98.0 (0.6)
	0.5	(400, 10)	48.3 (18.7)	61.9 (17.3)	53.9 (4.8)	88.9 (12.2)	92.4 (12.0)	90.3 (11.8)
		(800, 20)	50.2 (17.8)	59.5 (17.2)	54.1 (3.9)	97.9 (0.8)	98.1 (0.8)	98.0 (0.5)
2	-0.5	(400, 10)	48.7 (19.7)	60.6 (18.8)	53.6 (5.0)	66.3 (16.9)	72.5 (30.2)	68.8 (21.2)
		(800, 20)	50.7 (18.7)	58.8 (17.9)	54.0 (4.1)	66.2 (14.6)	71.9 (13.4)	68.6 (11.9)
	0	(400, 10)	49.3 (19.7)	59.6 (18.5)	53.6 (5.1)	64.4 (18.2)	69.2 (32.3)	66.3 (23.1)
		(800, 20)	51.6 (17.9)	58.5 (17.4)	54.5 (3.9)	66.2 (14.7)	72.1 (13.3)	68.6 (11.9)
	0.5	(400, 10)	49.2 (18.6)	60.6 (17.6)	53.7 (4.5)	68.7 (16.6)	74.9 (27.5)	71.1 (20.4)
		(800, 20)	50.8 (18.1)	58.8 (17.5)	54.0 (4.1)	66.3 (14.5)	72.3 (12.9)	68.7 (11.8)

Sensitivity, specificity and overall accuracy are given as percentages.
Reported values are means and standard deviations over 500 simulations.

size is larger. In scenario 1, the strong association between $\boldsymbol{Z}_i$ and $\boldsymbol{X}_i$ makes the transition likelihood much more informative. Therefore, the improvement in the classification performance is obvious in this scenario. However, in scenario 2, the association between $\boldsymbol{Z}_i$ and $\boldsymbol{X}_i$ is relatively weak. The transition likelihood contains less information. Hence, the second classifier improves little upon the first classifier, averaging merely 12% – 18% improvement in the overall accuracy, even when the sample size is larger.

When n and J are fixed, we can see that differences in the baseline hazard function $\lambda_0(t)$ barely affect the performance of both classifiers. This result is reasonable since the baseline hazard $\lambda_0(t)$ is canceled in (3.13). As long as the proportional hazards assumption stands, the classification accuracy is similar regardless of the true form of the baseline hazard $\lambda_0(t)$.

For high-dimensional settings, we set α to be 0, the first 10 components of regression coefficients in $\boldsymbol{\beta}$ to be $\log(1.5)$, and the rest to be 0. The remaining set-up was the same as in the low-dimensional setting. We repeat the simulation 500 times for each combination of (n, J) under two scenarios. The performance of the two classifiers is reported in Table 3.2.

In Table 3.2, we can see similar results as in Table 3.1. The first classifier behaves similarly under both scenarios. In scenario 1, the second classifier has perfect sensitivity and nearly perfect specificity. In scenario 2, the second classifier has similar overall accuracy as the first classifier, with slightly lower sensitivity and slightly higher specificity. The choice of the baseline hazard function $\lambda_0(t)$ barely affects the performance.

Table 3.2: Classification and variable selection of proposed classifiers with high-dimensional binary covariates.

			$I(\widehat{\xi}_i^{(0)} > 0.5)$			$I(\widehat{\xi}_i^{(1)} > 0.5)$			$\widehat{\alpha}$	$\widehat{\boldsymbol{\beta}}$		
Scenario	τ	$(n,\ J)$	Sensitivity	Specificity	Overall	Sensitivity	Specificity	Overall	Bias	Sensitivity	Specificity	Overall
1	-0.5	(100, 200)	98.1 (3.0)	4.4 (6.8)	76.1 (4.6)	100 (0)	96.3 (6.0)	99.5 (0.7)	0.48 (0.05)	75.2 (12.8)	57.1 (2.0)	58.0 (2.1)
		(200, 400)	96.7 (3.5)	8.4 (7.1)	75.3 (3.1)	100 (0)	100 (0)	100 (0)	0.51 (0.01)	87.2 (10.6)	65.7 (1.3)	66.7 (1.4)
	0	(100, 200)	97.8 (3.4)	5.3 (7.7)	74.8 (4.5)	100 (0)	97.4 (4.5)	99.6 (0.6)	0.49 (0.04)	72.9 (13.0)	58.2 (2.1)	59.0 (2.2)
		(200, 400)	95.7 (3.7)	9.2 (7.5)	75.1 (3.5)	100 (0)	100 (0)	100 (0)	0.51 (0.01)	87.0 (10.6)	65.7 (1.6)	65.9 (1.4)
	0.5	(100, 200)	97.6 (2.8)	4.6 (6.8)	75.3 (4.5)	100 (0)	96.7 (6.0)	99.5 (0.7)	0.49 (0.04)	72.9 (13.6)	58.0 (2.3)	58.7 (2.4)
		(200, 400)	95.8 (3.8)	9.8 (7.1)	75.0 (3.3)	100 (0)	99.9 (0.8)	100 (0)	0.51 (0.02)	87.1 (11.2)	65.4 (1.4)	66.0 (1.5)
2	-0.5	(100, 200)	97.9 (2.9)	4.9 (5.8)	75.8 (4.6)	91.8 (4.5)	13.0 (8.2)	73.1 (5.0)	0.49 (0.05)	78.3 (14.3)	62.2 (2.5)	61.1 (2.3)
		(200, 400)	96.2 (3.8)	8.8 (7.9)	75.3 (3.3)	90.7 (5.6)	14.9 (9.1)	72.7 (4.0)	0.50 (0.02)	73.8 (14.1)	67.2 (1.7)	66.3 (1.7)
	0	(100, 200)	97.5 (3.1)	6.4 (6.9)	74.9 (4.3)	91.9 (4.8)	14.3 (9.2)	72.7 (4.2)	0.50 (0.04)	79.2 (15.8)	62.6 (2.2)	61.4 (2.4)
		(200, 400)	95.8 (3.8)	8.8 (7.4)	74.8 (3.5)	90.6 (5.1)	15.4 (8.3)	72.5 (3.9)	0.51 (0.02)	75.3 (14.5)	67.5 (1.9)	66.5 (1.4)
	0.5	(100, 200)	97.4 (2.6)	5.7 (6.1)	75.4 (4.4)	91.5 (4.8)	13.6 (8.7)	72.8 (4.8)	0.51 (0.04)	79.0 (15.8)	61.6 (2.3)	60.5 (2.3)
		(200, 400)	95.6 (3.7)	9.5 (7.7)	74.7 (3.1)	90.3 (5.5)	16.3 (8.3)	72.2 (3.8)	0.51 (0.02)	73.1 (14.9)	66.2 (1.5)	65.4 (1.5)

Sensitivity, specificity and overall accuracy are given as percentages.
Reported values are means and standard deviations over 500 simulations.

We also evaluated the accuracy of coefficient estimates $\widehat{\boldsymbol{\theta}} = (\widehat{\alpha}, \widehat{\boldsymbol{\beta}}')'$ for the high-dimensional settings, where the bias of $\widehat{\alpha}$ and variable selection performance of $\widehat{\boldsymbol{\beta}}$ are reported in Table 3.2. Since we did not use transition likelihoods when estimating $\boldsymbol{\theta}$, the accuracy of $\widehat{\boldsymbol{\theta}}$ is similar under both scenarios. The baseline hazard function was canceled when calculating the partial likelihood function (3.5). Therefore, it has little influence on the performance of $\widehat{\boldsymbol{\theta}}$. One can see that as J gets larger, the bias of $\widehat{\alpha}$ increases. However, the performance of $\widehat{\boldsymbol{\beta}}$ improves since more variables are selected correctly.

3.5.2 Normally Distributed Covariates

In addition, we simulate for normally distributed $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$. For both low- and high-dimensional settings, we consider the same set-up for α and $\boldsymbol{\beta}$ as in the simulation study for binary covariates. We generated $\boldsymbol{X}_i$ and $\boldsymbol{W}_i$ independently from a standard normal distribution. The event time T_i^*, censoring time C_i, and observed time T_i were all generated with the same strategy as for the binary covariates. We generated $\boldsymbol{Z}_i$ based on the event type, following the transition model (3.11). We let $q_{j21} = q_{j21}^* = 0.9$ in scenario 1, where $\boldsymbol{Z}_i$ strongly associates with $\boldsymbol{X}_i$, and let $q_{j21} = q_{j21}^* = 0.001$ in scenario 2, where $\boldsymbol{Z}_i$ weakly associates with $\boldsymbol{X}_i$. We let $q_{j20} = 0.3$ and $\psi_{jk} = 1$ for each j under both scenarios. We repeated the simulation 500 times for each combination of n and J under both scenarios. The performance of two classifiers is reported in Tables 3.3 and 3.4 for low- and high-dimensional settings respectively. We also reported the estimation accuracy and variable selection performance of $\widehat{\boldsymbol{\theta}}$ in the high-dimensional settings in Table 3.4.

Table 3.3: Classification of proposed classifiers with low-dimensional continuous covariates.

			$I(\widehat{\xi}_i^{(0)} > 0.5)$			$I(\widehat{\xi}_i^{(1)} > 0.5)$		
Scenario	τ	$(n,\ J)$	Sensitivity	Specificity	Overall	Sensitivity	Specificity	Overall
1	-0.5	(400, 10)	67.8 (4.1)	54.8 (4.5)	61.6 (3.2)	97.5 (1.4)	97.5 (1.6)	97.6 (1.0)
		(800, 20)	64.9 (2.8)	58.6 (2.5)	61.8 (1.9)	99.8 (0.2)	99.8 (0.3)	99.7 (0.1)
	0	(400, 10)	65.9 (4.4)	57.1 (4.3)	61.6 (3.1)	97.6 (1.3)	97.5 (1.3)	97.5 (0.9)
		(800, 20)	63.6 (2.4)	59.5 (2.6)	61.6 (1.9)	99.7 (0.3)	99.7 (0.3)	99.7 (0.2)
	0.5	(400, 10)	64.7 (3.5)	60.2 (3.9)	62.4 (3.0)	97.6 (1.2)	97.4 (1.1)	97.5 (0.7)
		(800, 20)	62.5 (2.5)	60.4 (2.2)	61.5 (1.7)	99.7 (0.3)	99.7 (0.3)	99.7 (0.2)
2	-0.5	(400, 10)	67.6 (4.3)	54.9 (4.3)	61.5 (3.1)	68.5 (4.3)	56.2 (4.8)	62.5 (3.1)
		(800, 20)	64.6 (2.7)	58.4 (2.8)	61.8 (2.5)	67.9 (2.4)	62.7 (3.8)	65.4 (2.0)
	0	(400, 10)	65.7 (3.9)	57.4 (4.2)	61.7 (3.0)	67.0 (3.9)	59.1 (4.4)	63.1 (3.0)
		(800, 20)	63.6 (2.6)	59.9 (2.7)	61.8 (1.8)	67.3 (2.4)	64.0 (2.6)	65.6 (1.8)
	0.5	(400, 10)	63.9 (3.6)	59.6 (4.0)	61.8 (2.7)	65.5 (3.2)	61.1 (4.0)	63.5 (2.5)
		(800, 20)	62.8 (2.6)	60.6 (2.5)	61.7 (1.8)	66.4 (2.5)	64.6 (2.6)	65.5 (1.7)

Sensitivity, specificity and overall accuracy are given as percentages.
Reported values are means and standard deviations over 500 simulations.

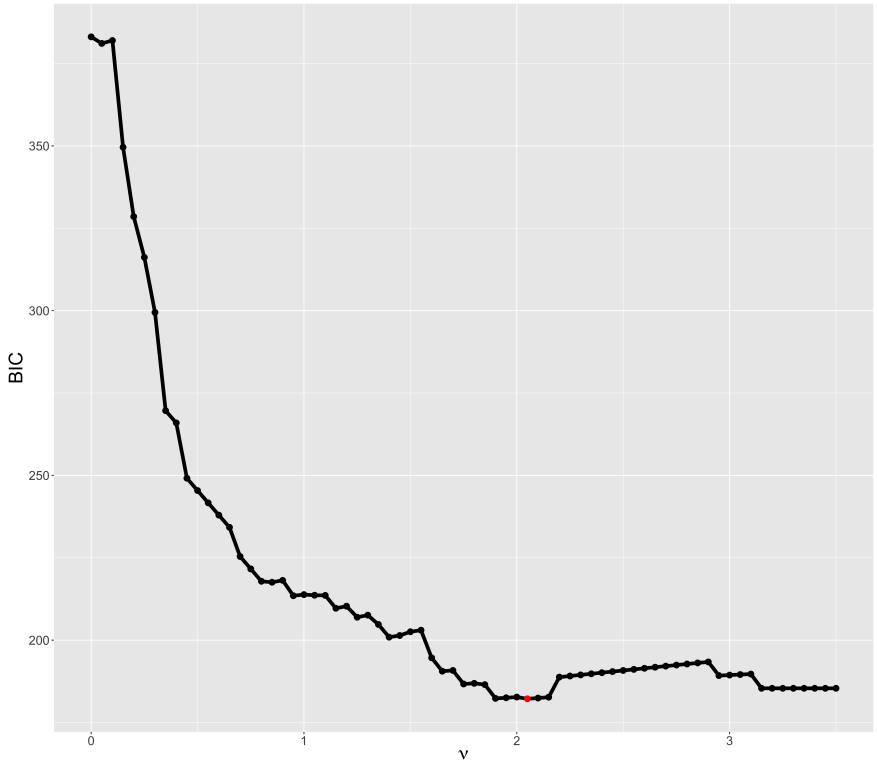
In Table 3.3, the first classifier performs similarly under both scenarios. The second classifier has better performance than the first classifier under scenario 1 but comparable performance under scenario 2. Also, the change of the baseline hazard function $\lambda_0(t)$ barely affects the performance of both classifiers. A similar pattern is also observed in Table 3.4 in high-dimensional settings. As for $\widehat{\boldsymbol{\theta}}$, it has similar accuracy with various baseline hazard functions $\lambda_0(t)$. However, when J gets larger, the bias of $\widehat{\alpha}$ increases a little, but the performance of $\widehat{\boldsymbol{\beta}}$ gets better. In summary, our classifiers perform similarly for both binary and normally distributed covariates.

Table 3.4: Classification and variable selection of proposed classifiers with high-dimensional continuous covariates.

			$I(\widehat{\xi}_i^{(0)} > 0.5)$			$I(\widehat{\xi}_i^{(1)} > 0.5)$			$\widehat{\alpha}$	$\widehat{\boldsymbol{\beta}}$		
Scenario	τ	$(n,\ J)$	Sensitivity	Specificity	Overall	Sensitivity	Specificity	Overall	Bias	Sensitivity	Specificity	Overall
1	-0.5	(100, 200)	85.8 (5.8)	29.8 (8.7)	59.2 (5.5)	98.7 (11.4)	99.7 (5.7)	99.5 (6.7)	0.44 (0.02)	69.5 (14.8)	57.3 (2.3)	58.5 (2.9)
		(200, 400)	88.7 (3.5)	27.1 (5.9)	60.0 (4.1)	100 (0)	100 (0)	100 (0)	0.47 (0.01)	82.0 (11.9)	60.4 (1.9)	60.9 (1.9)
	0	(100, 200)	83.4 (5.2)	33.6 (7.4)	59.0 (5.4)	99.0 (10.5)	99.6 (5.7)	99.1 (7.2)	0.45 (0.02)	70.8 (15.4)	57.3 (3.0)	57.9 (3.0)
		(200, 400)	85.2 (4.5)	31.9 (5.6)	59.6 (3.9)	100 (0)	100 (0)	100 (0)	0.47 (0.01)	82.7 (12.5)	59.3 (1.9)	59.9 (1.9)
	0.5	(100, 200)	81.9 (5.3)	37.5 (7.2)	60.1 (5.2)	98.3 (12.8)	99.7 (5.8)	99.0 (8.1)	0.44 (0.02)	71.4 (14.5)	56.1 (2.7)	56.9 (2.9)
		(200, 400)	84.5 (3.7)	34.0 (5.1)	59.6 (3.9)	100 (0)	100 (0)	100 (0)	0.47 (0.01)	85.0 (11.0)	58.6 (1.7)	59.2 (1.6)
2	-0.5	(100, 200)	85.5 (5.4)	29.3 (7.9)	58.9 (5.7)	94.2 (3.6)	23.4 (7.9)	60.8 (6.3)	0.43 (0.02)	62.3 (15.4)	64.8 (2.8)	64.6 (2.9)
		(200, 400)	84.0 (4.3)	32.0 (5.8)	59.5 (4.1)	96.3 (2.2)	31.7 (6.9)	65.8 (4.7)	0.47 (0.01)	75.6 (14.4)	68.0 (1.8)	68.2 (1.9)
	0	(100, 200)	82.9 (6.0)	34.2 (7.2)	59.5 (5.4)	92.9 (4.0)	27.7 (7.6)	61.7 (5.6)	0.44 (0.02)	64.8 (15.6)	64.1 (2.7)	64.1 (2.9)
		(200, 400)	81.3 (4.2)	36.5 (5.9)	59.6 (3.9)	95.7 (2.2)	35.7 (6.7)	66.5 (4.7)	0.47 (0.01)	76.8 (14.1)	67.4 (2.0)	67.7 (2.0)
	0.5	(100, 200)	82.0 (5.7)	37.8 (7.1)	60.1 (5.5)	92.5 (4.1)	31.0 (8.1)	62.1 (6.1)	0.45 (0.02)	63.8 (15.9)	63.7 (2.8)	63.7 (2.9)
		(200, 400)	79.9 (4.0)	38.5 (5.5)	59.5 (3.7)	95.1 (2.3)	37.8 (6.6)	66.9 (4.5)	0.46 (0.01)	77.1 (13.6)	66.7 (2.1)	67.4 (2.1)

Sensitivity, specificity and overall accuracy are given as percentages.
Reported values are means and standard deviations over 500 simulations.

Figure 3.3: The BIC curve with different values of the tuning parameter ν. The BIC attains its minimum at $\nu = 2.05$.



3.6 *Plasmodium vivax* Malaria Infection Study

As discussed in the introduction, it is essential to identify the cause of infection in *P. vivax* malaria research when the primary interest is treatment efficacy or effectiveness. In this section, we apply our proposed classifier to the *P. vivax* malaria data described in Section 3.1.1. We aim to classify the recurrent infection as either reinfection ($\epsilon_i = 1$) or relapse ($\epsilon_i = 2$). We first fit the cause-specific hazards model (3.3) and (3.4) with $\boldsymbol{X}_i$ as a vector of binary covariates that indicate whether a haplotype (genetic variant) is present or absent. Parameters $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}')'$ were estimated via the penalized partial likelihood function (3.5) with an L_1-penalty. To choose the optimal tuning parameter ν, we performed a grid search in the interval [0, 3.5] and calculated the corresponding Bayesian Information Criterion (BIC) values. Figure 3.3 shows the BIC curve with different values of the tuning parameter ν.

We report the classification results based on $\nu = 2.05$, where the BIC attains its minimum. In this case, two haplotypes (CAM.00 and CAM.04) were selected, with the proportional baseline coefficient $\exp(\widehat{\alpha}) = 0.686$. We also performed a sensitivity analysis by choosing $\nu = 0.8$, where the BIC curve begins

flatting out. In this case, 12 haplotypes (CAM.00, CAM.02 to CAM.10, CAM.12 and CAM.24) were selected with $\exp(\widehat{\alpha}) = 0.859$. The classification results based on $\nu = 0.8$ is reported in Section 3.8.

After we obtained $\widehat{\boldsymbol{\theta}}$, probabilities $\widehat{\xi}_{i1}^{(0)}$ and $\widehat{\xi}_{i2}^{(0)}$ were calculated based on formulae (3.6) and (3.7), respectively. For subjects with a recurrent infection, reading frequency for each haplotype presented at the baseline sequencing of the initial infection is used as the external covariate $\boldsymbol{W}_i$. Here, covariates $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are binary variables. When the recurrent infection is reinfection ($\epsilon_i = 1$), we assume $\boldsymbol{Z}_i$ is independent of $\boldsymbol{X}_i$ and $\boldsymbol{W}_i$, but follows the same distribution as $\boldsymbol{X}_i$. In this case, $\phi_i(1)$ can be estimated independently without using the pseudo-likelihood function (3.8), and the distribution of $\boldsymbol{Z}_i$ can be estimated using $\boldsymbol{X}_i$ alone.

To be specific, for $\epsilon_i = 1$, the transition likelihood function $\phi_i(1, \boldsymbol{\gamma}_1)$ can be written as

$$\phi_i(1, \boldsymbol{\gamma}_1) = f(\boldsymbol{z}_i | \epsilon_i = 1, dN_i(t) = 1, \boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{W}_i = \mathbf{w}_i) = \prod_{j=1}^{J} p_j^{z_{ij}} (1 - p_j)^{1 - z_{ij}},$$

where $p_j = P(X_{ij} = 1)$, $\boldsymbol{\gamma}_1 = (p_1, \ldots, p_J)'$. The parameter p_j can be consistently estimated by the sample mean $\widehat{p}_j = n^{-1} \sum_{i=1}^{n} x_{ij}$. Accordingly, the transition likelihood of reinfection can be estimated by $\phi_i(1, \widehat{\boldsymbol{\gamma}}_1) = \prod_{j=1}^{J} \widehat{p}_j^{z_{ij}} (1 - \widehat{p}_j)^{1 - z_{ij}}$.

For $\epsilon_i = 2$, when the recurrent infection is a relapse, we assume the transition likelihood follows the form of (3.10), that $\text{logit}(\mu_{ij2}) = q_{j20} + x_{ij} q_{j21} + w_{ij} x_{ij} q_{j21}^*$, with w_{ij} being the reading frequency of the jth haplotype of subject i when the haplotype is presented at the baseline sequencing, i.e., $x_{ij} = 1$. For computational simplicity, we assume that all haplotypes follow the same transition model, i.e., $q_{j20} = q_0$, $q_{j21} = q_1$, and $q_{j21}^* = q^*$ for all j. Then, we have $\phi_i(2, \boldsymbol{\gamma}_2) = f(\boldsymbol{z}_i | \epsilon_i = 2, dN_i(t) = 1, \boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{W}_i = \mathbf{w}_i) = \prod_{j=1}^{J} \mu_{ij2}^{z_{ij}} \big(1 - \mu_{ij2}\big)^{1 - z_{ij}}$, where $\mu_{ij2} = \exp(q_0 + x_{ij} q_1 + w_{ij} x_{ij} q^*) / \{1 + \exp(q_0 + x_{ij} q_1 + w_{ij} x_{ij} q^*)\}$ and $\boldsymbol{\gamma}_2 = (q_0, q_1, q^*)'$.

We replaced $\phi_i(1, \boldsymbol{\gamma}_1)$ in (3.8) by $\phi_i(1, \widehat{\boldsymbol{\gamma}}_1)$ and maximized the pseudo-likelihood function to obtain $\widehat{\boldsymbol{\gamma}}_2$. When using $\nu = 2.05$, we have $\widehat{q}_0 = -1.366$, $\widehat{q}_1 = 2.738$, and $\widehat{q}^* = 4.317$. When the recurrent infection is relapse, the parameter q_0 is the log odds of a subject whose baseline sequencing did not contain haplotype j ($x_{ij} = 0$) but the follow-up sequencing at the recurrence did ($z_{ij} = 1$). The estimate $\widehat{q}_0 = -1.366$ can be transformed into an estimated transition probability of 0.203, meaning there is 20% chance that the unseen haplotype at the baseline may show up at the recurrence when the cause is relapse. Since $\widehat{q}_0 + \widehat{q}_1 = 1.372$, it shows that there is around 80% chance of observing a haplotype again at the recurrence ($z_{ij} = 1$) when the cause is relapse and the haplotype appeared at the baseline ($x_{ij} = 1$). Since $\widehat{q}^* = 4.317$, it indicates that

there is more than 99% chance of observing the same haplotype again at the recurrence ($z_{ij} = 1$) when the reading frequency of the haplotype is more than 80% at the baseline ($w_{ij} = 0.8$). When using $\nu = 0.8$, we have $\widehat{q}_0 = -1.323$, $\widehat{q}_1 = 2.506$, and $\widehat{q}^* = 4.284$. These estimates are similar to those when using $\nu = 2.05$ and can be interpreted analogously.

Finally, we calculate $\widehat{\xi}_{ik}^{(1)}$ by (3.9) for $k = 1, 2$ and classify the recurrent event as relapse if $\widehat{\xi}_{i2}^{(1)} > \widehat{\xi}_{i1}^{(1)}$ and reinfection otherwise. Table 5 contains the classification results for the 23 subjects with recurrent infection based on our proposed method using $\nu = 2.05$. The tables include days to recurrence, baseline and recurrence haplotypes, the estimates $\widehat{\boldsymbol{\beta}}$, recurrence haplotype prevalence, two classification probabilities, and classification results from Lin, Li and Lin (2020), which analyzed the same data without utilizing the time to event information and external covariates in the estimation of transition likelihoods.

Our proposed method classifies 3 out of 23 recurrence pairs differently from Lin, Li and Lin (2020). The first pair is 87 $\rightarrow$ 87R, which was classified as relapse by Lin, Li and Lin (2020) but as reinfection by our classifier. Five variants showed up at the baseline sequencing, of which only CAM.00, the haplotype with the highest prevalence, showed up again in the recurrence sequencing. Also, the days to recurrence for this pair is 81 days, which is a relatively long time for relapse, suggesting that this recurrence event is more likely to be reinfection. The second pair is 123 $\rightarrow$ 123R, which was classified as reinfection by Lin, Li and Lin (2020) but as relapse by our classifier. Two haplotypes (CAM.00 and CAM.02) were observed at the baseline sequencing, and haplotype CAM.00 showed up again at the recurrence sequencing with CAM.01. Since only two haplotypes appeared at the recurrence, and CAM.00 is the most prevalent variant, the recurrent infection looks more likely to be a reinfection if not taking time to recurrent into consideration. However, the recurrent infection occurred only 26 days after the initial infection, which is a relatively short time compared to other reinfection cases. The only case classified as reinfection with a recurrent time less than 26 days was pair 160 $\rightarrow$ 160 R, with only 17 days to recurrence, but this is reasonable since there is no overlap between the baseline and recurrence variants. Notably, the pair 123 $\rightarrow$ 123R has 96% CAM.00 in the reading frequency at baseline, which supports the classification as relapse due to a high likelihood of observing the same variant in relapse if the variant has a high reading frequency at baseline, as suggested by large $\widehat{q}^*$. The last disparity comes from pair 153 $\rightarrow$ 153R, which was classified as relapse by Lin, Li and Lin (2020) but as reinfection by our classifier. There is no overlap between initial and recurrence variants. The time to recurrence is 115 days, which is longer than any case that was classified as relapse. The only case with days to recurrence longer than this pair is pair 151 $\rightarrow$ 151R, which was classified as reinfection by both Lin, Li and Lin (2020) and our

classifier. Therefore, it is more reasonable to classify pair 153 $\rightarrow$ 153R as reinfection. Overall, by considering the time to event and baseline haplotype reading frequency, our classifier achieves more consensus in this study.

Table 3.5: Classification of the first recurrent infection ($\nu = 2.05$).

Recurrence Pair	Days to Recurrence	Baseline Variants	$\widehat{\boldsymbol{\beta}}$	$\widehat{\xi}^{(0)}$	Recurrence Variants	Variant Prevalence	$\widehat{\xi}^{(1)}$	Class by our method	Class by Lin et al.
10 $\rightarrow$ 10R	84	CAM.00	0.907	0.783	CAM.00	0.590	0.995	Relapse	Relapse
		CAM.11	0		CAM.11	0.077			
					CAM.15	0.013			
31 $\rightarrow$ 31R	84	CAM.00	0.907	0.910	CAM.16	0.006	0.988	Relapse	Relapse
		CAM.02	0						
		CAM.04	1.026						
		CAM.31	0						
36 $\rightarrow$ 36R	99	CAM.00	0.907	0.910	CAM.01	0.269	0.645	Relapse	Relapse
		CAM.01	0		CAM.02	0.41			
		CAM.02	0		CAM.07	0.192			
		CAM.03	0		CAM.17	0.064			
		CAM.04	1.026						
		CAM.05	0						
		CAM.06	0						
		CAM.07	0						
		CAM.09	0						
		CAM.11	0						
68 $\rightarrow$ 68R	99	CAM.00	0.907	0.910	CAM.10	0.077	0.997	Relapse	Relapse
		CAM.02	0						
		CAM.04	1.026						
		CAM.10	0						
80 $\rightarrow$ 80R	56	CAM.00	0.907	0.910	CAM.00	0.590	0.000	Reinfection	Reinfection
		CAM.04	1.026		CAM.01	0.269			
		CAM.05	0		CAM.02	0.410			
		CAM.08	0		CAM.03	0.295			
		CAM.09	0		CAM.05	0.231			
		CAM.24	0		CAM.06	0.231			
		CAM.27	0		CAM.07	0.192			
					CAM.08	0.154			
					CAM.12	0.064			
					CAM.41	0.013			
81 $\rightarrow$ 81R	35	CAM.00	0.907	0.783	CAM.00	0.590	0.974	Relapse	Relapse
		CAM.01	0		CAM.01	0.269			

(Continued on next page)

Table 5 (continued from previous page)

Recurrence Pair	Days to Recurrence	Baseline Variants	$\widehat{\boldsymbol{\beta}}$	$\widehat{\xi}^{(0)}$	Recurrence Variants	Variant Prevalence	$\widehat{\xi}^{(1)}$	Class by our method	Class by Lin et al.
		CAM.51	0						
82 → 82R	56	CAM.00	0.907	0.910	CAM.00	0.590	0.674	Relapse	Relapse
		CAM.03	0		CAM.01	0.269			
		CAM.04	1.026		CAM.03	0.295			
		CAM.10	0		CAM.46	0.006			
87 → 87R	81	CAM.00	0.907	0.783	CAM.00	0.590	0.424	Reinfection	Relapse
		CAM.01	0		CAM.07	0.192			
		CAM.02	0		CAM.08	0.154			
		CAM.08	0		CAM.53	0.013			
		CAM.24	0						
89 → 89R	14	CAM.00	0.907	0.910	CAM.01	0.269	0.052	Reinfection	Reinfection
		CAM.04	1.026		CAM.09	0.077			
		CAM.06	0		CAM.20	0.026			
		CAM.08	0		CAM.27	0.038			
		CAM.10	0						
		CAM.12	0						
96 → 96R	71	CAM.00	0.907	0.910	CAM.00	0.590	0.983	Relapse	Relapse
		CAM.02	0		CAM.30	0.013			
		CAM.04	1.026						
		CAM.08	0						
112 → 112R	67	CAM.00	0.907	0.910	CAM.00	0.590	0.670	Relapse	Relapse
		CAM.01	0		CAM.01	0.269			
		CAM.02	0		CAM.02	0.410			
		CAM.04	1.026						
		CAM.07	0						
		CAM.12	0						
		CAM.40	0						
		CAM.42	0						
		CAM.60	0						
118 → 118R	89	CAM.08	0	0.593	CAM.01	0.269	0.008	Reinfection	Reinfection
					CAM.02	0.410			
					CAM.25	0.006			
					CAM.39	0.006			
123 → 123R	26	CAM.00	0.907	0.783	CAM.00	0.590	0.700	Relapse	Reinfection
		CAM.02	0		CAM.01	0.269			
125 → 125R	82	CAM.02	0	0.593	CAM.00	0.590	0.000	Reinfection	Reinfection
					CAM.01	0.269			

(Continued on next page)

Table 5 (continued from previous page)

Recurrence Pair	Days to Recurrence	Baseline Variants	$\widehat{\boldsymbol{\beta}}$	$\widehat{\xi}^{(0)}$	Recurrence Variants	Variant Prevalence	$\widehat{\xi}^{(1)}$	Class by our method	Class by Lin et al.
					CAM.02	0.410			
					CAM.04	0.346			
					CAM.09	0.077			
					CAM.13	0.006			
					CAM.14	0.026			
					CAM.38	0.006			
					CAM.45	0.006			
126 → 126R	85	CAM.00	0.907	0.910	CAM.01	0.269	0.975	Relapse	Relapse
		CAM.01	0		CAM.07	0.192			
		CAM.02	0		CAM.33	0.006			
		CAM.03	0						
		CAM.04	1.026						
		CAM.05	0						
		CAM.06	0						
		CAM.07	0						
		CAM.22	0						
		CAM.50	0						
130 → 130R	68	CAM.00	0.907	0.910	CAM.00	0.590	0.997	Relapse	Relapse
		CAM.02	0		CAM.04	0.346			
		CAM.03	0		CAM.12	0.064			
		CAM.04	1.026						
		CAM.12	0						
151 → 151R	126	CAM.03	0	0.593	CAM.00	0.590	0.325	Reinfection	Reinfection
		CAM.05	0		CAM.08	0.154			
		CAM.08	0		CAM.14	0.026			
					CAM.64	0.006			
152 → 152R	94	CAM.00	0.907	0.783	CAM.00	0.590	0.153	Reinfection	Reinfection
		CAM.01	0		CAM.01	0.269			
					CAM.05	0.231			
					CAM.07	0.192			
153 → 153R	115	CAM.00	0.907	0.910	CAM.02	0.410	0.425	Reinfection	Relapse
		CAM.04	1.026		CAM.20	0.026			
		CAM.07	0						
		CAM.55	0						
154 → 154R	64	CAM.00	0.907	0.783	CAM.03	0.295	0.116	Reinfection	Reinfection
		CAM.06	0		CAM.05	0.231			
		CAM.57	0		CAM.06	0.231			

(Continued on next page)

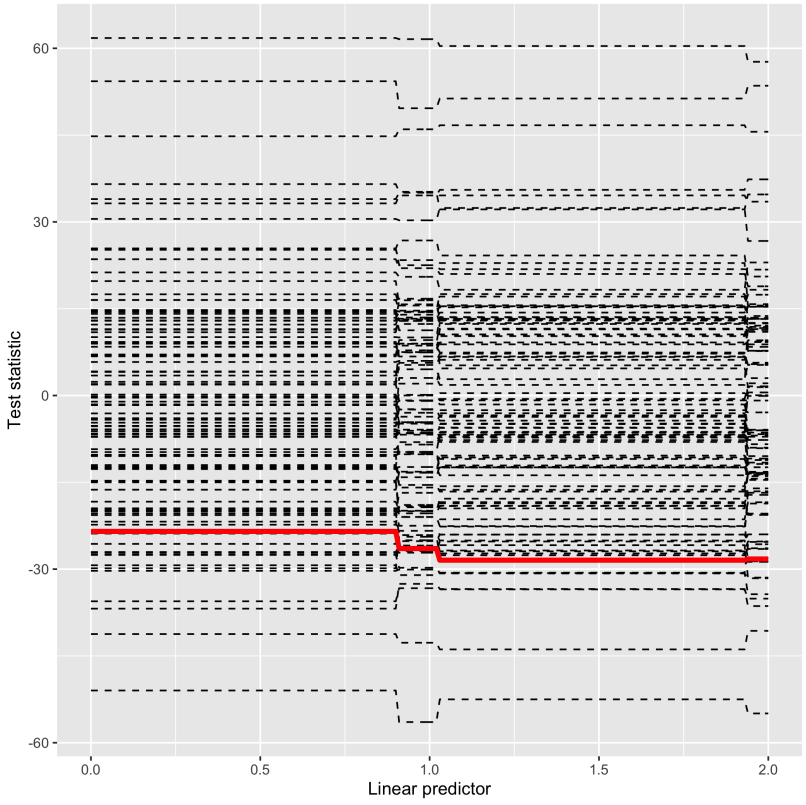
Table 5 (continued from previous page)

Recurrence Pair	Days to Recurrence	Baseline Variants	$\widehat{\boldsymbol{\beta}}$	$\widehat{\xi}^{(0)}$	Recurrence Variants	Variant Prevalence	$\widehat{\xi}^{(1)}$	Class by our method	Class by Lin et al.
160 → 160R	17	CAM.02	0	0.803	CAM.00	0.590	0.000	Reinfection	Reinfection
		CAM.04	1.026		CAM.03	0.295			
		CAM.07	0		CAM.05	0.231			
					CAM.10	0.077			
					CAM.61	0.006			
177 → 177R	84	CAM.00	0.907	0.910	CAM.01	0.269	0.773	Relapse	Relapse
		CAM.04	1.026						
		CAM.07	0						
179 → 179R	84	CAM.03	0	0.593	CAM.01	0.269	0.234	Reinfection	Reinfection
		CAM.05	0		CAM.13	0.006			
		CAM.07	0						
		CAM.09	0						
		CAM.17	0						
		CAM.22	0						

We evaluate the proposed models (3.3) and (3.4) for the *P. vivax* malaria data using the proposed method in Section 3.2. For a sequence of x in the range of the linear predictor $\widehat{\boldsymbol{\beta}}' \boldsymbol{X}_i$, we calculate the test statistic $T(x) = \sum_{i=1}^{n} I(\widehat{\boldsymbol{\beta}}' \boldsymbol{X}_i \leq x)\widehat{M}_i$, where $\widehat{M}_i$ is the martingale residual defined in Section 2. Using a Monte-Carlo simulation with $Q_i (i = 1, \ldots, n)$ sampled independently from the standard normal distribution, the confidence band for $T(x)$ can be constructed by calculating $T_Q(x) = \sum_{i=1}^{n} I(\widehat{\boldsymbol{\beta}}' \boldsymbol{X}_i \leq x)\widehat{M}_i Q_i$. We simulate the process of $T(x)$ by repeating the sampling. Using $\nu = 2.05$, the linear predictor $\widehat{\boldsymbol{\beta}}' \boldsymbol{X}_i$ ranges from 0 to 1.94. Figure 3.4 shows the result with observed $T(x)$ (thick solid line) and 100 simulated curves (dashed lines) for $x \in [0, 1.94]$. The test statistics are point-wisely within the simulated processes, with no significant indication of model violation. The model diagnosis result for the sensitivity analysis when $\nu = 0.8$ is provided in Section 3.8. Similarly, there is no significant model violation when using $\nu = 0.8$ as well.

Misidentification of unique haplotypes is a concern in the current analysis. Low-frequency minority genetic variants that only differ in sequence by one nucleotide base pair to common variants may represent false haplotypes generated by sequencing error. We adjusted the stringency of criteria used for calling haplotypes to "collapse" such variants together, reducing the total number of 67 unique haplotypes to 32 (Hathaway et al., 2018). As a sensitivity analysis, we also analyzed the data with this total number of 32 haplotypes, based on collapsing variants with 1-nucleotide apart within the same isolate. The classification

Figure 3.4: Goodness-of-fit model diagnosis for the *P. vivax* malaria data using $\nu = 2.05$



result has several disparities with that using 67 haplotypes but mostly agreed with the one based on the method in Lin, Li and Lin (2020) also using 32 haplotypes. It is not surprising to find the classification result sensitive to the identification of haplotype since our method relies on the modeling of the transition between variants. The collapse of variants and corresponding classification results using 32 haplotypes are provided in Section 3.8.

3.7 Discussion

We proposed a classification method for identifying the latent cause of events under competing risks set-up, which utilizes both time to event and transition likelihood information for better classification performance. By considering the transition likelihood, we utilize more information when constructing the classifier, which leads to better performance than the classifier using only baseline information. The method can be applied regardless of the true form of the baseline hazard function, and can also be applied to a variety of covariate data types. We examined the performance of our method through simulation studies under various settings as well as real data analysis, which shows high reliability of our method.

When modeling the outcomes of competing risks, we assumed a proportional hazards model with a common baseline hazard function for every cause-specific hazard. When the hazards share the same covariates, the model may not be identifiable. To avoid the identifiability issue when analyzing the *P. vivax* malaria data, we assume the reinfection process is independent of any baseline covariates in $\boldsymbol{X}_i$ but has a hazard function proportional to a baseline hazard $\lambda_0(t)$. This assumption is reasonable for our data but may not be ideal for a general case. Also, in our current approach, we assume the transition of covariates is independent of time. It will be of interest to generalize the transition model to be a function of time. A possible approach is to include time t_i as a covariate in the model for μ_{ijk}. This approach is somehow restricted to a linear function of time, which is subject to model misspecification.

Another topic worth investigating is the statistical inference of regression coefficients $\boldsymbol{\beta}$. The treatment effectiveness relies on formal statistical testing on the treatment effect. While the current method selects the variables with high accuracy, statistical inferences such as producing p-values and 95% confidence intervals need more work.

3.8 Additional results of *Plasmodium vivax* malaria infection study

We here give more details about the *Plasmodium vivax* Malaria Infection study. Table 3.6 shows the classification results given by our method when using 67 haplotypes and $\nu = 0.8$. Table 3.7 details the haplotypes that were collapsed with other haplotypes. Any other haplotypes not shown in the first column of the table were not collapsed. Table 3.8 shows the classification results given by our method when using 32 haplotypes, in which case the BIC attains its minimum at $\nu = 1.6$. Figure 3.5 shows the goodness-of-fit test result when using $\nu = 0.8$. The test statistics are point-wisely within the simulated processes, with no significant pattern of model violation.

Table 3.6: Classification of the first recurrent infection based on our proposed method ($\nu = 0.8$).

Recurrence Pair	Days to Recurrence	Baseline Variants	$\widehat{\beta}$	$\widehat{\xi}^{(0)}$	Recurrence Variants	Variant Prevalence	$\widehat{\xi}^{(1)}$	Class by our method	Class by Lin et al.
10 → 10R	84	CAM.00	1.194	0.793	CAM.00	0.590	0.996	Relapse	Relapse
		CAM.01	0		CAM.11	0.077			
					CAM.15	0.013			
31 → 31R	84	CAM.00	1.194	0.935	CAM.16	0.006	0.992	Relapse	Relapse
		CAM.02	0.075						
		CAM.04	1.245						
		CAM.31	0						
36 → 36R	99	CAM.00	1.194	0.897	CAM.01	0.269	0.628	Relapse	Relapse
		CAM.01	0		CAM.02	0.41			
		CAM.02	0.075		CAM.07	0.192			
		CAM.03	-0.293		CAM.17	0.064			
		CAM.04	1.245						
		CAM.05	-0.274						
		CAM.06	-0.292						
		CAM.07	0.287						
		CAM.09	0.068						
		CAM.11	0						
68 → 68R	99	CAM.00	1.194	0.936	CAM.10	0.077	0.998	Relapse	Relapse
		CAM.02	0.075						
		CAM.04	1.245						
		CAM.10	0.022						
80 → 80R	56	CAM.00	1.194	0.951	CAM.00	0.590	0.000	Reinfection	Reinfection
		CAM.04	1.245		CAM.01	0.269			
		CAM.05	-0.274		CAM.02	0.410			
		CAM.08	0.384		CAM.03	0.295			

(Continued on next page)

Table 3.6 (continued from previous page)

Recurrence Pair	Days to Recurrence	Baseline Variants	$\widehat{\beta}$	$\widehat{\xi}^{(0)}$	Recurrence Variants	Variant Prevalence	$\widehat{\xi}^{(1)}$	Class by our method	Class by Lin et al.
		CAM.09	0.068		CAM.05	0.231			
		CAM.24	0.207		CAM.06	0.231			
		CAM.27	0		CAM.07	0.192			
					CAM.08	0.154			
					CAM.12	0.064			
					CAM.41	0.013			
81 → 81R	35	CAM.00	1.194	0.793	CAM.00	0.590	0.975	Relapse	Relapse
		CAM.01	0		CAM.01	0.269			
		CAM.51	0						
82 → 82R	56	CAM.00	1.194	0.910	CAM.00	0.590	0.670	Relapse	Relapse
		CAM.03	-0.293		CAM.01	0.269			
		CAM.04	1.245		CAM.03	0.295			
		CAM.10	0.022		CAM.46	0.006			
87 → 87R	81	CAM.00	1.194	0.882	CAM.00	0.590	0.616	Relapse	Relapse
		CAM.01	0		CAM.07	0.192			
		CAM.02	0.075		CAM.08	0.154			
		CAM.08	0.384		CAM.53	0.013			
		CAM.24	0.207						
89 → 89R	14	CAM.00	1.194	0.953	CAM.01	0.269	0.109	Reinfection	Reinfection
		CAM.04	1.245		CAM.09	0.077			
		CAM.06	-0.292		CAM.20	0.026			
		CAM.08	0.384		CAM.27	0.038			
		CAM.10	0.022						
		CAM.12	0.307						
96 → 96R	71	CAM.00	1.194	0.955	CAM.00	0.590	0.992	Relapse	Relapse
		CAM.02	0.075		CAM.30	0.013			
		CAM.04	1.245						
		CAM.08	0.384						
112 → 112R	67	CAM.00	1.194	0.963	CAM.00	0.590	0.847	Relapse	Relapse
		CAM.01	0		CAM.01	0.269			
		CAM.02	0.075		CAM.02	0.410			
		CAM.04	1.245						
		CAM.07	0.287						
		CAM.12	0.307						
		CAM.40	0						
		CAM.42	0						
		CAM.60	0						

(Continued on next page)

Table 3.6 (continued from previous page)

Recurrence Pair	Days to Recurrence	Baseline Variants	$\widehat{\beta}$	$\widehat{\xi}^{(0)}$	Recurrence Variants	Variant Prevalence	$\widehat{\xi}^{(1)}$	Class by our method	Class by Lin et al.
118 → 118R	89	CAM.08	0.384	0.631	CAM.01	0.269	0.012	Reinfection	Reinfection
					CAM.02	0.410			
					CAM.25	0.006			
					CAM.39	0.006			
123 → 123R	26	CAM.00	1.194	0.805	CAM.00	0.590	0.720	Relapse	Reinfection
		CAM.02	0.075		CAM.01	0.269			
125 → 125R	82	CAM.02	0.075	0.556	CAM.00	0.590	0.000	Reinfection	Reinfection
					CAM.01	0.269			
					CAM.02	0.410			
					CAM.04	0.346			
					CAM.09	0.077			
					CAM.13	0.006			
					CAM.14	0.026			
					CAM.38	0.006			
					CAM.45	0.006			
126 → 126R	85	CAM.00	1.194	0.890	CAM.01	0.269	0.968	Relapse	Relapse
		CAM.01	0		CAM.07	0.192			
		CAM.02	0.075		CAM.33	0.006			
		CAM.03	-0.293						
		CAM.04	1.245						
		CAM.05	-0.274						
		CAM.06	-0.292						
		CAM.07	0.287						
		CAM.22	0						
		CAM.50	0						
130 → 130R	68	CAM.00	1.194	0.936	CAM.00	0.590	0.997	Relapse	Relapse
		CAM.02	0.075		CAM.04	0.346			
		CAM.03	-0.293		CAM.12	0.064			
		CAM.04	1.245						
		CAM.12	0.307						
151 → 151R	126	CAM.03	-0.293	0.492	CAM.00	0.590	0.242	Reinfection	Reinfection
		CAM.05	-0.274		CAM.08	0.154			
		CAM.08	0.384		CAM.14	0.026			
					CAM.64	0.006			
152 → 152R	94	CAM.00	1.194	0.793	CAM.00	0.590	0.157	Reinfection	Reinfection
		CAM.01	0		CAM.01	0.269			
					CAM.05	0.231			

(Continued on next page)

Table 3.6 (continued from previous page)

Recurrence Pair	Days to Recurrence	Baseline Variants	$\widehat{\beta}$	$\widehat{\xi}^{(0)}$	Recurrence Variants	Variant Prevalence	$\widehat{\xi}^{(1)}$	Class by our method	Class by Lin et al.
					CAM.07	0.192			
153 → 153R	115	CAM.00	1.194	0.947	CAM.02	0.410	0.586	Relapse	Relapse
		CAM.04	1.245		CAM.20	0.026			
		CAM.07	0.287						
		CAM.55	0						
154 → 154R	64	CAM.00	1.194	0.741	CAM.03	0.295	0.098	Reinfection	Reinfection
		CAM.06	-0.292		CAM.05	0.231			
		CAM.57	0		CAM.06	0.231			
160 → 160R	17	CAM.02	0.075	0.853	CAM.00	0.590	0.000	Reinfection	Reinfection
		CAM.04	1.245		CAM.03	0.295			
		CAM.07	0.287		CAM.05	0.231			
					CAM.10	0.077			
					CAM.61	0.006			
177 → 177R	84	CAM.00	1.194	0.947	CAM.01	0.269	0.864	Relapse	Relapse
		CAM.04	1.245						
		CAM.07	0.287						
179 → 179R	84	CAM.03	-0.293	0.485	CAM.01	0.269	0.165	Reinfection	Reinfection
		CAM.05	-0.274		CAM.13	0.006			
		CAM.07	0.287						
		CAM.09	0.068						
		CAM.17	0						
		CAM.22	0						

Table 3.7: Collapsing of original 67 haplotypes to 32 haplotypes.

Original haplotype	Collapse to
CAM.05	CAM.00
CAM.12	CAM.00
CAM.24	CAM.00
CAM.46	CAM.00
CAM.51	CAM.00
CAM.54	CAM.00
CAM.57	CAM.00
CAM.61	CAM.00
CAM.62	CAM.00
CAM.25	CAM.01
CAM.26	CAM.01
CAM.43	CAM.01
CAM.44	CAM.01
CAM.63	CAM.01
CAM.13	CAM.02
CAM.31	CAM.02
CAM.32	CAM.02
CAM.34	CAM.02
CAM.38	CAM.02
CAM.40	CAM.02
CAM.49	CAM.02
CAM.60	CAM.02
CAM.56	CAM.04
CAM.58	CAM.04
CAM.37	CAM.06
CAM.42	CAM.06
CAM.55	CAM.06
CAM.64	CAM.06
CAM.15	CAM.07
CAM.39	CAM.07
CAM.41	CAM.07
CAM.50	CAM.07
CAM.17	CAM.08
CAM.59	CAM.09
CAM.45	CAM.10

Table 3.8: Classification of first recurrent infection based on our proposed method when using 32 haplotypes ($\nu = 1.6$).

Recurrence Pair	Days to Recurrence	Baseline Variants	$\widehat{\beta}$	$\widehat{\xi}^{(0)}$	Recurrence Variants	Variant Prevalence	$\widehat{\xi}^{(1)}$	Class by our method	Class by Lin et al.
10 → 10R	84	CAM.00	0.862	0.768	CAM.00	0.679	0.927	Relapse	Relapse
		CAM.11	0		CAM.07	0.218			
					CAM.11	0.077			
31 → 31R	84	CAM.00	0.862	0.913	CAM.16	0.006	0.981	Relapse	Relapse
		CAM.02	0						
		CAM.04	1.157						
36 → 36R	99	CAM.00	0.862	0.913	CAM.01	0.321	0.393	Reinfection	Relapse
		CAM.01	0		CAM.02	0.449			
		CAM.02	0		CAM.07	0.218			
		CAM.03	0		CAM.08	0.218			
		CAM.04	1.157						
		CAM.06	0						
		CAM.07	0						
		CAM.09	0						
		CAM.11	0						
68 → 68R	99	CAM.00	0.862	0.913	CAM.10	0.077	0.995	Relapse	Relapse
		CAM.02	0						
		CAM.04	1.157						
		CAM.10	0						
80 → 80R	56	CAM.00	0.862	0.913	CAM.00	0.679	0	Reinfection	Reinfection
		CAM.04	1.157		CAM.01	0.321			
		CAM.08	0		CAM.02	0.449			
		CAM.09	0		CAM.03	0.295			
		CAM.27	0		CAM.06	0.269			
					CAM.07	0.218			
					CAM.08	0.218			
81 → 81R	35	CAM.00	0.862	0.768	CAM.00	0.679	0.942	Relapse	Relapse
		CAM.01	0		CAM.01	0.321			
82 → 82R	56	CAM.00	0.862	0.913	CAM.00	0.679	0.723	Relapse	Relapse
		CAM.03	0		CAM.01	0.321			
		CAM.04	1.157		CAM.03	0.295			
		CAM.10	0						
87 → 87R	81	CAM.00	0.862	0.768	CAM.00	0.679	0.461	Reinfection	Relapse
		CAM.01	0		CAM.07	0.218			
		CAM.02	0		CAM.08	0.218			
		CAM.08	0		CAM.53	0.006			

(Continued on next page)

Table 3.8 (continued from previous page)

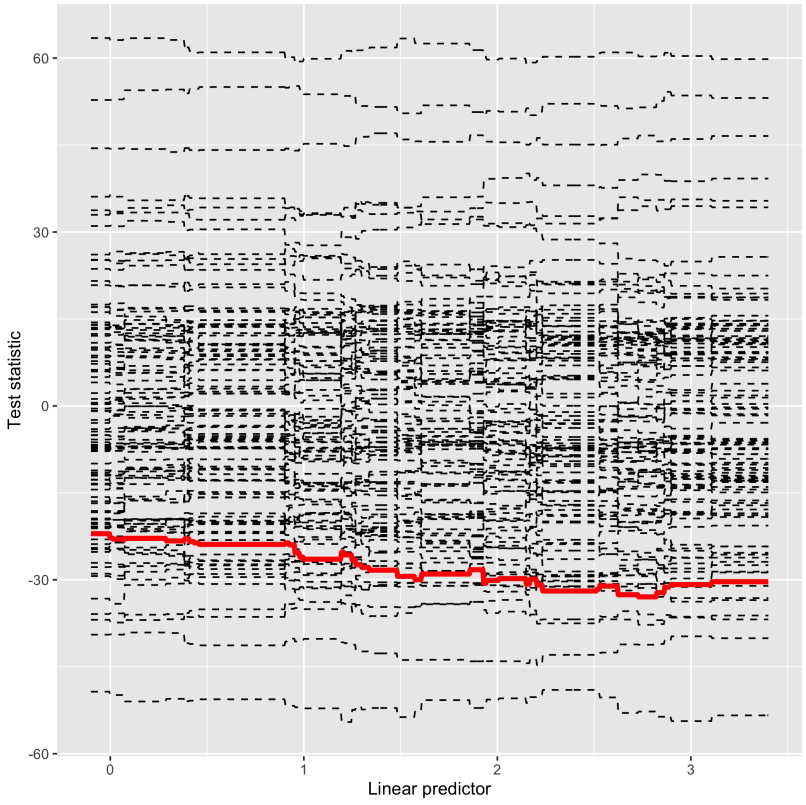
Recurrence Pair	Days to Recurrence	Baseline Variants	$\widehat{\beta}$	$\widehat{\xi}^{(0)}$	Recurrence Variants	Variant Prevalence	$\widehat{\xi}^{(1)}$	Class by our method	Class by Lin et al.
89 → 89R	14	CAM.00	0.862	0.913	CAM.01	0.321	0.7	Relapse	Reinfection
		CAM.04	1.157		CAM.09	0.077			
		CAM.06	0		CAM.20	0.026			
		CAM.08	0		CAM.27	0.038			
		CAM.10	0						
96 → 96R	71	CAM.00	0.862	0.913	CAM.00	0.679	0.989	Relapse	Relapse
		CAM.02	0		CAM.30	0.013			
		CAM.04	1.157						
		CAM.08	0						
112 → 112R	67	CAM.00	0.862	0.913	CAM.00	0.679	0.739	Relapse	Relapse
		CAM.01	0		CAM.01	0.321			
		CAM.02	0		CAM.02	0.449			
		CAM.04	1.157						
		CAM.06	0						
		CAM.07	0						
118 → 118R	89	CAM.08	0	0.583	CAM.01	0.321	0.003	Reinfection	Reinfection
					CAM.02	0.449			
					CAM.07	0.218			
123 → 123R	26	CAM.00	0.862	0.768	CAM.00	0.679	0.702	Relapse	Relapse
		CAM.02	0		CAM.01	0.321			
125 → 125R	82	CAM.02	0	0.583	CAM.00	0.679	0.001	Reinfection	Reinfection
					CAM.01	0.321			
					CAM.02	0.449			
					CAM.04	0.359			
					CAM.09	0.077			
					CAM.10	0.077			
					CAM.14	0.026			
126 → 126R	85	CAM.00	0.862	0.913	CAM.01	0.321	0.969	Relapse	Relapse
		CAM.01	0		CAM.07	0.218			
		CAM.02	0		CAM.33	0.006			
		CAM.03	0						
		CAM.04	1.157						
		CAM.06	0						
		CAM.07	0						
		CAM.22	0						
130 → 130R	68	CAM.00	0.862	0.913	CAM.00	0.679	0.98	Relapse	Relapse
		CAM.02	0		CAM.04	0.359			

(Continued on next page)

Table 3.8 (continued from previous page)

Recurrence Pair	Days to Recurrence	Baseline Variants	$\widehat{\beta}$	$\widehat{\xi}^{(0)}$	Recurrence Variants	Variant Prevalence	$\widehat{\xi}^{(1)}$	Class by our method	Class by Lin et al.
		CAM.03	0						
		CAM.04	1.157						
151 → 151R	126	CAM.00	0.862	0.768	CAM.00	0.679	0.828	Relapse	Relapse
		CAM.03	0		CAM.06	0.269			
		CAM.08	0		CAM.08	0.218			
					CAM.14	0.026			
152 → 152R	94	CAM.00	0.862	0.768	CAM.00	0.679	0.799	Relapse	Relapse
		CAM.01	0		CAM.01	0.321			
					CAM.07	0.218			
153 → 153R	115	CAM.00	0.862	0.913	CAM.02	0.449	0.76	Relapse	Relapse
		CAM.04	1.157		CAM.20	0.026			
		CAM.06	0						
		CAM.07	0						
154 → 154R	64	CAM.00	0.862	0.768	CAM.00	0.679	0.771	Relapse	Relapse
		CAM.06	0		CAM.03	0.295			
					CAM.06	0.269			
160 → 160R	17	CAM.02	0	0.816	CAM.00	0.679	0.009	Reinfection	Reinfection
		CAM.04	1.157		CAM.03	0.295			
		CAM.07	0		CAM.10	0.077			
177 → 177R	84	CAM.00	0.862	0.913	CAM.01	0.321	0.815	Relapse	Relapse
		CAM.04	1.157						
		CAM.07	0						
179 → 179R	84	CAM.00	0.862	0.768	CAM.01	0.321	0.046	Reinfection	Reinfection
		CAM.03	0		CAM.02	0.449			
		CAM.07	0						
		CAM.08	0						
		CAM.09	0						
		CAM.22	0						

Figure 3.5: Goodness-of-fit model diagnosis for the *P. vivax* malaria data using $\nu = 0.8$



CHAPTER 4: DECOMPOSITION OF CORRELATIONS OF MIXED VARIABLES BY A LATENT MIXED GAUSSIAN COPULA MODEL

4.1 Introduction

With the rapid development of technology, high-dimensional multi-omics data can be collected from the same subject, such as genomics (DNA methylation, copy number variation and single nucleotide polymorphism), transcriptomics (mRNA expression and microRNA expression), proteomics, and metabolomics data. Much evidence has demonstrated the benefit of integrating these data in an analysis. However, in practice, such an integrative analysis can be challenging because multi-omics data can be of different types and at different scales. It is especially challenging when seeking to identify the common and differential networks between two or more subject groups. An example is a *Chlamydia trachomatis* genital tract infection study. Chlamydia is the leading bacterial sexually transmitted infection in the United States and the infection is often asymptomatic. In up to 50% of women, untreated infection can ascend from the cervix to the upper genital tract and potentially lead to severe female reproductive morbidities. Identification of the commonly and differentially expressed genes and their underlying regulatory SNPs between women with and without ascending infection can greatly enhance the understanding of disease.

To study the correlations among mixed types of variables, many new methods have been developed. Fan et al. (2017) proposed a latent Gaussian copula model to measure the correlations between binary and continuous variables. They assumed that the observed binary and continuous variables are driven by some latent variables that follow the nonparanormal distribution (Liu, Lafferty and Wasserman, 2009). Under such an assumption, Fan et al. (2017) proposed to use Kendall's τ, a semiparametric rank-based correlation coefficient estimator, to measure the latent correlations between binary and continuous variables. Yoon, Carroll and Gaynanova (2020) further extended the method to incorporate truncated variables and developed a rank-based estimator that can be used in the Canonical Correlation Analysis. However, both works only considered the situation when there is only one population. They did not consider decomposing the variation into common and group-specific components.

To find the common and idiosyncratic variation of mixed variables for multiple groups, we propose a two-step method. First, we propose a Latent Mixed Gaussian Copula model to incorporate binary, categorical, continuous, and truncated variables under a unified framework. Then, we develop rank-based estimators of the correlation matrices for each group. Next, we propose to decompose such correlation matrices as a sum of a low-rank matrix that captures the group-specific variation for each group and a sparse matrix that captures the common variation across all groups. Such a decomposition is done by solving a penalized M-estimation problem. We propose to view the decomposition step as a de-noising process that after removing the shared variation, the low-rank group-specific components can give a clearer view of the differences between groups. Another benefit of our method is that it allows different types of data being analyzed in a unified framework.

The rest of this chapter is organized as follows. In Section 4.2, we describe the formulation and solution of our proposed method in details. In Section 4.3, we carry out extensive simulation studies to compare our method with some competitive methods. In Section 4.4, we apply our method to a *Chlamydia trachomatis* genital tract infection study to demonstrate how it can be used to find useful biomarkers that differentiate subtypes of patients.

4.2 Methodology

Without loss of generality, we consider two groups of subjects. For the g-th group, assume that we observe a p-dimensional vector $\boldsymbol{X}_g = (X_{g,1}, \ldots, X_{g,p})^T$ containing variables of mixed types, such as continuous, binary, categorical, or truncated variables. We assume that $\boldsymbol{X}_g$ is derived from a vector of latent continuous variables $\boldsymbol{Y}_g = (Y_{g,1}, \ldots, Y_{g,p})^T$ by the transformation function $\boldsymbol{h}_g = (h_{g,1}, \ldots, h_{g,p})^T$ that

$$X_{g,j} = h_{g,j}(Y_{g,j}) = \begin{cases} Y_{g,j}, & \text{if } j \in \mathscr{C}; \\ I(Y_{g,j} > C_{g,j}), & \text{if } j \in \mathscr{B}; \\ I(Y_{g,j} > D_{g,j}) Y_{g,j}, & \text{if } j \in \mathscr{T}; \\ \sum_{l=1}^{L-1} I(Y_{g,j} > C_{g,j,l}), & \text{if } j \in \mathscr{G}; \end{cases} \tag{4.1}$$

where $\mathscr{C}$, $\mathscr{B}$, $\mathscr{T}$, and $\mathscr{G}$ are the index sets for continuous, binary, truncated, and categorical variables respectively, and $\{C_{g,j}\}_{j\in\mathscr{B}}$, $\{D_{g,j}\}_{j\in\mathscr{T}}$ and $\{C_{g,j,l}\}_{j\in\mathscr{G}, 1\le l\le L_j-1}$ are the corresponding cut-offs. We assume that the latent $\boldsymbol{Y}_g$ follows a Gaussian Copula model proposed by Liu, Lafferty and Wasserman

(2009). More specifically, we assume that there exists some monotonically increasing functions $\boldsymbol{f}_g = (f_{g,1}, \ldots, f_{g,p})^T$ such that $(f_{g,1}(Y_{g,1}), \ldots, f_{g,p}(Y_{g,p}))^T \sim N(\mathbf{0}, \boldsymbol{R}_g)$, where $\boldsymbol{R}_g$ is a correlation matrix. We call (4.1) as the Latent Mixed Gaussian Copula (LMGC) model for mixed data. In the existing literature, Fan et al. (2017) studied the LMGC model for continuous and binary variables only. Yoon, Carroll and Gaynanova (2020) further extended it to incorporate truncated variables. In all these works, the authors developed consistent estimators of the latent correlation matrix, and further applied these estimators in some unsupervised problems, such as the Canonical Correlation Analysis. However, we would like to point out that these works only deal with a single set of samples.

Different from these works, we propose to use the LMGC model to decompose the latent correlation matrix into a low-rank and a sparse matrices that capture the group-specific and common variation among mixed variables respectively. The LMGC model transforms the observed mixed variables into latent multivariate normal variables. Then, we perform the decomposition based on the correlation matrix of the latent variables. We emphasize that even though the latent variables themselves are not observable, it is still feasible to decompose its correlation matrix. Indeed, such a decomposition is motivated by factor analysis. We assume that the latent variables $\boldsymbol{f}_g(\boldsymbol{Y}_g)$ follow a factor decomposition that

$$\boldsymbol{f}_g(\boldsymbol{Y}_g) = \boldsymbol{\Lambda}_g \boldsymbol{F}_g + \boldsymbol{U}, \tag{4.2}$$

where $\boldsymbol{F}_g \in \mathbb{R}^{r_g}$ is the group-specific latent factors from group g, $\boldsymbol{\Lambda}_g \in \mathbb{R}^{p \times r_g}$ is the loading matrix, r_g is the number of latent factors in group g, and $\boldsymbol{U} \in \mathbb{R}^p$ is the idiosyncratic component, which is assumed to be uncorrelated with $\boldsymbol{F}_g$. To avoid the identifiability issue, we adopt the standard identifiability conditions in the factor analysis literature by assuming that $\mathrm{cov}(\boldsymbol{F}_g) = \boldsymbol{I}_{r_g}$ and $\boldsymbol{\Lambda}_g' \boldsymbol{\Lambda}_g$ is a diagonal matrix for $g \in \{1, 2\}$. In factor decomposition (4.2), we assume that the group-specific variation is induced by the latent factor $\boldsymbol{F}_g$ and the common variation is induced by the idiosyncratic component $\boldsymbol{U}$ that is shared in the two groups. Then, it follows from (4.2) that

$$\boldsymbol{R}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Sigma}_U. \tag{4.3}$$

A similar idea of variation decomposition by factor model has also been used in Fan et al. (2018). We can think of (4.3) as a de-noising step of $\boldsymbol{R}_g$. That is, after removing the common $\boldsymbol{\Sigma}_U$, the real group-specific variation are contained in $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g'$.

Next, we show in Proposition 1 that (4.3) is a well-defined problem, in the sense that even if $\boldsymbol{f}_g(\boldsymbol{Y}_g)$ has different decomposition in (4.2), the decomposition of $\boldsymbol{R}_g$ in (4.3) is still unique. Furthermore, we demonstrate in Section 4.2.1 that the decomposition in (4.3) does not require $\boldsymbol{f}_g(\boldsymbol{Y}_g)$ to be observable.

Proposition 1 *For $g \in \{1, 2\}$, suppose r_g is fixed, and $\boldsymbol{f}_g(\boldsymbol{Y}_g) = \boldsymbol{\Lambda}_g \boldsymbol{F}_g + \boldsymbol{U} = \widetilde{\boldsymbol{\Lambda}}_g \widetilde{\boldsymbol{F}}_g + \widetilde{\boldsymbol{U}}$, where $(\boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{U})$ and $(\widetilde{\boldsymbol{F}}_1, \widetilde{\boldsymbol{F}}_2, \widetilde{\boldsymbol{U}})$ are both mutually uncorrelated. Then, $\boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g' = \widetilde{\boldsymbol{\Lambda}}_g \widetilde{\boldsymbol{\Lambda}}_g'$ and $\widetilde{\boldsymbol{\Sigma}}_U = \boldsymbol{\Sigma}_U$.*

4.2.1 Decomposition of the latent correlation matrices

To solve the decomposition problem (4.3), we need an estimator of $\boldsymbol{R}_g$. In Section 4.2.2, we give more details on how such an estimator can be obtained using some rank-based method. Given such an estimator, we let $\ell(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_U) = (1/2)\|\widehat{\boldsymbol{R}}_1 + \widehat{\boldsymbol{R}}_2 - \boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2 - 2\boldsymbol{\Sigma}_U\|_F^2$ and propose to solve a regularized M-estimation problem that

$$\big(\widehat{\boldsymbol{\Sigma}}_1, \widehat{\boldsymbol{\Sigma}}_2, \widehat{\boldsymbol{\Sigma}}_U\big) = \underset{(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_U)}{\operatorname{argmin}} \Big\{ \ell(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_U) + \nu_1 \|\boldsymbol{\Sigma}_1\|_* + \nu_2 \|\boldsymbol{\Sigma}_2\|_* + \nu_3 \|\boldsymbol{\Sigma}_U\|_1 \Big\}, \tag{4.4}$$

where ν_1, ν_2 and ν_3 are all non-negative tuning parameters, whose optimal values can be chosen by cross-validation. $\|\boldsymbol{M}\|_F$, $\|\boldsymbol{M}\|_1$ and $\|\boldsymbol{M}\|_*$ represents the Frobenius, L_1- and nuclear norms of $\boldsymbol{M}$, which are defined as $\|\boldsymbol{M}\|_F = \sqrt{\sum_i \sum_j M_{ij}^2}$, $\|\boldsymbol{M}\|_1 = \sum_{i,j} |M_{i,j}|$, and $\|\boldsymbol{M}\|_* = \sum_k \lambda_k(\boldsymbol{M})$, where $\lambda_k(\boldsymbol{M})$ is the k-th largest eigenvalue of $\boldsymbol{M}$ for any real matrix $\boldsymbol{M} = (M_{ij}) \in \mathbb{R}^{n \times p}$. In (4.4), we use the nuclear norm penalty to regularize the ranks of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, and use the L_1-penalty to induce a sparse estimator of $\boldsymbol{\Sigma}_U$. The nuclear norm penalty has been shown to be useful to recover the low-rank structure (Candès and Recht, 2009; Candès and Tao, 2010; Mazumder, Hastie and Tibshirani, 2010). The L_1-penalty is a well-known penalty function to render a sparse solution (Tibshirani, 1996). We also remark that if there are G groups ($G > 2$), we can solve a similar problem as

$$\big(\widehat{\boldsymbol{\Sigma}}_g, \widehat{\boldsymbol{\Sigma}}_U\big) = \underset{(\boldsymbol{\Sigma}_g, \boldsymbol{\Sigma}_U)}{\operatorname{argmin}} \Big\{ \ell(\boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\Sigma}_G, \boldsymbol{\Sigma}_U) + \sum_{g=1}^{G} \nu_g \|\boldsymbol{\Sigma}_g\|_* + \nu_{G+1} \|\boldsymbol{\Sigma}_U\|_1 \Big\}. \tag{4.5}$$

Such a problem is essentially the same as (4.4).

Next, we discuss how to solve (4.4). Let $\boldsymbol{\Theta} = (\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_U)$ and write $\widehat{\boldsymbol{R}} = \widehat{\boldsymbol{R}}_1 + \widehat{\boldsymbol{R}}_2$. We have $\ell(\boldsymbol{\Theta}) = (1/2)\mathrm{tr}\{(\widehat{\boldsymbol{R}} - \boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2 - 2\boldsymbol{\Sigma}_U)^T (\widehat{\boldsymbol{R}} - \boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2 - 2\boldsymbol{\Sigma}_U)\}$. We propose to iteratively fix two components in $\boldsymbol{\Theta}$ and solve for the other. To start the iterations, we need to obtain an initial estimator $\widehat{\boldsymbol{\Theta}}^{(0)} = (\widehat{\boldsymbol{\Sigma}}_1^{(0)}, \widehat{\boldsymbol{\Sigma}}_2^{(0)}, \widehat{\boldsymbol{\Sigma}}_U^{(0)})$.

First, we estimate r_g by $\widehat{r}_g = \text{argmax}_{j \leq \min\{(n_g,p)\}} \lambda_{j-1}(\widehat{\boldsymbol{R}}_g)/\lambda_j(\widehat{\boldsymbol{R}}_g)$, where $\lambda_j(\widehat{\boldsymbol{R}}_g)$ is the j-th largest eigenvalue of $\widehat{\boldsymbol{R}}_g$. Let $\widehat{\boldsymbol{V}}_g = (\widehat{\boldsymbol{v}}_g^1, \ldots, \widehat{\boldsymbol{v}}_g^{\widehat{r}_g})$, where $\widehat{\boldsymbol{v}}_g^j$ is the eigenvector corresponding to $\lambda_j(\widehat{\boldsymbol{R}}_g)$ and $\widehat{\boldsymbol{D}}_g = \text{diag}(\lambda_1(\widehat{\boldsymbol{R}}_g), \ldots, \lambda_{\widehat{r}_g}(\widehat{\boldsymbol{R}}_g))$. Then, we let $\widehat{\boldsymbol{\Sigma}}_g^{(0)} = \widehat{\boldsymbol{V}}_g \widehat{\boldsymbol{D}}_g \widehat{\boldsymbol{V}}_g^T$ for $g \in \{1, 2\}$ and $\widehat{\boldsymbol{\Sigma}}_U^{(0)} = (1/2)\big(\widehat{\boldsymbol{R}} - \widehat{\boldsymbol{\Sigma}}_1^{(0)} - \widehat{\boldsymbol{\Sigma}}_2^{(0)}\big)$. Denote the solution of $\boldsymbol{\Theta}$ at the h-th iteration as $\widehat{\boldsymbol{\Theta}}^{(h)} = (\widehat{\boldsymbol{\Sigma}}_1^{(h)}, \widehat{\boldsymbol{\Sigma}}_2^{(h)}, \widehat{\boldsymbol{\Sigma}}_U^{(h)})$. At the $(h+1)$-th iteration, we first fix $\widehat{\boldsymbol{\Sigma}}_2^{(h)}, \widehat{\boldsymbol{\Sigma}}_U^{(h)}$ and solve for $\widehat{\boldsymbol{\Sigma}}_1^{(h+1)}$. This becomes a spectral regularization problem. As shown in Mazumder, Hastie and Tibshirani (2010), this problem can be solved by a hard-thresholding Singular Value Decomposition (SVD). In particular, let $\widehat{\boldsymbol{R}} - \widehat{\boldsymbol{\Sigma}}_2^{(h)} - 2\widehat{\boldsymbol{\Sigma}}_U^{(h)} = \boldsymbol{U}_{\widehat{r}_1}^{(h)} \boldsymbol{D}_{\widehat{r}_1}^{(h)} \boldsymbol{V}_{\widehat{r}_1}^{(h)T}$ be the rank-$\widehat{r}_1$ SVD. We have $\widehat{\boldsymbol{\Sigma}}_1^{(h+1)} = \boldsymbol{U}_{\widehat{r}_1}^{(h)} \boldsymbol{S}_{\nu_1}(\boldsymbol{D}_{\widehat{r}_1}^{(h)}) \boldsymbol{V}_{\widehat{r}_1}^{(h)T}$, where $\boldsymbol{S}_{\nu_1}(\boldsymbol{D}_{\widehat{r}_1}^{(h)}) = \text{diag}[(\lambda_1^{(h)} - \nu_1)_+, \ldots, (\lambda_{\widehat{r}_1}^{(h)} - \nu_1)_+]$ and $\lambda_j^{(h)}$ is the j-th largest singular value of $\boldsymbol{D}_{\widehat{r}_1}^{(h)}$. Then, we check if $\widehat{\boldsymbol{\Sigma}}_1^{(h+1)}$ is positive definite. If not, we project it to the nearest positive definite matrix by solving

$$\underset{\lambda_{\min}(\boldsymbol{A})>0}{\text{argmin}} \ \|\widehat{\boldsymbol{\Sigma}}_1^{(h+1)} - \boldsymbol{A}\|_F, \tag{4.6}$$

where $\lambda_{\min}(\boldsymbol{A})$ is the smallest eigenvalue of $\boldsymbol{A}$. With a slight abuse of notation, we still denote the solution to (4.6) as $\widehat{\boldsymbol{\Sigma}}_1^{(h+1)}$. Then, we fix $(\widehat{\boldsymbol{\Sigma}}_1^{(h+1)}, \widehat{\boldsymbol{\Sigma}}_U^{(h)})$ and solve for $\widehat{\boldsymbol{\Sigma}}_2^{(h+1)}$, which can be done using the same hard-thresholding SVD. Lastly, we fix $(\widehat{\boldsymbol{\Sigma}}_1^{(h+1)}, \widehat{\boldsymbol{\Sigma}}_2^{(h+1)})$ and solve for $\widehat{\boldsymbol{\Sigma}}_U^{(h+1)}$. In this step, we use the proximal gradient descent algorithm (Parikh and Boyd, 2014) to solve the corresponding L_1-penalized problem. The solution is given by $\widehat{\boldsymbol{\Sigma}}_U^{(h+1)} = s(\widehat{\boldsymbol{\Sigma}}_U^{(h)} - d\nabla_{\boldsymbol{\Sigma}_U} \ell(\widehat{\boldsymbol{\Sigma}}_1^{(h+1)}, \widehat{\boldsymbol{\Sigma}}_2^{(h+1)}, \widehat{\boldsymbol{\Sigma}}_U^{(h)}), \nu_3 d)$, where d is the step size, $\nabla_{\boldsymbol{\Sigma}_U} \ell(\widehat{\boldsymbol{\Sigma}}_1^{(h+1)}, \widehat{\boldsymbol{\Sigma}}_2^{(h+1)}, \widehat{\boldsymbol{\Sigma}}_U^{(h)}) = 4\widehat{\boldsymbol{\Sigma}}_U^{(h)} - 2(\widehat{\boldsymbol{R}} - \widehat{\boldsymbol{\Sigma}}_1^{(h+1)} - \widehat{\boldsymbol{\Sigma}}_2^{(h+1)})$ and $s(\boldsymbol{x}, \pi)$ is the element-wise soft-thresholding operator, whose (i,j)-th element is defined as $s(\boldsymbol{x}, \pi)_{i,j} = \text{sign}(x_{i,j})(|x_{i,j}| - \pi)_+$. As for the choice of d, we follow Parikh & Boyd (2014, section 4.2) to perform a backtracking line search. That is, we iteratively decrease d until $\ell(\widehat{\boldsymbol{\Theta}}^{(h+1)}) \leq Q_d\big(\widehat{\boldsymbol{\Theta}}^{(h+1)}; \widehat{\boldsymbol{\Theta}}^{(h)}\big)$, where $Q_d\big(\widehat{\boldsymbol{\Theta}}^{(h+1)}; \widehat{\boldsymbol{\Theta}}^{(h)}\big) = \ell(\widehat{\boldsymbol{\Sigma}}_1^{(h+1)}, \widehat{\boldsymbol{\Sigma}}_2^{(h+1)}, \widehat{\boldsymbol{\Sigma}}_U^{(h)}) + \langle \widehat{\boldsymbol{\Sigma}}_U^{(h+1)} - \widehat{\boldsymbol{\Sigma}}_U^{(h)}, \nabla_{\boldsymbol{\Sigma}_U} \ell(\widehat{\boldsymbol{\Sigma}}_1^{(h+1)}, \widehat{\boldsymbol{\Sigma}}_2^{(h+1)}, \widehat{\boldsymbol{\Sigma}}_U^{(h)}) \rangle + \frac{1}{2d} ||\widehat{\boldsymbol{\Sigma}}_U^{(h+1)} - \widehat{\boldsymbol{\Sigma}}_U^{(h)}||_F^2$. We stop the iterations when the proportion of the maximal changes of $(\widehat{\boldsymbol{\Sigma}}_1, \widehat{\boldsymbol{\Sigma}}_2, \widehat{\boldsymbol{\Sigma}}_U)$ between two consecutive iterations is less than ζ, where $\zeta \in (0, 1)$ is a user-defined stopping threshold, set to be 0.1. A detailed algorithm for solving (4.4) is given here:

Algorithm1: The Proximal Gradient Algorithm for solving (4.4)

Input: $\boldsymbol{X}_1 \in \mathbb{R}^{n_1 \times p}, \boldsymbol{X}_2 \in \mathbb{R}^{n_2 \times p}$.

Output: $\widehat{\boldsymbol{\Sigma}}_1, \widehat{\boldsymbol{\Sigma}}_2$ and $\widehat{\boldsymbol{\Sigma}}_U$.

Initialization: Compute $\widehat{\boldsymbol{R}}_1$, $\widehat{\boldsymbol{R}}_2$ and let $\widehat{\boldsymbol{R}} = \widehat{\boldsymbol{R}}_1 + \widehat{\boldsymbol{R}}_2$.

For $g = 1, 2$, let $\widehat{r}_g = \text{argmax}_{j \leq \min\{(n_g, p)\}} \lambda_{j-1}(\widehat{\boldsymbol{R}}_g)/\lambda_j(\widehat{\boldsymbol{R}}_g)$ and $\widehat{\boldsymbol{\Sigma}}_g^{(0)} = \widehat{\boldsymbol{V}}_g \widehat{\boldsymbol{D}}_g \widehat{\boldsymbol{V}}_g^T$, where $\widehat{\boldsymbol{D}}_g = \text{diag}(\lambda_1(\widehat{\boldsymbol{R}}_g), \ldots, \lambda_{\widehat{r}_g}(\widehat{\boldsymbol{R}}_g))$, $\lambda_j(\widehat{\boldsymbol{R}}_g)$ is the j-th eigenvalue of $\widehat{\boldsymbol{R}}_g$, $\widehat{\boldsymbol{v}}_g^j$ is the corresponding eigenvector and $\widehat{\boldsymbol{V}}_g = (\widehat{\boldsymbol{v}}_g^1, \ldots, \widehat{\boldsymbol{v}}_g^{\widehat{r}_g})$.

Let $\widehat{\boldsymbol{\Sigma}}_U^{(0)} = \frac{1}{2}\big(\widehat{\boldsymbol{R}} - \widehat{\boldsymbol{\Sigma}}_1^{(0)} - \widehat{\boldsymbol{\Sigma}}_2^{(0)}\big)$. Set the step size d at $d = d^{(0)} \in \mathbb{R}^+$.

At the $(h+1)$th iteration, let $d = d^{(h)}$ and repeat the following steps.

Let $\widehat{\boldsymbol{\Sigma}}_1^{(h+1)} = \boldsymbol{U}_{\widehat{r}_1}^{(h)} \boldsymbol{S}_{\nu_1}(\boldsymbol{D}_{\widehat{r}_1}^{(h)}) \boldsymbol{V}_{\widehat{r}_1}^{(h)T}$, where $\widehat{\boldsymbol{R}} - \widehat{\boldsymbol{\Sigma}}_2^{(h)} - 2\widehat{\boldsymbol{\Sigma}}_U^{(h)} = \boldsymbol{U}_{\widehat{r}_1}^{(h)} \boldsymbol{D}_{\widehat{r}_1}^{(h)} \boldsymbol{V}_{\widehat{r}_1}^{(h)T}$ and $\boldsymbol{S}_{\nu_1}(\boldsymbol{D}_{\widehat{r}_1}^{(h)}) = \text{diag}[(\lambda_1(\boldsymbol{D}_{\widehat{r}_1}^{(h)}) - \nu_1)_+, \ldots, (\lambda_{\widehat{r}_1}(\boldsymbol{D}_{\widehat{r}_1}^{(h)}) - \nu_1)_+]$.

If $\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}_1^{(h+1)}) \geq 0$ then go to the next step, else let $\widehat{\boldsymbol{\Sigma}}_1^{(h+1)} = \boldsymbol{A}$, where $\boldsymbol{A}$ solves $\text{argmin}_{\lambda_{\min}(\boldsymbol{A})>0} \|\widehat{\boldsymbol{\Sigma}}_1^{(h+1)} - \boldsymbol{A}\|_F$.

Let $\widehat{\boldsymbol{\Sigma}}_2^{(h+1)} = \boldsymbol{U}_{\widehat{r}_2}^{(h)} \boldsymbol{S}_{\nu_2}(\boldsymbol{D}_{\widehat{r}_2}^{(h)}) \boldsymbol{V}_{\widehat{r}_2}^{(h)T}$, where $\widehat{\boldsymbol{R}} - \widehat{\boldsymbol{\Sigma}}_1^{(h+1)} - 2\widehat{\boldsymbol{\Sigma}}_U^{(h)} = \boldsymbol{U}_{\widehat{r}_2}^{(h)} \boldsymbol{D}_{\widehat{r}_2}^{(h)} \boldsymbol{V}_{\widehat{r}_2}^{(h)T}$ and $\boldsymbol{S}_{\nu_2}(\boldsymbol{D}_{\widehat{r}_2}^{(h)}) = \text{diag}[(\lambda_1(\boldsymbol{D}_{\widehat{r}_2}^{(h)}) - \nu_2)_+, \ldots, (\lambda_{\widehat{r}_2}(\boldsymbol{D}_{\widehat{r}_2}^{(h)}) - \nu_2)_+]$.

If $\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}_2^{(h+1)}) \geq 0$ then go to the next step, else let $\widehat{\boldsymbol{\Sigma}}_2^{(h+1)} = \boldsymbol{A}$, where $\boldsymbol{A}$ solves $\text{argmin}_{\lambda_{\min}(\boldsymbol{A})>0} \|\widehat{\boldsymbol{\Sigma}}_2^{(h+1)} - \boldsymbol{A}\|_F$.

Let

$$\widehat{\boldsymbol{\Sigma}}_U^{(h+1)} = s(\widehat{\boldsymbol{\Sigma}}_U^{(h)} - d\nabla_{\boldsymbol{\Sigma}_U} \ell(\widehat{\boldsymbol{\Sigma}}_1^{(h+1)}, \widehat{\boldsymbol{\Sigma}}_2^{(h+1)}, \widehat{\boldsymbol{\Sigma}}_U^{(h)}), \nu_3 d), \tag{4.7}$$

where $s(\boldsymbol{x}, \pi)_{i,j} = \text{sign}(x_{i,j})(|x_{i,j}| - \pi)_+$.

If $\ell(\widehat{\boldsymbol{\Theta}}^{(h+1)}) \leq Q_d\{\widehat{\boldsymbol{\Theta}}^{(h+1)}; \widehat{\boldsymbol{\Theta}}^{(h)}\}$, proceed to the next iteration,

else let $d = 0.8d$ and return to equation (4.7).

Iterate until $\max\left\{ \frac{\|\widehat{\boldsymbol{\Sigma}}_1^{(h+1)}\|_F - \|\widehat{\boldsymbol{\Sigma}}_1^{(h)}\|_F}{\|\widehat{\boldsymbol{\Sigma}}_1^{(h)}\|_F}, \frac{\|\widehat{\boldsymbol{\Sigma}}_2^{(h+1)}\|_F - \|\widehat{\boldsymbol{\Sigma}}_2^{(h)}\|_F}{\|\widehat{\boldsymbol{\Sigma}}_2^{(h)}\|_F}, \frac{\|\widehat{\boldsymbol{\Sigma}}_U^{(h+1)}\|_F - \|\widehat{\boldsymbol{\Sigma}}_U^{(h)}\|_F}{\|\widehat{\boldsymbol{\Sigma}}_U^{(h)}\|_F} \right\} \leq \zeta$.

4.2.2 Rank-based Latent Correlation Matrix Estimator

In this section, we propose a rank-based estimator of $\boldsymbol{R}_g$ for mixed data to be used in (4.4). Since such an estimator is separately calculated for each group, for the sake of simplicity, we omit the notation g in the subscript. We denote $\Phi_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as the cumulative distribution function (c.d.f.) of the p-dimensional multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. In particular, we write $\Phi_2(\mu_1, \mu_2; \sigma_{12})$

as the c.d.f. of a two-dimensional normal distribution with mean $(\mu_1, \mu_2)^T$ and σ_{12} is the covariance between the two variables.

Here, we derive the formulae of bridge functions for the latent correlations between three-level categorical variables and continuous/binary/truncated variables, which is one of our paper's contributions. We also prove that all these bridge functions are monotone so that they are invertible.

Theorem 2 *The following results hold.*

(a) For $j \in \mathscr{G}$ and $k \in \mathscr{C}$,

$$F_{jk}(R_{jk}; \Delta_{j1}, \Delta_{j2}) = 2\Phi_2(\Delta_{j2}, 0; \frac{R_{jk}}{\sqrt{2}}) - 2\Phi_2(\Delta_{j2}, 0; -\frac{R_{jk}}{\sqrt{2}}) \\ - 2\Phi_3(\Delta_{j1}, \Delta_{j2}, 0; \boldsymbol{R}_{3d}) + 2\Phi_3(\Delta_{j2}, \Delta_{j1}, 0; \boldsymbol{R}_{3d}),$$

where $\Delta_{j1} = f_j(C_{j1}), \Delta_{j2} = f_j(C_{j2})$, and

$$\boldsymbol{R}_{3d} = \begin{bmatrix} 1 & 0 & -\frac{R_{jk}}{\sqrt{2}} \\ 0 & 1 & \frac{R_{jk}}{\sqrt{2}} \\ -\frac{R_{jk}}{\sqrt{2}} & \frac{R_{jk}}{\sqrt{2}} & 1 \end{bmatrix}.$$

(b) For $j \in \mathscr{G}$ and $k \in \mathscr{B}$,

$$F_{jk}(R_{jk}) = 2\Phi_2(\Delta_{j2}, \Delta_k; R_{jk}) - 2\Phi_1(\Delta_{j2})\Phi_1(\Delta_k) - 2\Phi_1(\Delta_{j1})\Phi_2(\Delta_{j2}, \Delta_k; R_{jk}) \\ + 2\Phi_1(\Delta_{j2})\Phi_2(\Delta_{j1}, \Delta_k; R_{jk}),$$

where $\Delta_{j1} = f_j(C_{j1}), \Delta_{j2} = f_j(C_{j2})$, and $\Delta_k = f_k(C_k)$.

(c) For $j \in \mathscr{G}$ and $k \in \mathscr{T}$,

$$
\begin{aligned}
&F_{jk}(R_{jk}; \Delta_{j1}, \Delta_{j2}, \Delta_k) \\
&= 2\big\{ -2\Phi_1(\Delta_k)\Phi_1(\Delta_{j2}) - \Phi_1(\Delta_{j2}) + 2\Phi_2(\Delta_{j2}, \Delta_k; R_{jk}) \\
&\quad - \Phi_1(\Delta_{j1})\Phi_1(\Delta_{j2}) + \Phi_2(0, \Delta_{j2}; -R_{jk}/\sqrt{2}) - 2\Phi_1(\Delta_{j1})\Phi_2(\Delta_{j2}, \Delta_k; R_{jk}) \\
&\quad - \Phi_1(\Delta_k)\Phi_2(\Delta_{j2}, \Delta_k; R_{jk}) + 2\Phi_3(0, \Delta_{j2}, \Delta_k; \boldsymbol{R}_{3e}) - 2\Phi_3(0, \Delta_{j2}, \Delta_k; \boldsymbol{R}_{3f}) \\
&\quad + 2\Phi_3(\Delta_{j2}, \Delta_{j1}, 0; \boldsymbol{R}_{3d}) + 2\Phi_4(0, \Delta_{j2}, \Delta_k, \Delta_k; \boldsymbol{R}_{4c}) \\
&\quad + 2\Phi_4(0, \Delta_{j2}, \Delta_{j1}, \Delta_k; \boldsymbol{R}_{4d}) + 2\Phi_4(0, \Delta_{j2}, \Delta_{j1}, \Delta_k; \boldsymbol{R}_{4e}) \\
&\quad + 2\Phi_4(0, \Delta_{j1}, \Delta_{j1}, \Delta_k; \boldsymbol{R}_{4e}) - 2\Phi_2(\Delta_{j1}, \Delta_k; R_{jk})\Phi_2(\Delta_{j2}, \Delta_k; R_{jk}) \\
&\quad - 2\Phi_5(0, \Delta_{j1}, \Delta_{j1}, \Delta_k, \Delta_k; \boldsymbol{R}_5) + 2\Phi_5(0, \Delta_{j2}, \Delta_{j1}, \Delta_k, \Delta_k; \boldsymbol{R}_5)\big\},
\end{aligned}
$$

where $\Delta_{j1} = f_j(C_{j1}), \Delta_{j2} = f_j(C_{j2}), \Delta_k = f_k(C_k)$,

$$
\boldsymbol{R}_{3e} = \begin{bmatrix} 1 & \frac{R_{jk}}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{R_{jk}}{\sqrt{2}} & 1 & 0 \\ -\frac{1}{\sqrt{2}} & 0 & 1 \end{bmatrix}, \quad \boldsymbol{R}_{3f} = \begin{bmatrix} 1 & -\frac{R_{jk}}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ -\frac{R_{jk}}{\sqrt{2}} & 1 & R_{jk} \\ -\frac{1}{\sqrt{2}} & R_{jk} & 1 \end{bmatrix},
$$

$$
\boldsymbol{R}_{4c} = \begin{bmatrix} 1 & -\frac{R_{jk}}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{R_{jk}}{\sqrt{2}} & 1 & R_{jk} & 0 \\ -\frac{1}{\sqrt{2}} & R_{jk} & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{R}_{4d} = \begin{bmatrix} 1 & \frac{R_{jk}}{\sqrt{2}} & -\frac{R_{jk}}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{R_{jk}}{\sqrt{2}} & 1 & 0 & 0 \\ -\frac{R_{jk}}{\sqrt{2}} & 0 & 1 & R_{jk} \\ -\frac{1}{\sqrt{2}} & 0 & R_{jk} & 1 \end{bmatrix},
$$

$$
\boldsymbol{R}_{4e} = \begin{bmatrix} 1 & \frac{R_{jk}}{\sqrt{2}} & -\frac{R_{jk}}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{R_{jk}}{\sqrt{2}} & 1 & 0 & R_{jk} \\ -\frac{R_{jk}}{\sqrt{2}} & 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & R_{jk} & 0 & 1 \end{bmatrix} \text{ and } \boldsymbol{R}_5 = \begin{bmatrix} 1 & -\frac{R_{jk}}{\sqrt{2}} & \frac{R_{jk}}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{R_{jk}}{\sqrt{2}} & 1 & 0 & R_{jk} & 0 \\ \frac{R_{jk}}{\sqrt{2}} & 0 & 1 & 0 & R_{jk} \\ -\frac{1}{\sqrt{2}} & R_{jk} & 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & R_{jk} & 0 & 1 \end{bmatrix}.
$$

(d) For $j \in \mathscr{G}$ and $k \in \mathscr{G}$,

$$\begin{aligned} F(R_{jk}) = {} & 2\Phi_2(\Delta_{j2}, \Delta_{k2}; R_{jk}) - 2\Phi_1(\Delta_{j2})\Phi_1(\Delta_{k2}) \\ & - 4\Phi_2(\Delta_{k2}, \Delta_{j2}; R_{jk})\Phi_1(\Delta_{j1}) + 4\Phi_2(\Delta_{j1}, \Delta_{k2}; R_{jk})\Phi_1(\Delta_{j2}) \\ & + 2\Phi_2(\Delta_{j2}, \Delta_{k2}; R_{jk})\Phi_2(\Delta_{j1}, \Delta_{k1}; R_{jk}) \\ & - 2\Phi_2(\Delta_{j2}, \Delta_{k1}; R_{jk})\Phi_2(\Delta_{j1}, \Delta_{k2}; R_{jk}), \end{aligned}$$

where $\Delta_{j1} = f_j(C_{j1}), \Delta_{j2} = f_j(C_{j2}), \Delta_{k1} = f_k(C_{k1})$, $\Delta_{k2} = f_k(C_{k2})$.

Proposition 2 *All bridge functions in Theorem 2 are strictly increasing functions of $R_{jk} \in (-1, 1)$ for any given constants $\Delta_j, \Delta_k, \Delta_{j1}, \Delta_{j2}, \Delta_{k1}$, and Δ_{k2}.*

For a three-level categorical variable, $\Delta_{j1} = f_j(C_{j1})$ and $\Delta_{j2} = f_j(C_{j2})$ are unknown in practice. We observe that

$$\begin{aligned} \mathbb{E}\{I(X_{ij} = 2)\} &= \mathbb{P}(X_{ij} = 2) = \mathbb{P}(f_j(C_{j2}) > \Delta_{j2}) = 1 - \Phi_1(\Delta_{j2}), \\ \mathbb{E}\{I(X_{ij} = 0)\} &= \mathbb{P}(X_{ij} = 0) = \mathbb{P}(f_j(C_{j1}) < \Delta_{j1}) = \Phi_1(\Delta_{j1}). \end{aligned}$$

Then, we can estimate Δ_{j1} and Δ_{j2} by the moment estimators $\widehat{\Delta}_{j1} = \Phi^{-1}(n_0/n)$ and $\widehat{\Delta}_{j2} = \Phi^{-1}(1-n_2/n)$, where $n_0 = \sum_{i=1}^n I(X_{ij} = 0), n_1 = \sum_{i=1}^n I(X_{ij} = 1), n_2 = \sum_{i=1}^n I(X_{ij} = 2)$, and $n_0 + n_1 + n_2 = n$.

Using all formulae given in the above, we can estimate each element of $\boldsymbol{R}$. However, the resulting estimator $\widehat{\boldsymbol{R}}$ is not guaranteed to be positive definite. In that case, we can use the same technique in (4.6) to further project it to the nearest positive definite matrix.

4.3 Simulation experiments

To evaluate the numerical performance of our proposed method, we compare it with the following methods: (i) separate decomposition of sample Pearson correlation matrix in each group, which solves the problem of

$$(\breve{\boldsymbol{\Sigma}}_g^\dagger, \breve{\boldsymbol{\Sigma}}_{U_g}^\dagger) = \underset{(\boldsymbol{\Sigma}_g, \boldsymbol{\Sigma}_{U_g})}{\operatorname{argmin}} \|\breve{\boldsymbol{R}}_g - \boldsymbol{\Sigma}_g - \boldsymbol{\Sigma}_{U_g}\|_F^2 + \nu_g\|\boldsymbol{\Sigma}_g\|_* + \nu_3\|\boldsymbol{\Sigma}_{U_g}\|_1, \text{ for } g \in \{1, 2\},$$

where $\breve{\boldsymbol{R}}_g$ is the sample Pearson correlation matrix of observed variables for the g-th group. (ii) joint decomposition based on sample Pearson correlation matrix, which solves

$$\left(\breve{\boldsymbol{\Sigma}}_1, \breve{\boldsymbol{\Sigma}}_2, \breve{\boldsymbol{\Sigma}}_U\right) = \underset{(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_U)}{\operatorname{argmin}} \frac{1}{2}\|\breve{\boldsymbol{R}}_1 + \breve{\boldsymbol{R}}_2 - \boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2 - 2\boldsymbol{\Sigma}_U\|_F^2 + \nu_1\|\boldsymbol{\Sigma}_1\|_* + \nu_2\|\boldsymbol{\Sigma}_2\|_* + \nu_3\|\boldsymbol{\Sigma}_U\|_1;$$

(iii) separate decomposition of the correlation matrix of latent variables in each group,

$$(\widehat{\boldsymbol{\Sigma}}_g^{\dagger}, \widehat{\boldsymbol{\Sigma}}_{U_g}^{\dagger}) = \underset{(\boldsymbol{\Sigma}_g, \boldsymbol{\Sigma}_{U_g})}{\operatorname{argmin}} \|\widehat{\boldsymbol{R}}_g - \boldsymbol{\Sigma}_g - \boldsymbol{\Sigma}_{U_g}\|_F^2 + \nu_g\|\boldsymbol{\Sigma}_g\|_* + \nu_3\|\boldsymbol{\Sigma}_{U_g}\|_1, \text{ for } g \in \{1, 2\}.$$

We carry out simulation studies under both low- and high-dimensional settings and consider three different correlation structures under each setting. In all cases, ranks of the low-rank matrices were set as $r_1 = 3$ and $r_2 = 2$. For the low-dimensional settings, we set $n_g = 100$ for $g \in \{1, 2\}$ and $p = 60$. We consider the following three scenarios.

Scenario 1: We set $\boldsymbol{\Sigma}_1 = \boldsymbol{Q}_1\boldsymbol{D}_1\boldsymbol{Q}_1^T$, where $\boldsymbol{D}_1 = \operatorname{diag}(7, 6.5, 6)$ and $\boldsymbol{Q}_1 \in \mathbb{R}^{p\times r_1}$ is an orthonormal matrix. To generate $\boldsymbol{Q}_1$, we start with $\boldsymbol{A}_1 = (\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3) \in \mathbb{R}^{p\times r_1}$, where the first 30 elements of $\boldsymbol{a}_1$ range from 0.02 to 0.6 with increments of 0.02 and the next 30 elements of $\boldsymbol{a}_1$ range from 0.71 to 1 with increments of 0.01, elements of $\boldsymbol{a}_2$ range from -1.16 to 1.2 with increments of 0.04, and elements of $\boldsymbol{a}_3$ range from 0.12 to 1.3 with increments of 0.02. We then apply the Gram-Schmidt normalization to $\boldsymbol{A}_1$ to obtain $\boldsymbol{Q}_1$. We set diagonal elements of $\boldsymbol{\Sigma}_U$ to be 1 minus the diagonal elements of $\boldsymbol{\Sigma}_1$, and its (i, j)-th off-diagonal element as $\sigma_{u,ij} = \sigma_{u,i}\sigma_{u,j}\rho^{|i-j|}$ if $|i - j| = 1$ and $\sigma_{u,ij} = 0$ otherwise, where $\rho = 0.5$ and $\sigma_{u,j}$ denotes the j-th diagonal element of $\boldsymbol{\Sigma}_U$. We set $\boldsymbol{\Sigma}_2 = \mathbf{w}_2\mathbf{w}_2^T$, where $\mathbf{w}_2 = (\mathbf{w}_{21}, \mathbf{w}_{22})$ with the (1:4, 31:35, 54:60)-th the elements of $\mathbf{w}_{21}$ being equal to 0.65 and the rest being zeros. Here, we use $i : j$ to denote a sequence of consecutive integers from i to j. The j-th element of $\mathbf{w}_{22}$ is set as $w_{22;j} = \sqrt{1 - \sigma_{u,j} - w_{21;j}^2}$, for $1 \le j \le 60$, where $w_{22;j}$ denotes the j-th element of w_{22}. Under this scenario, $\|\boldsymbol{\Sigma}_U\|_F \approx 44.89\%\|\boldsymbol{R}_1\|_F$ and $\|\boldsymbol{\Sigma}_U\|_F \approx 36.82\%\|\boldsymbol{R}_2\|_F$, meaning the common variation captures 44.89% and 36.82% total variation in these two groups.

Scenario 2: We set $\boldsymbol{D}_1 = \operatorname{diag}(3.5, 3.5, 3.5)$. Using the same $\boldsymbol{Q}_1$ as in Scenario 1, we let $\boldsymbol{\Sigma}_1 = \boldsymbol{Q}_1\boldsymbol{D}_1\boldsymbol{Q}_1^T$. We set the diagonal elements of $\boldsymbol{\Sigma}_U$ to be 1 minus the diagonal elements of $\boldsymbol{\Sigma}_1$, and its (i, j)-th off-diagonal element $\sigma_{u,ij} = \sigma_{u,i}\sigma_{u,j}\rho^{|i-j|}$ if $|i - j| \le 2$, $\sigma_{u,ij} = 0$ otherwise, and $\rho = 0.5$. We set $\boldsymbol{\Sigma}_2$ the same as in Scenario 1 with the (1:3, 31:35, 55:60)-th elements of $\mathbf{w}_{21}$ been chosen as 0.503 and the rest as 0. Under this

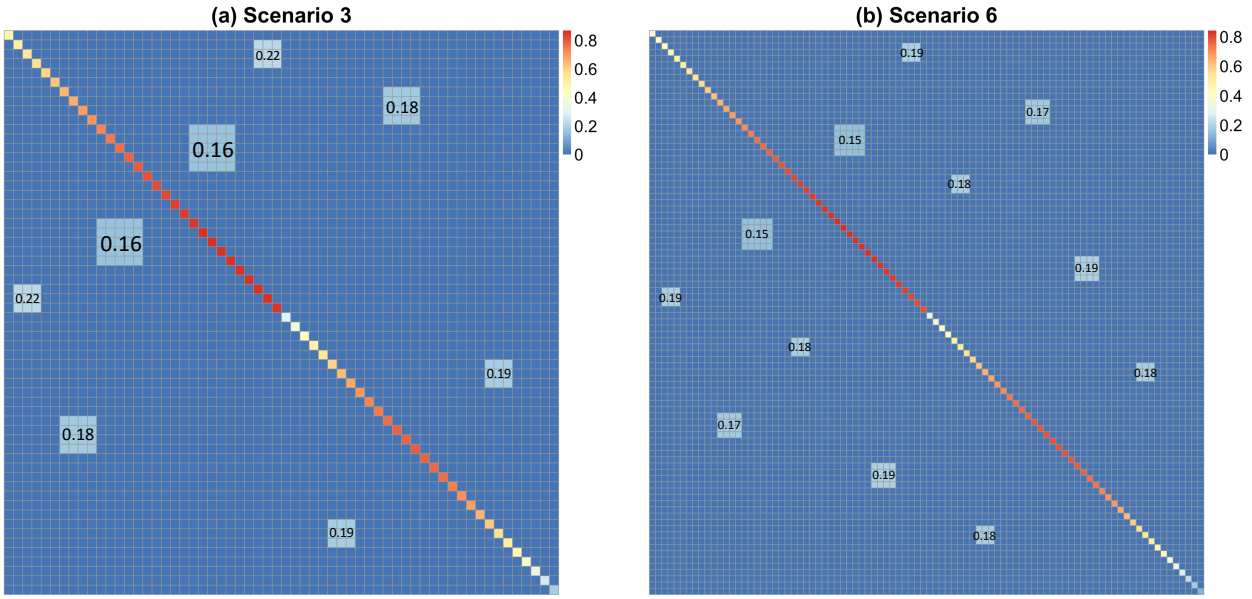


Figure 4.1: Heatmaps of $\boldsymbol{\Sigma}_U$ for Scenarios 3 and 6.

scenario, $\|\boldsymbol{\Sigma}_U\|_F \approx 68.41\%\|\boldsymbol{R}_1\|_F$ and $\|\boldsymbol{\Sigma}_U\|_F \approx 62.72\%\|\boldsymbol{R}_2\|_F$.

Scenario 3: We set $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ the same as in Scenario 1 with the (1:4, 31:35, 54:60)-th elements of $\mathbf{w}_{21}$ to be 0.65 and the rest to be 0. We set the diagonal elements of $\boldsymbol{\Sigma}_U$ to be 1 minus the diagonal elements of $\boldsymbol{\Sigma}_1$. The off-diagonal elements of $\boldsymbol{\Sigma}_U$ have a block-wise sparse structure; see Figure 4.1 for more details. In this scenario, $\|\boldsymbol{\Sigma}_U\|_F \approx 42.09\%\|\boldsymbol{R}_1\|_F$ and $\|\boldsymbol{\Sigma}_U\|_F \approx 33.31\%\|\boldsymbol{R}_2\|_F$.

The $\boldsymbol{\Sigma}_U$s in first two scenarios have similar banded structures, while in Scenario 2, the common variation, measured by $\|\boldsymbol{\Sigma}_U\|_F$, takes a larger proportion of the total variation, measured by $\|\boldsymbol{R}_g\|_F$. The third scenario has a different structure of $\boldsymbol{\Sigma}_U$, but the proportion of its common variation is comparable to that of Scenario 1. We use these settings to see how our method performs for different cases of $\boldsymbol{\Sigma}_U$ and $\boldsymbol{\Sigma}_g$.

For the high-dimensional settings, we set $n_g = 50$ for $g \in \{1, 2\}$, $p = 90$ and consider the following three scenarios.

Scenario 4: We set $\boldsymbol{\Sigma}_1 = \boldsymbol{Q}_1\boldsymbol{D}_1\boldsymbol{Q}_1^T$, where $\boldsymbol{D}_1 = \text{diag}(11, 10.5, 10)$ and $\boldsymbol{Q}_1 \in \mathbb{R}^{p\times r_1}$ is an orthonormal matrix. To generate $\boldsymbol{Q}_1$, we start with $\boldsymbol{A}_1 = (\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3)$, where the first 45 elements of $\boldsymbol{a}_1$ range from 0.02 to 0.9 with increments of 0.02 and the rest elements range from 1.01 to 1.45 with increments of 0.01, elements of $\boldsymbol{a}_2$ range from -1.76 to 1.8 with increments of 0.04, and elements of $\boldsymbol{a}_3$ range from 0.12 to

1.9 with increments of 0.02. We then apply Gram-Schmidt normalization to $\boldsymbol{A}_1$ to obtain $\boldsymbol{Q}_1$. We set diagonal elements of $\boldsymbol{\Sigma}_U$ to be 1 minus the diagonal elements of $\boldsymbol{\Sigma}_1$, and its (i,j)-th off-diagonal element as $\sigma_{u,ij} = \sigma_{u,i}\sigma_{u,j}\rho^{|i-j|}$ if $|i-j| = 1$ and $\sigma_{u,ij} = 0$ otherwise, where $\rho = 0.5$. We set $\boldsymbol{\Sigma}_2 = \mathbf{w}_2\mathbf{w}_2^T$, where $\mathbf{w}_2 = (\mathbf{w}_{21}, \mathbf{w}_{22})$ with the (1:8, 46:52, 81:90)-th elements of $\mathbf{w}_{21}$ being set as 0.65 and the rest as 0. The j-th element of $\mathbf{w}_{22}$ is set as $w_{22;j} = \sqrt{1 - \sigma_{u,j} - w_{21;j}^2}$. Under this scenario, $\|\boldsymbol{\Sigma}_U\|_F \approx 36.37\%\|\boldsymbol{R}_1\|_F$ and $\|\boldsymbol{\Sigma}_U\|_F \approx 28.83\%\|\boldsymbol{R}_2\|_F$.

Scenario 5: This scenario is the same as Scenario 2, except that we set $\boldsymbol{D}_1 = \mathrm{diag}(4.5, 4.5, 4.5)$ and use the same $\boldsymbol{Q}_1$ as in Scenario 4. We set $\boldsymbol{\Sigma}_2 = \mathbf{w}_2\mathbf{w}_2^T$, where $\mathbf{w}_2 = (\mathbf{w}_{21}, \mathbf{w}_{22})$, where the (1:6, 46:51, 82:90)-th elements of $\mathbf{w}_{21}$ are set to be 0.459 and the rest to be 0. Under this scenario, $\|\boldsymbol{\Sigma}_U\|_F \approx 69.39\%\|\boldsymbol{R}_1\|_F$ and $\|\boldsymbol{\Sigma}_U\|_F \approx 62.82\%\|\boldsymbol{R}_2\|_F$.

Scenario 6: We set $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ the same as in Scenario 4 with the (1:5, 46:50, 83:90)-th elements of $\mathbf{w}_{21}$ to be 0.72 and the rest to be 0. The off-diagonal elements of $\boldsymbol{\Sigma}_U$ have a block-wise sparse structure; see Figure 4.1 for more details. Under this scenario, $\|\boldsymbol{\Sigma}_U\|_F \approx 33.2\%\|\boldsymbol{R}_1\|_F$ and $\|\boldsymbol{\Sigma}_U\|_F \approx 25.47\%\|\boldsymbol{R}_2\|_F$.

Under each scenario, we first generate n_g i.i.d samples of $\boldsymbol{Z}_g$ from $N(\mathbf{0}, \boldsymbol{R}_g)$ for $g \in \{1,2\}$, and consider three LMGC models. In all these models, we set $\mathscr{C} = \{1, \ldots, p/3\}$, $\mathscr{B} = \{p/3+1, \ldots, 2p/3\}$, and $\mathscr{G} = \{2p/3+1, \ldots, p\}$.

- **Model 1:** For $g \in \{1,2\}$, $\boldsymbol{Y}_g = \boldsymbol{Z}_g$, $\boldsymbol{X}_g = \boldsymbol{h}_g(\boldsymbol{Y}_g)$, where $C_{1,j} = 0.3$, $C_{2,j} = 0.1$ for $j \in \mathscr{B}$, and $C_{1,j,1} = -0.7$, $C_{1,j,2} = 0.3$, $C_{2,j,1} = -0.5$, $C_{2,j,2} = 0.5$ for $j \in \mathscr{G}$.
- **Model 2:** $\boldsymbol{Y}_1 = \exp(\boldsymbol{Z}_1)$ and $\boldsymbol{Y}_2 = \boldsymbol{Z}_2$; $\boldsymbol{X}_g = \boldsymbol{h}_g(\boldsymbol{Y}_g)$, where $C_{1,j} = 1.5$, $C_{2,j} = 0.1$ for $j \in \mathscr{B}$, and $C_{1,j,1} = 0.6$, $C_{1,j,2} = 1.4$, $C_{2,j,1} = -0.5$, $C_{2,j,2} = 0.5$ for $j \in \mathscr{G}$.
- **Model 3:** $\boldsymbol{Y}_1 = \exp(\boldsymbol{Z}_1)$ and $\boldsymbol{Y}_2 = \boldsymbol{Z}_2^3$; $\boldsymbol{X}_g = \boldsymbol{h}_g(\boldsymbol{Y}_g)$, where $C_{1,j} = 1.5$, $C_{2,j} = 0.1$ for $j \in \mathscr{B}$, and $C_{1,j,1} = 0.6$, $C_{1,j,2} = 1.4$, $C_{2,j,1} = -0.5$, $C_{2,j,2} = 0.5$ for $j \in \mathscr{G}$.

We denote $\breve{\boldsymbol{R}}_g$ as the sample Pearson correlation coefficient of $\boldsymbol{X}_g$, where its (j,k)-th element is calculated as:

$\breve{R}_{g;(j,k)} = \sum_{i=1}^{n_g}(x_{g;i,j} - \bar{\boldsymbol{X}}_{g;j})(x_{g;i,k} - \bar{\boldsymbol{X}}_{g;k}) / \sqrt{\{\sum_{i=1}^{n_g}(x_{g;i,j} - \bar{\boldsymbol{X}}_{g;j})^2\}\{\sum_{i=1}^{n_g}(x_{g;i,k} - \bar{\boldsymbol{X}}_{g;k})^2\}}$, where $\bar{\boldsymbol{X}}_{g;j} = (1/n_g)\sum_{i=1}^{n_g} x_{g;i,j}$ for $j = 1, \ldots, p$.

Figure 4.2 gives the boxplots of $\|\widehat{\boldsymbol{R}}_g - \boldsymbol{R}_g\|_F$ and $\|\breve{\boldsymbol{R}}_g - \boldsymbol{R}_g\|_F$. It is seen that $\widehat{\boldsymbol{R}}_g$ performs much better than $\breve{\boldsymbol{R}}_g$ in all scenarios.

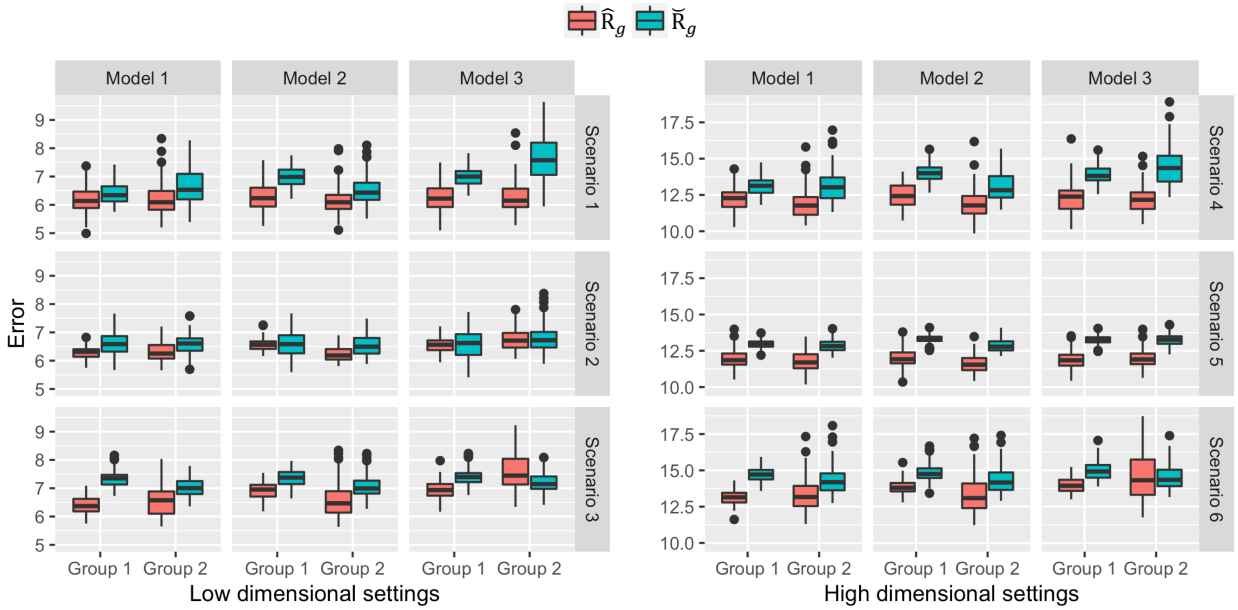


Figure 4.2: Estimation errors of Kendall's τ and Pearson correlation based estimators of $\boldsymbol{R}_g$.

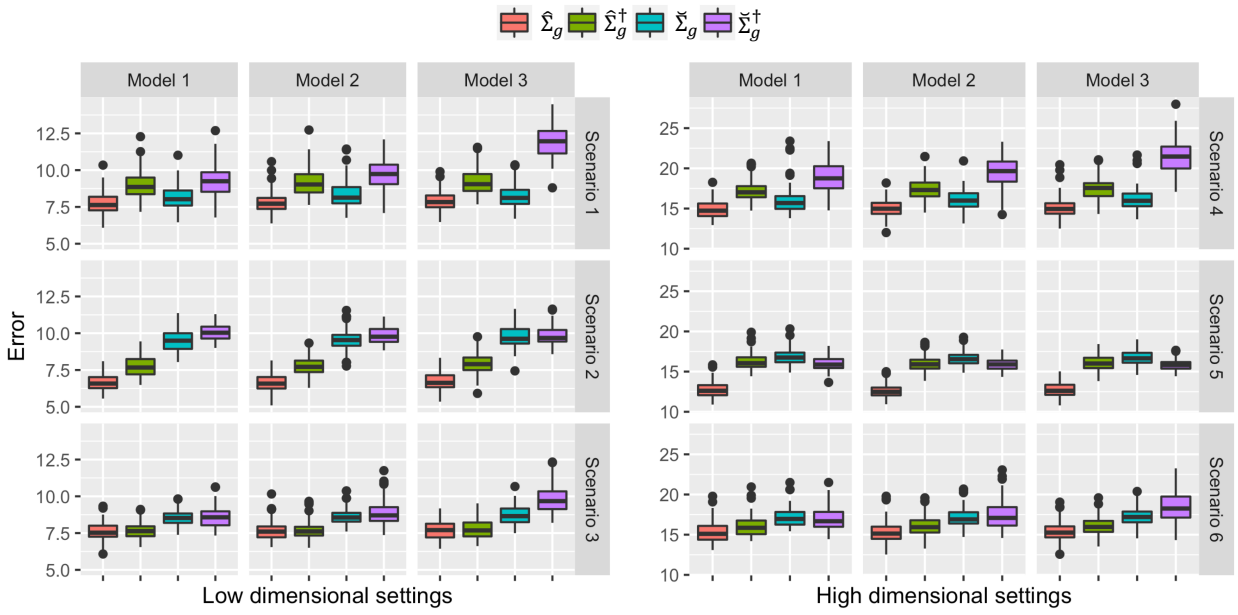
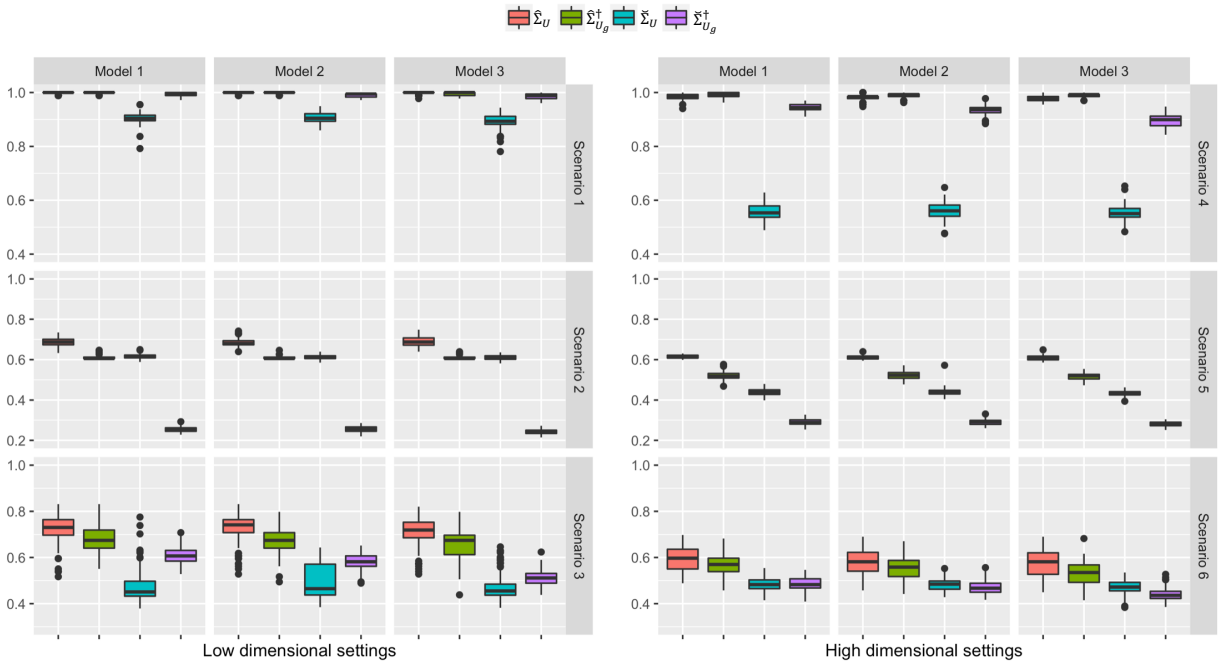
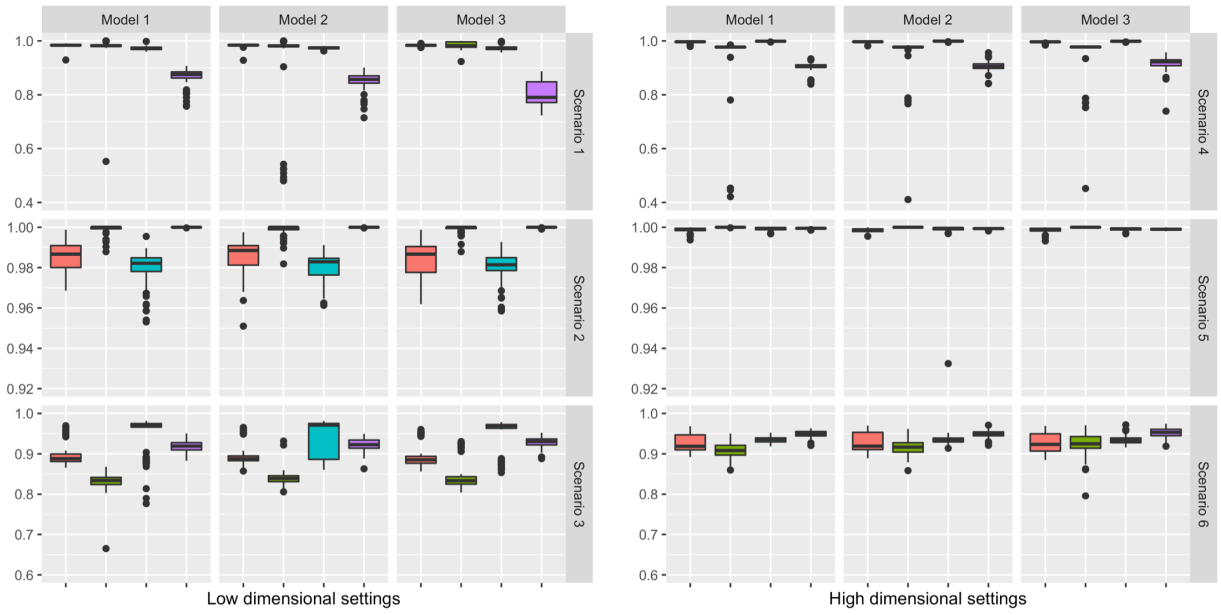


Figure 4.3: Estimation errors of $\boldsymbol{\Sigma}_g$ given by the four methods.



(a) Sensitivity



(b) Specificity

Figure 4.4: Variable selection accuracy of $\boldsymbol{\Sigma}_U$ given by the four methods.

Next, we compare the estimation error of the low-rank components by the four methods, which is measured by $\|\boldsymbol{A}_1 + \boldsymbol{A}_2 - \boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\|_F$, where $\boldsymbol{A}_g$ denotes one of $\check{\boldsymbol{\Sigma}}_g^{\dagger}$, $\check{\boldsymbol{\Sigma}}_g$, $\widehat{\boldsymbol{\Sigma}}_g^{\dagger}$ and $\widehat{\boldsymbol{\Sigma}}_g$ for $g \in \{1, 2\}$. It is seen from Figure 4.3 that our method performs the best in all scenarios .

Moreover, we compare the sensitivity and specificity of the four methods on recovering the nonzero elements of $\boldsymbol{\Sigma}_U$. Sensitivity is defined as the proportion of non-zero entries in $\boldsymbol{\Sigma}_U$ being estimated as non-zeros and specificity is defined as the proportion of zero entries in $\boldsymbol{\Sigma}_U$ being estimated as zeros. Figure 4.4 demonstrates the sensitivity and specificity of four competitors over 100 simulations. In Scenarios 1 and 4, $\check{\boldsymbol{\Sigma}}_{U_g}^{\dagger}, \check{\boldsymbol{\Sigma}}_U$, and $\widehat{\boldsymbol{\Sigma}}_U$ have high and comparable sensitivities, but $\widehat{\boldsymbol{\Sigma}}_U$'s specificity is higher than the other two. The sensitivity of $\widehat{\boldsymbol{\Sigma}}_{U_g}^{\dagger}$ is low in these two scenarios, suggesting that separately decompose the latent correlation matrices in two groups may not be capable of recovering the shared variation. For Scenarios 2 and 5, the sensitivity of all four methods reduces a lot. This is because their $\boldsymbol{\Sigma}_U$s have more complicated structures than the $\boldsymbol{\Sigma}_U$s in Scenarios 1 and 4. However, our estimator still has much higher sensitivity compared with the other methods . For Scenarios 3 and 6, $\boldsymbol{\Sigma}_U$ has blocks of small nonzero elements. Under these challenging settings, our method still outperforms the other three competitors. All these simulation studies suggest that our method can have good recovery of the group-specific low-rank and the shared sparse matrices for a variety of copula models.

4.4 *Chlamydia trachomatis* Genital Tract Infection Study

We applied our method to the multimodal data from the T cell Response Against Chlamydia (TRAC) cohort (Russell et al., 2016), which is designed for studying chlamydial genital tract infection. *Chlamydia trachomatis* can ascend from the cervix to the uterus and fallopian tubes in some women, and potentially result in pelvic inflammatory disease and infertility. Leveraging the TRAC cohort, we previously analyzed the association of 48 cytokines examined in cervical secretions with endometrial infection (Poston et al., 2019) and identified the cytokine regulatory network associated with chlamydial ascending infection by a graphical modelling approach (Zhong et al., 2020), but the genetic factors that drive the dysregulated cytokine network are still unclear.

To reveal the underlying genetic factors, we jointly analyzed the data on 48 cervical cytokines and genotype data from 128 women in TRAC, who had both cervical and endometrial infection (Endo+ group, n = 60) or had infection limited to the cervix (Endo- group, n = 68). Descriptions of the TRAC cohort, processing

and quality control of genotype and cervical cytokine expression data have been published in detail previously (Poston et al., 2019, Zhong et al., 2019). Directly genotyped single nucleotide polymorphisms (SNPs) were used in this study, while imputed genotypes were excluded. We treated the genotypes as ordinal variables with 3 levels. Cytokine levels were determined using Milliplex Magnetic Bead Assay. The cytokine values were log2 transformed, and treated as normally distributed continuous variables.

Expression quantitative trait loci (eQTLs) are the SNPs that influence expression levels of mRNA transcripts, which provide functional interpretation of the correlation between SNPs and cytokines. We thus primarily focused on SNPs that were cis-eQTLs of the cytokines, defined as SNPs within 1MB region flanking the gene that encodes the tested cytokine. eQTLs outside this region were defined as trans-eQTLs. We identified 300 SNP-cytokine cis-eQTL pairs, including 277 unique SNPs and 42 unique cytokines by Matrix eQTL (Shabalin, 2012) at significance level of 0.02. Next, we pruned the SNPs in high linkage disequilibrium with other SNPs in the list (squared correlation coefficient > 0.6) by PriorityPruner (v0.1.4, Edlund et al., 2016), and preferentially kept the most significant SNPs in the cis-eQTL detection. A total of 218 SNPs remained for further analysis. In each group, we further filtered SNPs whose latent correlation with another SNP is greater than the upper 0.1% quantile of the absolute value of the latent correlation matrix, while keeping the more significant SNPs in the cis-eQTL detection. Our final data set for each group had a total of 227 variables, including 42 cytokines and 185 SNP variables.

We apply our proposed method on this data set to obtain $\widehat{\boldsymbol{R}}_g$, $\widehat{\boldsymbol{\Sigma}}_g$ and $\widehat{\boldsymbol{\Sigma}}_U$. Figures 4.5 and 4.6 represent their heatmaps. Rows and columns of the heatmaps in Figure 4.5 were ordered by applying hierarchical clustering to the absolute value of $\widehat{\boldsymbol{\Sigma}}_1$, and those in Figure 4.6 were ordered by applying clustering result to the absolute value of $\widehat{\boldsymbol{\Sigma}}_2$.

We highlighted the cluster of variables that is most distinct from the other variables in $\widehat{\boldsymbol{\Sigma}}_1$ for the Endo- group (Figure 4.5B) and the same group of variables in $\widehat{\boldsymbol{\Sigma}}_2$ for Endo+ (Figure 4.5D) with green squares, namely Block $\boldsymbol{A}$, which consists 7 cytokines (CXCL13, EGF, IL17A, IL23A, CXCL10, CCL7, CCL23) and 40 SNPs. Among the 40 SNPs, 28 SNPs (70%) are cis-eQTLs of these 7 cytokines, and the remaining 12 SNPs are trans-eQTLs of these cytokines.

These 7 cytokines formed two sub-networks, one includes IL17A, IL23A, CXCL10, CXCL13, and their eQTLs. These 4 cytokines are associated with the aggregation of plasma cells and induction of Th17 cells, which are important immune cells involved in the host response to chlamydial genital tract infection (Andrew et al., 2013; Darville et al., 2019). IL17A is the signature cytokine of Th17 cells; IL-23 induces

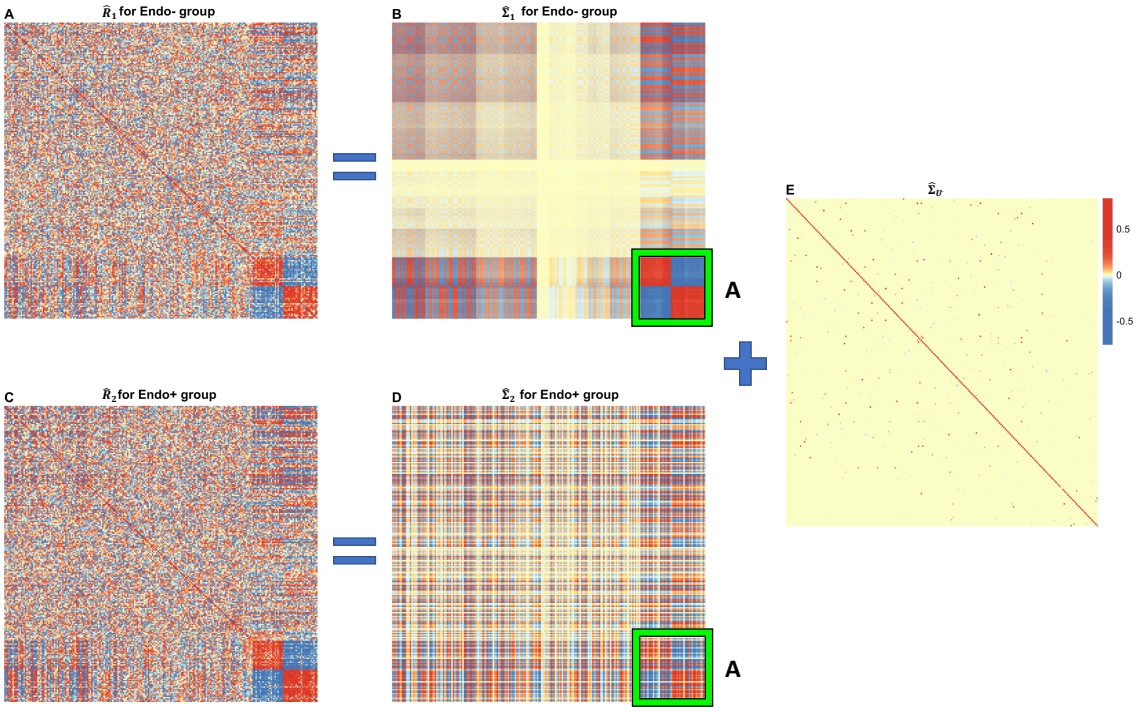


Figure 4.5: Heatmaps of $\widehat{\boldsymbol{R}}_g$, $\widehat{\boldsymbol{\Sigma}}_g$ and $\widehat{\boldsymbol{\Sigma}}_U$ for Endo- ($g = 1$) and Endo+ ($g = 2$) groups. Rows and columns of all heatmaps were ordered by applying clustering to the absolute value of $\widehat{\boldsymbol{\Sigma}}_1$. The cluster that is most distinct from all other clusters in $\widehat{\boldsymbol{\Sigma}}_1$ is highlighted in the green square. The same group of variables in $\widehat{\boldsymbol{\Sigma}}_2$ is also highlighted in the green square.

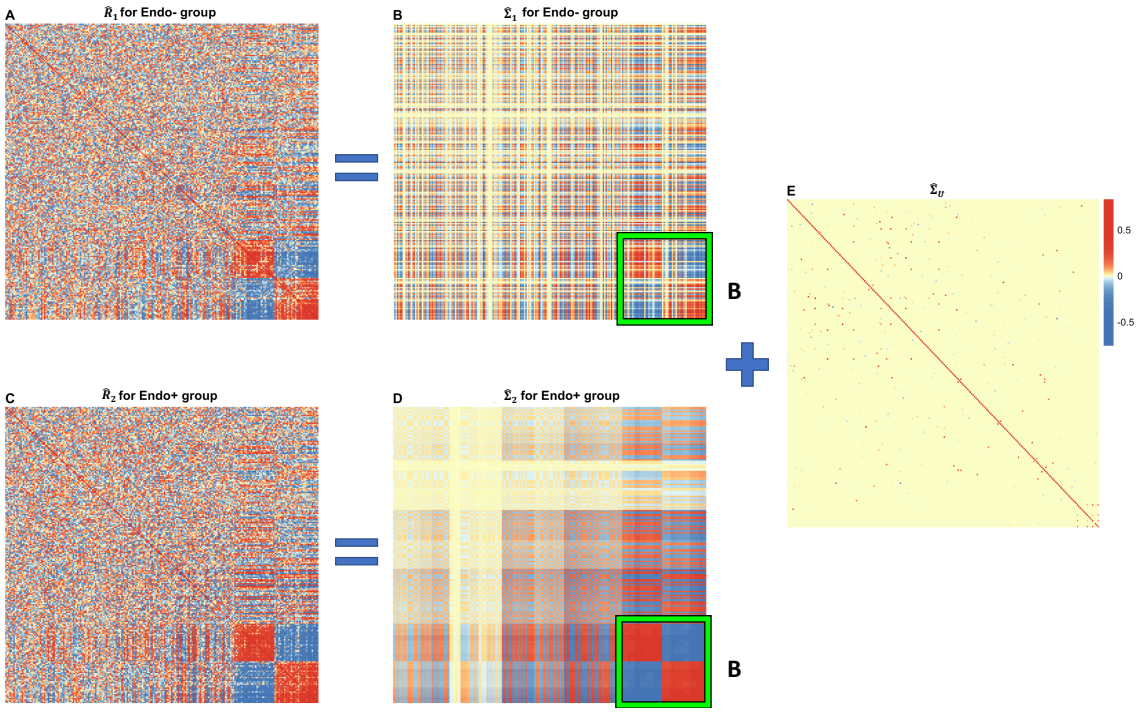


Figure 4.6: Heatmaps of $\widehat{\boldsymbol{R}}_g$, $\widehat{\boldsymbol{\Sigma}}_g$ and $\widehat{\boldsymbol{\Sigma}}_U$ for Endo- ($g = 1$) and Endo+ ($g = 2$) groups. Rows and columns of all heatmaps were ordered by applying clustering to the absolute value of $\widehat{\boldsymbol{\Sigma}}_2$. The cluster that is most distinct from all other clusters in $\widehat{\boldsymbol{\Sigma}}_2$ is highlighted in the green square. The same group of variables in $\widehat{\boldsymbol{\Sigma}}_1$ is also highlighted in the green square.

the differentiation of naive CD4+ T cells into Th17 cells (Iwakura, Ishigame et al., 2006); CXCL10 is a chemoattractant for CXCR3-positive Th17 cells and has also previously been correlated with detection of plasma cells in patients with inflammation and fibrosis (Nastase et al., 2018). CXCL13 levels are associated with plasma cell aggregates in tissues obtained from chlamydial induced endometrial inflammation (Kiviat et al., 1990). Additionally, the connectivity of CXCL13 and IL-17A has been evidenced experimentally (Rangel-Moreno et al., 2011).

We found enhanced eQTL effects and stronger correlation among these 4 cytokines in Endo- group, compared to Endo+ group. To validate the differential eQTL effects between Endo+ and Endo- group, we used a mediation test (GSMUT, Zhong et al., 2019) to examine whether any cis-eQTLs significantly differentially affect the outcome mediated through altering the cytokine expression. We found that the eQTL effects for rs4859453 and rs344108, i.e., the cis-eQTL of CXCL10 and CXCL13 respectively, only exist in Endo- group, but not in Endo+ group (GSMUT P value < 0.05, Figure 4.7). The diminished eQTL effects in Endo+ group were mostly mediated through increased expression of these two cytokines in Endo+ women who carried TT and/or AT genotype. Our method also revealed additional differential eQTLs, for example rs12941575, i.e., cis-eQTL of CCL7, which is not identified by mediation test (Figure 4.8). These findings suggest that our method can simultaneously identify the dysregulated cytokine networks and differential eQTLs. In addition, we can disclose the differential eQTLs with even mild to moderate effects, which might be overlooked by mediation test at single cytokine level.

The other sub-network in block $\boldsymbol{A}$ includes CCL7, CCL23, EGF and their eQTLs. These 3 cytokines are predominately associated with the recruitment of monocytes to sites of inflammation and regulation of host inflammatory responses. CCL23 and CCL7 are ligands for the chemokine receptor CCR1, which is critical for recruitment of monocytes. CCR1 is a target of the EGF signaling axis, which can induce and enhance CCR1 expression (Shin et al., 2017). In addition, CCL23 can mediate EGF receptor activation (Keates et al., 2007).

Next, we highlighted the cluster that is most distinct from all other clusters in the unique low rank part $\widehat{\boldsymbol{\Sigma}}_2$ for Endo+ group (Figure 4.6D) and the same group of variables in the unique low rank part $\widehat{\boldsymbol{\Sigma}}_1$ for Endo- (Figure 4.6B) with green squares, namely Block $\boldsymbol{B}$. Block $\boldsymbol{B}$ consists of 5 cytokines (CSF3, FLT3LG, TNFSF10, CCL5, CCL23,) and 56 SNPs.

All these 5 cytokines are involved in host immune and inflammatory responses to an infection. CSF3 and FLT3LG play synergistic roles in the physiological steady state for maintenance of neutrophil and dendritic

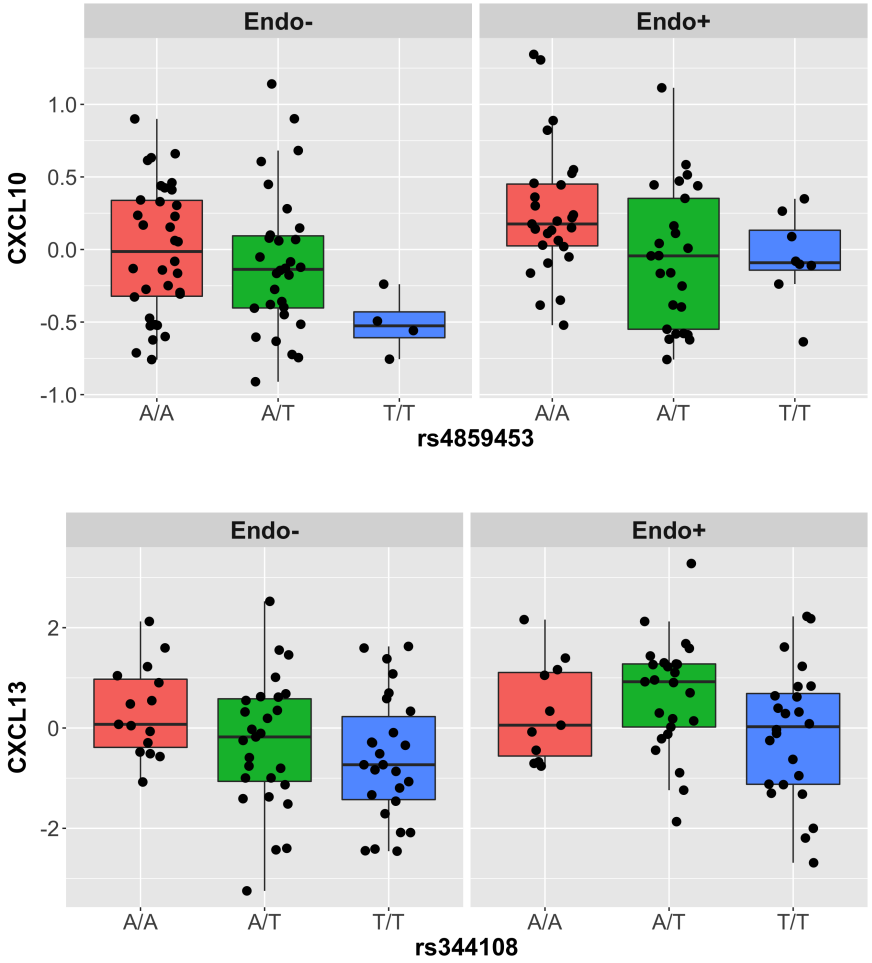


Figure 4.7: Expression of CXCL13 and CXCL10

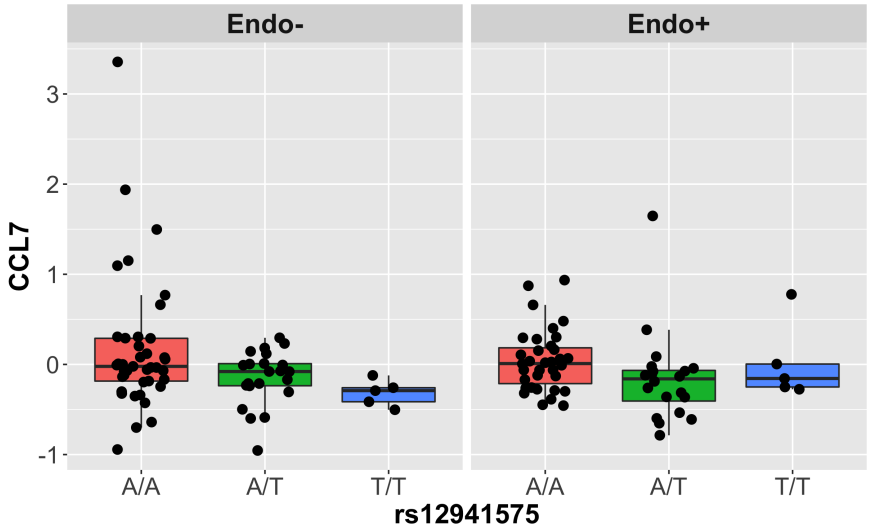


Figure 4.8: Expression of CCL7

Table 4.1: Estimate of all non-zero gene and gene elements of $\widehat{\mathbf{\Sigma}}_U$ in real data analysis

Gene1, gene2	Estimate
TNFSF13, CXCL5	-0.037
TNFSF13, CCL23	0.203
CXCL14, PDGFB	-0.004
CXCL9, CCL2	-0.026
EGF, CCL7	0.001
EGF, TNF	-0.025
CXCL5, IFNA2	-0.063
CCL11, CCL7	0.057
FGF2, IL15	-0.009
FGF2, PDGFA	-0.018
FGF2, PDGFB	-0.015
CX3CL1, CXCL1	-0.021
IFNA2, IL13	0.021
IL13, IL6	-0.080
IL13, PDGFB	0.076
IL16, PDGFA	-0.106
IL6, CCL23	0.026
IL6, PDGFB	-0.024
CCL3, CCL4	0.097

cell populations (Bhattacharya et al., 2015). TNFSF10 is critical in promoting infection-induced inflammation (Starkey et al., 2014), and experiments showed that G-CSF treatment increased the amount of TNFSF10 and the infiltration of neutrophils and mononuclear cells (Marino et al., 2009). CCL5 plays an important role in sustaining CD8 cytotoxic T cell responses and CCL23 is highly chemotactic for monocytes. It has been reported that neutrophils, monocytes and CD8 cytotoxic T cells contribute to chlamydial-induced upper genital tract inflammation (Lijek et al., 2018).

Finally, we demonstrated the shared cytokine and eQTL networks between Endo- and Endo+ groups, where the details are given in Tables 4.1 and 4.2. The cytokine networks among CXCL14, IL15, IL-16, PDGF-A, PDGF-B have been consistently identified by our previous graphic modeling algorithm and evidenced by biological function (Zhong et al., 2020). The preserved eQTL networks revealed important constitutional eQTLs despite different disease groups, such as rs11176892 for IFNG, which is a critical cytokine for controlling chlamydial infection.

Table 4.2: Estimate of all non-zero gene and SNP elements of $\widehat{\mathbf{\Sigma}}_U$ in real data analysis

Gene, SNP	Estimate
TNFSF13, rs4227	-0.006
CXCL14, rs2112186	-0.041
EGF, rs2081466	-0.060
CXCL5, rs10002688	0.035
CXCL5, rs13139174	0.075
CCL11, rs280045	0.088
CCL11, rs8070999	0.003
CX3CL1, rs9935360	0.003
CX3CL1, rs1466133	-0.005
IFNG, rs11176892	-0.030
IL12B, rs11952950	0.031
IL12B, rs7734683	0.043
IL15, rs12331218	-0.012
IL15, rs1425520	0.041
IL16, rs7178382	-0.001
IL16, rs6495518	0.005
IL16, rs4778906	-0.071
CCL22, rs11076198	-0.003
TGFA, rs1871241	-0.002

4.5 Discussion

We proposed a novel method to decompose the correlation of mixed types of variables from multiple groups by a Latent Mixed Gaussian Copula model. Our method can analyze binary, continuous, truncated and categorical variables simultaneously. By solving a penalized M estimation problem, our method can identify both the group-specific variation and the common variation. Various simulation experiments show that our method is more accurate in identifying the shared and group-specific structures, compared with the Pearson correlation based methods. The application of our method to the *Chlamydia trachomatis* genital tract infection study reveals differential and shared eQTL networks between the Endo- and Endo+ patients.

4.6 Technical Details

Proof of Proposition 1.

Since $(\boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{U})$ and $(\widetilde{\boldsymbol{F}}_1, \widetilde{\boldsymbol{F}}_2, \widetilde{\boldsymbol{U}})$ are both mutually uncorrelated, we have

$$\begin{aligned}
&\mathrm{cov}(\boldsymbol{\Lambda}_1\boldsymbol{F}_1, \boldsymbol{\Lambda}_2\boldsymbol{F}_2) = 0, \quad \mathrm{cov}(\boldsymbol{\Lambda}_1\boldsymbol{F}_1, \boldsymbol{U}) = 0, \quad \mathrm{cov}(\boldsymbol{\Lambda}_2\boldsymbol{F}_2, \boldsymbol{U}) = 0, \\
&\mathrm{cov}(\widetilde{\boldsymbol{\Lambda}}_1\widetilde{\boldsymbol{F}}_1, \widetilde{\boldsymbol{\Lambda}}_2\widetilde{\boldsymbol{F}}_2) = 0, \quad \mathrm{cov}(\widetilde{\boldsymbol{\Lambda}}_1\widetilde{\boldsymbol{F}}_1, \widetilde{\boldsymbol{U}}) = 0, \quad \mathrm{cov}(\widetilde{\boldsymbol{\Lambda}}_2\widetilde{\boldsymbol{F}}_2, \widetilde{\boldsymbol{U}}) = 0.
\end{aligned}$$

Suppose $\boldsymbol{\Lambda}_g\boldsymbol{F}_g + \boldsymbol{U} = \widetilde{\boldsymbol{\Lambda}}_g\widetilde{\boldsymbol{F}}_g + \widetilde{\boldsymbol{U}}$ for $g \in \{1, 2\}$. Let $\boldsymbol{W} = \widetilde{\boldsymbol{U}} - \boldsymbol{U}$. Then, $\widetilde{\boldsymbol{\Lambda}}_1\widetilde{\boldsymbol{F}}_1 = \boldsymbol{\Lambda}_1\boldsymbol{F}_1 - \boldsymbol{W}$ and $\widetilde{\boldsymbol{\Lambda}}_2\widetilde{\boldsymbol{F}}_2 = \boldsymbol{\Lambda}_2\boldsymbol{F}_2 - \boldsymbol{W}$. We have

$$\begin{aligned}
0 = \mathrm{cov}(\widetilde{\boldsymbol{\Lambda}}_1\widetilde{\boldsymbol{F}}_1, \widetilde{\boldsymbol{\Lambda}}_2\widetilde{\boldsymbol{F}}_2) &= \mathrm{cov}(\boldsymbol{\Lambda}_1\boldsymbol{F}_1 - \boldsymbol{W}, \boldsymbol{\Lambda}_2\boldsymbol{F}_2 - \boldsymbol{W}) \\
&= \mathrm{cov}(\boldsymbol{\Lambda}_1\boldsymbol{F}_1, \boldsymbol{\Lambda}_2\boldsymbol{F}_2) + \mathrm{Var}(\boldsymbol{W}) - \mathrm{cov}(\boldsymbol{\Lambda}_1\boldsymbol{F}_1, \boldsymbol{W}) - \mathrm{cov}(\boldsymbol{\Lambda}_2\boldsymbol{F}_2, \boldsymbol{W}),
\end{aligned}$$

which implies that

$$\mathrm{Var}(\boldsymbol{W}) = \mathrm{cov}(\boldsymbol{\Lambda}_1\boldsymbol{F}_1, \boldsymbol{W}) + \mathrm{cov}(\boldsymbol{\Lambda}_2\boldsymbol{F}_2, \boldsymbol{W}). \tag{4.8}$$

Similarly, we have

$$\begin{aligned}
0 = \mathrm{cov}(\widetilde{\boldsymbol{\Lambda}}_g\widetilde{\boldsymbol{F}}_g, \widetilde{\boldsymbol{U}}) &= \mathrm{cov}(\boldsymbol{\Lambda}_g\boldsymbol{F}_g - \boldsymbol{W}, \boldsymbol{U} + \boldsymbol{W}) \\
&= \mathrm{cov}(\boldsymbol{\Lambda}_g\boldsymbol{F}_g, \boldsymbol{U}) - \mathrm{Var}(\boldsymbol{W}) + \mathrm{cov}(\boldsymbol{\Lambda}_g\boldsymbol{F}_g, \boldsymbol{W}) - \mathrm{cov}(\boldsymbol{U}, \boldsymbol{W}),
\end{aligned}$$

which implies that

$$\mathrm{Var}(\boldsymbol{W}) = -\mathrm{cov}(\boldsymbol{U}, \boldsymbol{W}) + \mathrm{cov}(\boldsymbol{\Lambda}_g\boldsymbol{F}_g, \boldsymbol{W}). \tag{4.9}$$

By (4.9), we also have

$$\mathrm{cov}(\boldsymbol{\Lambda}_1\boldsymbol{F}_1, \boldsymbol{W}) = \mathrm{cov}(\boldsymbol{\Lambda}_2\boldsymbol{F}_2, \boldsymbol{W}). \tag{4.10}$$

By (4.8) and (4.10), we have $\mathrm{Var}(\boldsymbol{W}) = 2\mathrm{cov}(\boldsymbol{\Lambda}_1\boldsymbol{F}_1, \boldsymbol{W}) = 2\mathrm{cov}(\boldsymbol{\Lambda}_2\boldsymbol{F}_2, \boldsymbol{W})$.

By (4.8) and (4.9), we have $-\mathrm{cov}(\boldsymbol{U}, \boldsymbol{W}) = \mathrm{cov}(\boldsymbol{\Lambda}_1\boldsymbol{F}_1, \boldsymbol{W}) = \mathrm{cov}(\boldsymbol{\Lambda}_2\boldsymbol{F}_2, \boldsymbol{W})$.

Then, we have $\mathrm{Var}(\boldsymbol{W}) = -2\mathrm{cov}(\boldsymbol{U}, \boldsymbol{W}) = 2\mathrm{cov}(\boldsymbol{\Lambda}_1\boldsymbol{F}_1, \boldsymbol{W}) = 2\mathrm{cov}(\boldsymbol{\Lambda}_2\boldsymbol{F}_2, \boldsymbol{W})$.

Therefore,

$$\widetilde{\boldsymbol{\Lambda}}_g\widetilde{\boldsymbol{\Lambda}}_g' = \mathrm{Var}(\widetilde{\boldsymbol{\Lambda}}_g\widetilde{\boldsymbol{F}}_g) = \mathrm{Var}(\boldsymbol{\Lambda}_g\boldsymbol{F}_g - \boldsymbol{W}) = \mathrm{Var}(\boldsymbol{\Lambda}_g\boldsymbol{F}_g) + \mathrm{Var}(\boldsymbol{W}) - 2\mathrm{cov}(\boldsymbol{\Lambda}_g\boldsymbol{F}_g, \boldsymbol{W})$$
$$= \mathrm{Var}(\boldsymbol{\Lambda}_g\boldsymbol{F}_g) = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g'.$$

Thus, $\mathrm{Var}(\widetilde{\boldsymbol{U}}) = \mathrm{Var}(\boldsymbol{U} + \boldsymbol{W}) = \mathrm{Var}(\boldsymbol{U}) + \mathrm{Var}(\boldsymbol{W}) + 2\mathrm{cov}(\boldsymbol{U}, \boldsymbol{W}) = \mathrm{Var}(\boldsymbol{U})$.

Proof of Theorem 2.

(a) By the definition,

$$\tau_{jk} = \mathbb{E}(\widehat{\tau}_{jk}) = \mathbb{E}\Big[\frac{2}{n(n-1)} \sum_{1\le i<i'\le n} \mathrm{sign}\left\{(X_{ij} - X_{i'j})(X_{ik} - X_{i'k})\right\}\Big].$$

For the simplicity of notation, we omit i and i' from the subscripts and write X_{ij} and $X_{i'j}$ as X_j and X_j', where we treat them as two independent realizations from the same distribution.

Since X_j and X_j' are three-level categorical variables,

$$\begin{aligned}
\mathrm{sign}(X_j - X_j') = {} & I(X_j = 2, X_j' = 0) + I(X_j = 2, X_j' = 1) + I(X_j = 1, X_j' = 0) \\
& - I(X_j = 1, X_j' = 2) - I(X_j = 0, X_j' = 1) - I(X_j = 0, X_j' = 2) \\
= {} & I(X_j = 2) - I(X_j = 2, X_j' = 2) \\
& + I(X_j = 1, X_j' = 0) - I(X_j = 0, X_j' = 1) - I(X_j' = 2) + I(X_j = 2, X_j' = 2) \\
= {} & I(X_j = 2) - I(X_j' = 2) + I(X_j = 1, X_j' = 0) - I(X_j = 0, X_j' = 1)
\end{aligned}$$

Define $\boldsymbol{Z} = \boldsymbol{f}(\boldsymbol{Y})$, where $\boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$. Since $\mathrm{sign}(x) = 2I(x > 0) - 1$, we have

$$\begin{aligned}
\tau_{jk} &= \mathbb{E}\Big[\,\mathrm{sign}(X_j - X_j')\,\mathrm{sign}(X_k - X_k')\Big] = \mathbb{E}\Big[\,\mathrm{sign}(X_j - X_j')\{2I(X_k > X_k') - 1\}\Big] \\
&= \mathbb{E}\Big[2\,\mathrm{sign}(X_j - X_j')I(X_k > X_k') - \mathrm{sign}(X_j - X_j')\Big] \\
&= \mathbb{E}\Big[2I(X_j = 2)I(X_k > X_k') - 2I(X_j' = 2)I(X_k > X_k') \\
&\qquad + 2I(X_j = 1, X_j' = 0)I(X_k > X_k') - 2I(X_j = 0, X_j' = 1)I(X_k > X_k')
\end{aligned}$$

$$
\begin{aligned}
&\quad - I(X_j = 2) + I(X'_j = 2) - I(X_j = 1, X'_j = 0) + I(X_j = 0, X'_j = 1)\Big] \\
&= \mathbb{E}\Big[2I(Z_j > \Delta_{j2}, Z_k - Z'_k > 0) - 2I(Z'_j > \Delta_{j2}, Z_k - Z'_k > 0) \\
&\quad + 2I(\Delta_{j1} < Z_j < \Delta_{j2}, Z'_j < \Delta_{j1}, Z_k - Z'_k > 0) \\
&\quad - 2I(Z_j < \Delta_{j1}, \Delta_{j1} < Z'_j < \Delta_{j2}, Z_k - Z'_k > 0) \\
&\quad - I(Z_j > \Delta_{j2}) + I(Z'_j > \Delta_{j2}) \\
&\quad - I(\Delta_{j1} < Z_j < \Delta_{j2}, Z'_j < \Delta_{j1}) + I(Z_j < \Delta_{j1}, \Delta_{j1} < Z'_j < \Delta_{j2})\Big] \\
&= 2\mathbb{P}(Z_j > \Delta_{j2}, Z_k - Z'_k > 0) - 2\mathbb{P}(Z'_j > \Delta_{j2}, Z_k - Z'_k > 0) \\
&\quad + 2\mathbb{P}(\Delta_{j1} < Z_j < \Delta_{j2}, Z'_j < \Delta_{j1}, Z_k - Z'_k > 0) \\
&\quad - 2\mathbb{P}(Z_j < \Delta_{j1}, \Delta_{j1} < Z'_j < \Delta_{j2}, Z_k - Z'_k > 0) \\
&\quad - \mathbb{P}(Z_j > \Delta_{j2}) + \mathbb{P}(Z'_j > \Delta_{j2}) \\
&\quad - \mathbb{P}(\Delta_{j1} < Z_j < \Delta_{j2}, Z'_j < \Delta_{j1}) + \mathbb{P}(Z_j < \Delta_{j1}, \Delta_{j1} < Z'_j < \Delta_{j2}) \\
&= 2\mathbb{P}(Z'_k - Z_k < 0) - 2\mathbb{P}(Z_j < \Delta_{j2}, Z'_k - Z_k < 0) \\
&\quad - 2\mathbb{P}(Z'_k - Z_k < 0) + 2\mathbb{P}(Z'_j < \Delta_{j2}, Z'_k - Z_k < 0) \\
&\quad + 2\mathbb{P}(Z_j < \Delta_{j2}, Z'_j < \Delta_{j1}, Z'_k - Z_k < 0) - 2\mathbb{P}(Z_j < \Delta_{j1}, Z'_j < \Delta_{j1}, Z'_k - Z_k < 0) \\
&\quad - 2\mathbb{P}(Z_j < \Delta_{j1}, Z'_j < \Delta_{j2}, Z'_k - Z_k < 0) + 2\mathbb{P}(Z_j < \Delta_{j1}, Z'_j < \Delta_{j1}, Z'_k - Z_k < 0) \\
&= -2\mathbb{P}(Z_j < \Delta_{j2}, Z'_k - Z_k < 0) + 2\mathbb{P}(Z'_j < \Delta_{j2}, Z'_k - Z_k < 0) \\
&\quad + 2\mathbb{P}(Z_j < \Delta_{j2}, Z'_j < \Delta_{j1}, Z'_k - Z_k < 0) - 2\mathbb{P}(Z_j < \Delta_{j1}, Z'_j < \Delta_{j2}, Z'_k - Z_k < 0) \\
&= 2\Phi_2(\Delta_{j2}, 0; \frac{R_{jk}}{\sqrt{2}}) - 2\Phi_2(\Delta_{j2}, 0; -\frac{R_{jk}}{\sqrt{2}}) - 2\Phi_3(\Delta_{j1}, \Delta_{j2}, 0; \boldsymbol{R}_{3d}) + 2\Phi_3(\Delta_{j2}, \Delta_{j1}, 0; \boldsymbol{R}_{3d}).
\end{aligned}
$$

(b) Since X_k and X'_k are binary variables,

$$
\operatorname{sign}(X_k - X'_k) = I(Y_k > C_k) - I(Y'_k > C_k).
$$

We have

$$
\begin{aligned}
\tau_{jk} &= \mathbb{E}\Big[\operatorname{sign}(X_j - X_j')\big\{I(Y_k > C_k) - I(Y_k' > C_k)\big\}\Big] \\
&= \mathbb{E}\Big[I(X_j = 2, Y_k > C_k) - I(X_j' = 2, Y_k > C_k) + I(X_j = 1, X_j' = 0, Y_k > C_k) \\
&\quad - I(X_j = 0, X_j' = 1, Y_k > C_k) - I(X_j = 2, Y_k' > C_k) + I(X_j' = 2, Y_k' > C_k) \\
&\quad - I(X_j = 1, X_j' = 0, Y_k' > C_k) + I(X_j = 0, X_j' = 1, Y_k' > C_k)\Big] \\
&= \mathbb{E}\Big[I(Y_j > C_{j2}, Y_k > C_k) - I(Y_j' > C_{j2}, Y_k > C_k) \\
&+ I(C_{j1} < Y_j < C_{j2}, Y_j' < C_{j1}, Y_k > C_k) \\
&\quad - I(Y_j < C_{j1}, C_{j1} < Y_j' < C_{j2}, Y_k > C_k) - I(Y_j > C_{j2}, Y_k' > C_k) \\
&+ I(Y_j' > C_{j2}, Y_k' > C_k) \\
&\quad - I(C_{j1} < Y_j < C_{j2}, Y_j' < C_{j1}, Y_k' > C_k) \\
&+ I(Y_j < C_{j1}, C_{j1} < Y_j' < C_{j2}, Y_k' > C_k)\Big] \\
&= \mathbb{P}(Z_j > \Delta_{j2}, Z_k > \Delta_k) - \mathbb{P}(Z_j' > \Delta_{j2}, Z_k > \Delta_k) \\
&+ \mathbb{P}(\Delta_{j1} < Z_j < \Delta_{j2}, Z_j' < \Delta_{j1}, Z_k > \Delta_k) \\
&\quad - \mathbb{P}(Z_j < \Delta_{j1}, \Delta_{j1} < Z_j' < \Delta_{j2}, Z_k > \Delta_k) - \mathbb{P}(Z_j > \Delta_{j2}, Z_k' > \Delta_k) \\
&+ \mathbb{P}(Z_j' > \Delta_{j2}, Z_k' > \Delta_k) - \mathbb{P}(\Delta_{j1} < Z_j < \Delta_{j2}, Z_j' < \Delta_{j1}, Z_k' > \Delta_k) \\
&+ \mathbb{P}(Z_j < \Delta_{j1}, \Delta_{j1} < Z_j' < \Delta_{j2}, Z_k' > \Delta_k) \\
&= \mathbb{P}(Z_k > \Delta_k) - \mathbb{P}(Z_j < \Delta_{j2}, Z_k > \Delta_k) - \mathbb{P}(Z_k > \Delta_k) + \mathbb{P}(Z_j' < \Delta_{j2}, Z_k > \Delta_k) \\
&\quad + \mathbb{P}(Z_j < \Delta_{j2}, Z_j' < \Delta_{j1}, Z_k > \Delta_k) - \mathbb{P}(Z_j < \Delta_{j1}, Z_j' < \Delta_{j1}, Z_k > \Delta_k) \\
&\quad - \mathbb{P}(Z_j < \Delta_{j1}, Z_j' < \Delta_{j2}, Z_k > \Delta_k) + \mathbb{P}(Z_j < \Delta_{j1}, Z_j' < \Delta_{j1}, Z_k > \Delta_k) \\
&\quad - \mathbb{P}(Z_k' > \Delta_k) + \mathbb{P}(Z_j < \Delta_{j2}, Z_k' > \Delta_k) + \mathbb{P}(Z_k' > \Delta_k) - \mathbb{P}(Z_j' < \Delta_{j2}, Z_k' > \Delta_k) \\
&\quad - \mathbb{P}(Z_j < \Delta_{j2}, Z_j' < \Delta_{j1}, Z_k' > \Delta_k) + \mathbb{P}(Z_j < \Delta_{j1}, Z_j' < \Delta_{j1}, Z_k' > \Delta_k) \\
&\quad + \mathbb{P}(Z_j < \Delta_{j1}, Z_j' < \Delta_{j2}, Z_k' > \Delta_k) - \mathbb{P}(Z_j < \Delta_{j1}, Z_j' < \Delta_{j1}, Z_k' > \Delta_k) \\
&= -2\mathbb{P}(Z_j < \Delta_{j2}, Z_k > \Delta_k) + 2\mathbb{P}(Z_j' < \Delta_{j2}, Z_k > \Delta_k) \\
&\quad + 2\mathbb{P}(Z_j < \Delta_{j2}, Z_j' < \Delta_{j1}, Z_k > \Delta_k) - 2\mathbb{P}(Z_j < \Delta_{j1}, Z_j' < \Delta_{j1}, Z_k > \Delta_k) \\
&\quad - 2\mathbb{P}(Z_j < \Delta_{j1}, Z_j' < \Delta_{j2}, Z_k > \Delta_k) + 2\mathbb{P}(Z_j < \Delta_{j1}, Z_j' < \Delta_{j1}, Z_k > \Delta_k)
\end{aligned}
$$

$$
\begin{aligned}
&= -2\mathbb{P}(Z_j < \Delta_{j2}) + 2\mathbb{P}(Z_j < \Delta_{j2}, Z_k < \Delta_k) + 2\mathbb{P}(Z'_j < \Delta_{j2}) \\
&- 2\mathbb{P}(Z'_j < \Delta_{j2}, Z_k < \Delta_k) \\
&\quad + 2\mathbb{P}(Z_j < \Delta_{j2}, Z'_j < \Delta_{j1}) - 2\mathbb{P}(Z_j < \Delta_{j2}, Z'_j < \Delta_{j1}, Z_k < \Delta_k) \\
&\quad - 2\mathbb{P}(Z_j < \Delta_{j1}, Z'_j < \Delta_{j1}) + 2\mathbb{P}(Z_j < \Delta_{j1}, Z'_j < \Delta_{j1}, Z_k < \Delta_k) \\
&\quad - 2\mathbb{P}(Z_j < \Delta_{j1}, Z'_j < \Delta_{j2}) + 2\mathbb{P}(Z_j < \Delta_{j1}, Z'_j < \Delta_{j2}, Z_k < \Delta_k) \\
&\quad + 2\mathbb{P}(Z_j < \Delta_{j1}, Z'_j < \Delta_{j1}) - 2\mathbb{P}(Z_j < \Delta_{j1}, Z'_j < \Delta_{j1}, Z_k < \Delta_k) \\
&= 2\mathbb{P}(Z_j < \Delta_{j2}, Z_k < \Delta_k) - 2\mathbb{P}(Z'_j < \Delta_{j2}, Z_k < \Delta_k) \\
&\quad - 2\mathbb{P}(Z_j < \Delta_{j2}, Z'_j < \Delta_{j1}, Z_k < \Delta_k) + 2\mathbb{P}(Z_j < \Delta_{j1}, Z'_j < \Delta_{j2}, Z_k < \Delta_k) \\
&= 2\Big[\mathbb{P}(Z_j < \Delta_{j2}, Z_k < \Delta_k) - \mathbb{P}(Z'_j < \Delta_{j2})\mathbb{P}(Z_k < \Delta_k) \\
&\quad - \mathbb{P}(Z_j < \Delta_{j2}, Z'_j < \Delta_{j1}, Z_k < \Delta_k) + \mathbb{P}(Z_j < \Delta_{j1}, Z'_j < \Delta_{j2}, Z_k < \Delta_k)\Big] \\
&= 2\Phi_2(\Delta_{j2}, \Delta_k; R_{jk}) - 2\Phi_1(\Delta_{j2})\Phi_1(\Delta_k) - 2\Phi_1(\Delta_{j1})\Phi_2(\Delta_{j2}, \Delta_k; R_{jk}) \\
&\quad + 2\Phi_1(\Delta_{j2})\Phi_2(\Delta_{j1}, \Delta_k; R_{jk}.)
\end{aligned}
$$

(c) Since X_j and X'_j are three-level categorical variables,

$$\operatorname{sign}(X_j - X'_j) = I(X_j = 2) - I(X'_j = 2) + I(X_j = 1, X'_j = 0) - I(X_j = 0, X'_j = 1).$$

Since X_k and X'_k are truncated variables,

$$
\begin{aligned}
\operatorname{sign}(X_k - X'_k) &= -I(X_k = 0, X'_k > 0) + I(X_k > 0, X'_k = 0) \\
&\quad + I(X_k > 0, X'_k > 0)\operatorname{sign}(X_k - X'_k) \\
&= -I(X_k = 0) + I(X'_k = 0) + I(X_k > 0, X'_k > 0)\operatorname{sign}(X_k - X'_k).
\end{aligned}
$$

Since $\text{sign}(x) = 2I(x > 0) - 1$, we have

$$
\begin{aligned}
\tau_{jk} &= \mathbb{E}\Big[\,\text{sign}(X_j - X_j')\,\text{sign}(X_k - X_k')\Big] \\
&= \mathbb{E}\Big[2I(X_j = 2)I(X_k' = 0) - 2I(X_j = 2)I(X_k = 0) + 2I(X_j = 1, X_j' = 0)I(X_k' = 0) \\
&\quad - 2I(X_j = 1, X_j' = 0)I(X_k = 0) + 2I(X_k > X_k')I(X_j = 2)I(X_k > 0, X_k' > 0) \\
&\quad - 2I(X_k > X_k')I(X_j' = 2)I(X_k > 0, X_k' > 0) \\
&\quad + 2I(X_k > X_k')I(X_j = 1, X_j' = 0)I(X_k > 0, X_k' > 0) \\
&\quad - 2I(X_k > X_k')I(X_j = 0, X_j' = 1)I(X_k > 0, X_k' > 0)\Big] \\
&= \mathbb{E}\Big[2I(Z_j > \Delta_{j2})I(Z_k' < \Delta_k) - 2I(Z_j > \Delta_{j2})I(Z_k < \Delta_k) \\
&\quad + 2I(\Delta_{j1} < Z_j < \Delta_{j2}, Z_j' < \Delta_{j1})I(Z_k' < \Delta_k) \\
&\quad - 2I(\Delta_{j1} < Z_j < \Delta_{j2}, Z_j' < \Delta_{j1})I(Z_k < \Delta_k) \\
&\quad + 2I(Z_k' - Z_k < 0)I(Z_j > \Delta_{j2})I(Z_k > \Delta_k, Z_k' > \Delta_k) \\
&\quad - 2I(Z_k' - Z_k < 0)I(Z_j' > \Delta_{j2})I(Z_k > \Delta_k, Z_k' > \Delta_k) \\
&\quad + 2I(Z_k' - Z_k < 0)I(\Delta_{j1} < Z_j < \Delta_{j2}, Z_j' < \Delta_{j1})I(Z_k > \Delta_k, Z_k' > \Delta_k) \\
&\quad - 2I(Z_k' - Z_k < 0)I(\Delta_{j1} < Z_j' < \Delta_{j2}, Z_j < \Delta_{j1})I(Z_k > \Delta_k, Z_k' > \Delta_k)\Big] \\
&= 2\Big[\mathbb{P}(Z_j > \Delta_{j2}, Z_k' < \Delta_k) - \mathbb{P}(Z_j > \Delta_{j2}, Z_k < \Delta_k) \\
&\quad + \mathbb{P}(\Delta_{j1} < Z_j < \Delta_{j2}, Z_j' < \Delta_{j1}, Z_k' < \Delta_k) \\
&\quad - \mathbb{P}(\Delta_{j1} < Z_j < \Delta_{j2}, Z_j' < \Delta_{j1}, Z_k < \Delta_k) \\
&\quad + \mathbb{P}(Z_k' - Z_k < 0, Z_j > \Delta_{j2}, Z_k > \Delta_k, Z_k' > \Delta_k) \\
&\quad - \mathbb{P}(Z_k' - Z_k < 0, Z_j' > \Delta_{j2}, Z_k > \Delta_k, Z_k' > \Delta_k) \\
&\quad + \mathbb{P}(Z_k' - Z_k < 0, \Delta_{j1} < Z_j < \Delta_{j2}, Z_j' < \Delta_{j1}, Z_k > \Delta_k, Z_k' > \Delta_k) \\
&\quad - \mathbb{P}(Z_k' - Z_k < 0, \Delta_{j1} < Z_j' < \Delta_{j2}, Z_j < \Delta_{j1}, Z_k > \Delta_k, Z_k' > \Delta_k)\Big] \\
&= 2\Big[\mathbb{P}(Z_k' < \Delta_k)\mathbb{P}(Z_j > \Delta_{j2}) - \mathbb{P}(Z_k < \Delta_k) - \mathbb{P}(Z_j < \Delta_{j2}) \\
&\quad + 2\mathbb{P}(Z_j < \Delta_{j2}, Z_k < \Delta_k) - \mathbb{P}(Z_j < \Delta_{j2})\mathbb{P}(Z_j' < \Delta_{j1}) \\
&\quad - \mathbb{P}(Z_j' < \Delta_{j2})\mathbb{P}(Z_k < \Delta_k) + 2\mathbb{P}(Z_k' - Z_k < 0, Z_j < \Delta_{j2}) \\
&\quad - 2\mathbb{P}(Z_j' < \Delta_{j1})\mathbb{P}(Z_j < \Delta_{j2}, Z_k < \Delta_k) - \mathbb{P}(Z_k' < \Delta_k)\mathbb{P}(Z_j < \Delta_{j2}, Z_k < \Delta_k)
\end{aligned}
$$

$$
\begin{aligned}
&+ 2\mathbb{P}(Z'_k - Z_k < 0, Z'_j < \Delta_{j2}, Z_k < \Delta_k) - 2\mathbb{P}(Z'_k - Z_k < 0, Z_j < \Delta_{j2}, Z_k < \Delta_k) \\
&+ 2\mathbb{P}(Z'_k - Z_k < 0, Z_j < \Delta_{j2}, Z'_j < \Delta_{j1}) + 2\mathbb{P}(Z'_k - Z_k < 0, Z_j < \Delta_{j2}, Z_k < \Delta_k, Z'_k < \Delta_k) \\
&+ 2\mathbb{P}(Z'_k - Z_k < 0, Z'_j < \Delta_{j2}, Z_j < \Delta_{j1}, Z_k < \Delta_k) \\
&+ 2\mathbb{P}(Z'_k - Z_k < 0, Z'_j < \Delta_{j2}, Z_j < \Delta_{j1}, Z'_k < \Delta_k) \\
&+ 2\mathbb{P}(Z'_k - Z_k < 0, Z_j < \Delta_{j1}, Z'_j < \Delta_{j1}, Z'_k < \Delta_k) \\
&- 2\mathbb{P}(Z'_k - Z_k < 0, Z_j < \Delta_{j1}, Z'_j < \Delta_{j1}, Z_k < \Delta_k, Z'_k < \Delta_k) \\
&+ 2\mathbb{P}(Z'_k - Z_k < 0, Z_j < \Delta_{j2}, Z'_j < \Delta_{j1}, Z_k < \Delta_k, Z'_k < \Delta_k) \\
&- \mathbb{P}(Z_j < \Delta_{j2}, Z'_j < \Delta_{j1}, Z_k < \Delta_k, Z'_k < \Delta_k)\Big] \\
= 2\big\{ &- 2\Phi_1(\Delta_k)\Phi_1(\Delta_{j2}) - \Phi_1(\Delta_{j2}) + 2\Phi_2(\Delta_{j2}, \Delta_k; R_{jk}) - \Phi_1(\Delta_{j1})\Phi_1(\Delta_{j2}) \\
&+ \Phi_2(0, \Delta_{j2}; -\frac{R_{jk}}{\sqrt{2}}) - 2\Phi_1(\Delta_{j1})\Phi_2(\Delta_{j2}, \Delta_k; R_{jk}) - \Phi_1(\Delta_k)\Phi_2(\Delta_{j2}, \Delta_k; R_{jk}) \\
&+ 2\Phi_3(0, \Delta_{j2}, \Delta_k; \boldsymbol{R}_{3e}) - 2\Phi_3(0, \Delta_{j2}, \Delta_k; \boldsymbol{R}_{3f}) + 2\Phi_3(\Delta_{j2}, \Delta_{j1}, 0; \boldsymbol{R}_{3d}) \\
&+ 2\Phi_4(0, \Delta_{j2}, \Delta_k, \Delta_k; \boldsymbol{R}_{4c}) + 2\Phi_4(0, \Delta_{j2}, \Delta_{j1}, \Delta_k; \boldsymbol{R}_{4d}) + 2\Phi_4(0, \Delta_{j2}, \Delta_{j1}, \Delta_k; \boldsymbol{R}_{4e}) \\
&+ 2\Phi_4(0, \Delta_{j1}, \Delta_{j1}, \Delta_k; \boldsymbol{R}_{4e}) - 2\Phi_2(\Delta_{j1}, \Delta_k; R_{jk})\Phi_2(\Delta_{j2}, \Delta_k; R_{jk}) \\
&- 2\Phi_5(0, \Delta_{j1}, \Delta_{j1}, \Delta_k, \Delta_k; \boldsymbol{R}_5) + 2\Phi_5(0, \Delta_{j2}, \Delta_{j1}, \Delta_k, \Delta_k; \boldsymbol{R}_5)\big\}.
\end{aligned}
$$

(d) We have

$$
\begin{aligned}
\tau_{jk} &= \mathbb{E}\Big[\operatorname{sign}(X_j - X'_j)\operatorname{sign}(X_k - X'_k)\Big] \\
&= \mathbb{E}\Big[\{I(X_j = 2) - I(X'_j = 2) + I(X_j = 1, X'_j = 0) - I(X_j = 0, X'_j = 1)\} \\
&\qquad \{I(X_k = 2) - I(X'_k = 2) + I(X_k = 1, X'_k = 0) - I(X_k = 0, X'_k = 1)\}\Big] \\
&= \mathbb{E}\Big[I(X_j = 2, X_k = 2) - I(X_j = 2, X'_k = 2) \\
&\quad - I(X'_j = 2, X_k = 2) + I(X'_j = 2, X'_k = 2) \\
&\quad + I(X_j = 2, X_k = 1, X'_k = 0) - I(X_j = 2, X_k = 0, X'_k = 1) \\
&\quad + I(X'_j = 2, X_k = 0, X'_k = 1) - I(X'_j = 2, X_k = 1, X'_k = 0) \\
&\quad + I(X_j = 1, X'_j = 0, X_k = 2) - I(X_j = 1, X'_j = 0, X'_k = 2)
\end{aligned}
$$

$$
\begin{aligned}
&+ I(X_j = 0, X'_j = 1, X'_k = 2) - I(X_j = 0, X'_j = 1, X_k = 2) \\
&+ I(X_j = 1, X'_j = 0, X_k = 1, X'_k = 0) - I(X_j = 1, X'_j = 0, X_k = 0, X'_k = 1) \\
&+ I(X_j = 0, X'_j = 1, X_k = 0, X'_k = 1) - I(X_j = 0, X'_j = 1, X_k = 1, X'_k = 0)\Big] \\
= \mathbb{E}\Big[&2I(X_j = 2, X_k = 2) - 2I(X'_j = 2, X_k = 2) \\
&+ 2I(X_j = 2, X_k = 1, X'_k = 0) - 2I(X_j = 2, X_k = 0, X'_k = 1) \\
&+ 2I(X_j = 1, X'_j = 0, X_k = 2) - 2I(X_j = 1, X'_j = 0, X'_k = 2) \\
&+ 2I(X_j = 1, X'_j = 0, X_k = 1, X'_k = 0) - 2I(X_j = 1, X'_j = 0, X_k = 0, X'_k = 1)\Big] \\
= 2\mathbb{P}(&X_j = 2, X_k = 2) - 2\mathbb{P}(X'_j = 2, X_k = 2) \\
&+ 2\mathbb{P}(X_j = 2, X_k = 1, X'_k = 0) - 2\mathbb{P}(X_j = 2, X_k = 0, X'_k = 1) \\
&+ 2\mathbb{P}(X_j = 1, X'_j = 0, X_k = 2) - 2\mathbb{P}(X_j = 0, X'_j = 1, X_k = 2) \\
&+ 2\mathbb{P}(X_j = 1, X'_j = 0, X_k = 1, X'_k = 0) - 2\mathbb{P}(X_j = 1, X'_j = 0, X_k = 0, X'_k = 1) \\
= 2\mathbb{P}(&Z_j > \Delta_{j2}, Z_k > \Delta_{k2}) - 2\mathbb{P}(Z'_j > \Delta_{j2}, Z_k > \Delta_{k2}) \\
&+ 2\mathbb{P}(Z_j > \Delta_{j2}, \Delta_{k1} < Z_k < \Delta_{k2}, Z'_k < \Delta_{k1}) - 2\mathbb{P}(Z_j > \Delta_{j2}, Z_k < \Delta_{k1}, \Delta_{k1} < Z'_k < \Delta_{k2}) \\
&+ 2\mathbb{P}(\Delta_{j1} < Z_j < \Delta_{j2}, Z'_j < \Delta_{j1}, Z_k > \Delta_{k2}) - 2\mathbb{P}(Z_j < \Delta_{j1}, \Delta_{j1} < Z'_j < \Delta_{j2}, Z_k > \Delta_{k2}) \\
&+ 2\mathbb{P}(\Delta_{j1} < Z_j < \Delta_{j2}, Z'_j < \Delta_{j1}, \Delta_{k1} < Z_k < \Delta_{k2}, Z'_k < \Delta_{k1}) \\
&- 2\mathbb{P}(\Delta_{j1} < Z_j < \Delta_{j2}, Z'_j < \Delta_{j1}, Z_k < \Delta_{k1}, \Delta_{k1} < Z'_k < \Delta_{k2}) \\
= 2\mathbb{P}(&Z_j < \Delta_{j2}, Z_k < \Delta_{k2}) - 2\mathbb{P}(Z'_j < \Delta_{j2}, Z_k < \Delta_{k2}) \\
&+ 2\mathbb{P}(Z_k < \Delta_{k2}, Z'_k < \Delta_{k1}) - 2\mathbb{P}(Z_j < \Delta_{j2}, Z_k < \Delta_{k2}, Z'_k < \Delta_{k1}) \\
&- 2\mathbb{P}(Z_k < \Delta_{k1}, Z'_k < \Delta_{k1}) + 2\mathbb{P}(Z_j < \Delta_{j2}, Z_k < \Delta_{k1}, Z'_k < \Delta_{k1}) \\
&- 2\mathbb{P}(Z_k < \Delta_{k1}, Z'_k < \Delta_{k2}) + 2\mathbb{P}(Z_j < \Delta_{j2}, Z_k < \Delta_{k1}, Z'_k < \Delta_{k2}) \\
&+ 2\mathbb{P}(Z_k < \Delta_{k1}, Z'_k < \Delta_{k1}) - 2\mathbb{P}(Z_j < \Delta_{j2}, Z_k < \Delta_{k1}, Z'_k < \Delta_{k1})
\end{aligned}
$$

$$
\begin{aligned}
&+ 2\mathbb{P}(Z_j < \Delta_{j2}, Z'_j < \Delta_{j1}) - 2\mathbb{P}(Z_j < \Delta_{j2}, Z'_j < \Delta_{j1}, Z_k < \Delta_{k2}) \\
&- 2\mathbb{P}(Z_j < \Delta_{j1}, Z'_j < \Delta_{j1}) + 2\mathbb{P}(Z_j < \Delta_{j1}, Z'_j < \Delta_{j1}, Z_k < \Delta_{k2}) \\
&- 2\mathbb{P}(Z_j < \Delta_{j1}, Z'_j < \Delta_{j2}) + 2\mathbb{P}(Z_j < \Delta_{j1}, Z'_j < \Delta_{j2}, Z_k < \Delta_{k2}) \\
&+ 2\mathbb{P}(Z_j < \Delta_{j1}, Z'_j < \Delta_{j1}) - 2\mathbb{P}(Z_j < \Delta_{j1}, Z'_j < \Delta_{j1}, Z_k < \Delta_{k2}) \\
&+ 2\mathbb{P}(Z_j < \Delta_{j2}, Z'_j < \Delta_{j1}, Z_k < \Delta_{k2}, Z'_k < \Delta_{k1}) \\
&- 2\mathbb{P}(Z_j < \Delta_{j2}, Z'_j < \Delta_{j1}, Z_k < \Delta_{k1}, Z'_k < \Delta_{k1}) \\
&- 2\mathbb{P}(Z_j < \Delta_{j1}, Z'_j < \Delta_{j1}, Z_k < \Delta_{k2}, Z'_k < \Delta_{k1}) \\
&+ 2\mathbb{P}(Z_j < \Delta_{j1}, Z'_j < \Delta_{j1}, Z_k < \Delta_{k1}, Z'_k < \Delta_{k1}) \\
&- 2\mathbb{P}(Z_j < \Delta_{j2}, Z'_j < \Delta_{j1}, Z_k < \Delta_{k1}, Z'_k < \Delta_{k2}) \\
&+ 2\mathbb{P}(Z_j < \Delta_{j2}, Z'_j < \Delta_{j1}, Z_k < \Delta_{k1}, Z'_k < \Delta_{k1}) \\
&+ 2\mathbb{P}(Z_j < \Delta_{j1}, Z'_j < \Delta_{j1}, Z_k < \Delta_{k1}, Z'_k < \Delta_{k2}) \\
&- 2\mathbb{P}(Z_j < \Delta_{j1}, Z'_j < \Delta_{j1}, Z_k < \Delta_{k1}, Z'_k < \Delta_{k1}) \\
&= 2\mathbb{P}(Z_j < \Delta_{j2}, Z_k < \Delta_{k2}) - 2\mathbb{P}(Z'_j < \Delta_{j2}, Z_k < \Delta_{k2}) - 2\mathbb{P}(Z_j < \Delta_{j2}, Z_k < \Delta_{k2}, Z'_k < \Delta_{k1}) \\
&+ 2\mathbb{P}(Z_j < \Delta_{j2}, Z_k < \Delta_{k1}, Z'_k < \Delta_{k2}) - 2\mathbb{P}(Z_j < \Delta_{j2}, Z'_j < \Delta_{j1}, Z_k < \Delta_{k2}) \\
&+ 2\mathbb{P}(Z_j < \Delta_{j1}, Z'_j < \Delta_{j2}, Z_k < \Delta_{k2}) + 2\mathbb{P}(Z_j < \Delta_{j2}, Z'_j < \Delta_{j1}, Z_k < \Delta_{k2}, Z'_k < \Delta_{k1}) \\
&- 2\mathbb{P}(Z_j < \Delta_{j2}, Z'_j < \Delta_{j1}, Z_k < \Delta_{k1}, Z'_k < \Delta_{k2}) \\
&= 2\Phi_2(\Delta_{j2}, \Delta_{k2}; R_{jk}) - 2\Phi_1(\Delta_{j2})\Phi_1(\Delta_{k2}) \\
&- 2\Phi_2(\Delta_{j2}, \Delta_{k2}; R_{jk})\Phi_1(\Delta_{k1}) + 2\Phi_2(\Delta_{j2}, \Delta_{k1}; R_{jk})\Phi_1(\Delta_{k2}) \\
&- 2\Phi_2(\Delta_{j2}, \Delta_{k2}; R_{jk})\Phi_1(\Delta_{j1}) + 2\Phi_2(\Delta_{j1}, \Delta_{k2}; R_{jk})\Phi_1(\Delta_{j2}) \\
&+ 2\Phi_2(\Delta_{j2}, \Delta_{k2}; R_{jk})\Phi_2(\Delta_{j1}, \Delta_{k1}; R_{jk}) - 2\Phi_2(\Delta_{j2}, \Delta_{k1}; R_{jk})\Phi_2(\Delta_{j1}, \Delta_{k2}; R_{jk}) \\
&= 2\Phi_2(\Delta_{j2}, \Delta_{k2}; R_{jk}) - 2\Phi_1(\Delta_{j2})\Phi_1(\Delta_{k2}) \\
&- 4\Phi_2(\Delta_{k2}, \Delta_{j2}; R_{jk})\Phi_1(\Delta_{j1}) + 4\Phi_2(\Delta_{j1}, \Delta_{k2}; R_{jk})\Phi_1(\Delta_{j2}) \\
&+ 2\Phi_2(\Delta_{j2}, \Delta_{k2}; R_{jk})\Phi_2(\Delta_{j1}, \Delta_{k1}; R_{jk}) - 2\Phi_2(\Delta_{j2}, \Delta_{k1}; R_{jk})\Phi_2(\Delta_{j1}, \Delta_{k2}; R_{jk})
\end{aligned}
$$

Proof of Proposition 2.

To prove this lemma, we note that Fan et al. (2017) showed that the distribution function $\Phi_2(\cdot,\cdot;t)$ of a bivariate random variable (X_j, X_k) is strictly increasing with t. Therefore, we have $\partial\Phi_2(\Delta_j,\Delta_k;t)/\partial t > 0$ for fixed constants Δ_j and Δ_k. Also, Yoon, Carroll and Gaynanova (2020) proved that for any constants $a_1,\ldots,a_d$, let $\Phi_d(a_1,\ldots,a_d;\mathbf{\Sigma}_d(r))$ be the CDF of d-dimensional multivariate normal distribution with covariance matrix

$$\mathbf{\Sigma}_d(r) = \begin{pmatrix} 1 & \sigma_{12}(r) & \sigma_{13}(r) & \ldots & \sigma_{1d}(r) \\ \sigma_{21}(r) & 1 & \sigma_{23}(r) & \ldots & \sigma_{2d}(r) \\ & & 1 & & \\ \vdots & & & \ddots & \vdots \\ \sigma_{d1}(r) & & \ldots & & 1 \end{pmatrix}.$$

Then there exists $s_{ij}(r) > 0$ for all $r \in (-1,1)$ such that

$$\frac{\partial\Phi_d(a_1,\ldots,a_d;\mathbf{\Sigma}_d(r))}{\partial r} = \sum_{i=1}^{d-1}\sum_{j=i+1}^{d} s_{ij}(r)\frac{\partial\sigma_{ij}(r)}{\partial r}.$$

(a) Let

$$\boldsymbol{R}_{3g} = \begin{bmatrix} 1 & 0 & \frac{R_{jk}}{\sqrt{2}} \\ 0 & 1 & -\frac{R_{jk}}{\sqrt{2}} \\ \frac{R_{jk}}{\sqrt{2}} & -\frac{R_{jk}}{\sqrt{2}} & 1 \end{bmatrix}, \quad \boldsymbol{R}_{3h} = \begin{bmatrix} 1 & 0 & -\frac{R_{jk}}{\sqrt{2}} \\ 0 & 1 & -\frac{R_{jk}}{\sqrt{2}} \\ -\frac{R_{jk}}{\sqrt{2}} & -\frac{R_{jk}}{\sqrt{2}} & 1 \end{bmatrix},$$

we have $\Phi_3(\Delta_{j2},\Delta_{j1},0;\boldsymbol{R}_{3d}) = \Phi_3(\Delta_{j1},\Delta_{j2},0;\boldsymbol{R}_{3g})$, $\Phi(\Delta_{j2}) = \Phi_2(\Delta_{j2},0;\frac{R_{jk}}{\sqrt{2}}) + \Phi_2(\Delta_{j2},0;-\frac{R_{jk}}{\sqrt{2}})$, and $\Phi_2(\Delta_{j1},\Delta_{j2};0) = \Phi_3(\Delta_{j1},\Delta_{j2},0;\boldsymbol{R}_{3d}) + \Phi_3(\Delta_{j1},\Delta_{j2},0;\boldsymbol{R}_{3g})$.

Therefore,

$$
\begin{aligned}
&\frac{\partial F(t;\Delta_{j1},\Delta_{j2})}{\partial t} \\
&= 2\partial\Big\{\Phi_2(\Delta_{j2},0;\frac{t}{\sqrt{2}}) - \Phi_2(\Delta_{j2},0;-\frac{t}{\sqrt{2}}) - \Phi_3(\Delta_{j1},\Delta_{j2},0;\boldsymbol{R}_{3d}) + \Phi_3(\Delta_{j2},\Delta_{j1},0;\boldsymbol{R}_{3d})\Big\}/\partial t \\
&= 2\partial\Big\{\Phi_2(\Delta_{j2},0;\frac{t}{\sqrt{2}}) - \big\{\Phi(\Delta_{j2}) - \Phi_2(\Delta_{j2},0;\frac{t}{\sqrt{2}})\big\} \\
&\qquad\qquad + \Phi_3(\Delta_{j2},\Delta_{j1},0;\boldsymbol{R}_{3d}) - \big\{\Phi_2(\Delta_{j1},\Delta_{j2};0) - \Phi_3(\Delta_{j1},\Delta_{j2},0;\boldsymbol{R}_{3g})\big\}\Big\}/\partial t \\
&= 2\partial\Big\{2\Phi_2(\Delta_{j2},0;\frac{t}{\sqrt{2}}) - \Phi(\Delta_{j2}) - \Phi(\Delta_{j1})\Phi(\Delta_{j2}) + 2\Phi_3(\Delta_{j1},\Delta_{j2},0;\boldsymbol{R}_{3g})\Big\}/\partial t \\
&= 4\frac{\partial\Phi_2(\Delta_{j2},0;\frac{t}{\sqrt{2}})}{\partial t} + 4\frac{\partial\Phi_3(\Delta_{j1},\Delta_{j2},0;\boldsymbol{R}_{3g})}{\partial t}.
\end{aligned}
$$

Then, we only need to prove that $\partial\Phi_3(\Delta_{j1},\Delta_{j2},0;\boldsymbol{R}_{3g})/\partial t \geq 0$.

By the chain rule,

$$
\begin{aligned}
\frac{\partial\Phi_3(\Delta_{j1},\Delta_{j2},0;\boldsymbol{R}_{3g})}{\partial t} &= \sum_{i<k}\left\{\frac{\partial\Phi_3(\Delta_{j1},\Delta_{j2},0;\boldsymbol{R}_{3g})}{\partial\sigma_{ik}(r)}\frac{\partial\sigma_{ik}(r)}{\partial r}\right\} \\
&= \frac{1}{\sqrt{2}}\left\{\frac{\partial\Phi_3(\Delta_{j1},\Delta_{j2},0;\boldsymbol{R}_{3g})}{\partial\sigma_{13}(r)} - \frac{\partial\Phi_3(\Delta_{j1},\Delta_{j2},0;\boldsymbol{R}_{3g})}{\partial\sigma_{23}(r)}\right\}
\end{aligned}
$$

where σ_{ik} denotes the (i,k)-th element of $\boldsymbol{R}_{3g}$.

By equation (3) in Plackett (1954), we have the result that

$$
\partial\phi_d/\partial\sigma_{ik} = \partial^2\phi_d/(\partial x_i\partial x_k)
$$

Hence,

$$
\begin{aligned}
&\frac{\partial\Phi_3(\Delta_{j1},\Delta_{j2},0;\boldsymbol{R}_{3g})}{\partial t} = \frac{1}{\sqrt{2}}\left\{\frac{\partial\Phi_3(\Delta_{j1},\Delta_{j2},0;\boldsymbol{R}_{3g})}{\partial\sigma_{13}(r)} - \frac{\partial\Phi_3(\Delta_{j1},\Delta_{j2},0;\boldsymbol{R}_{3g})}{\partial\sigma_{23}(r)}\right\} \\
&= \frac{1}{\sqrt{2}}\left\{\int_{-\infty}^{\Delta_{j1}}\int_{-\infty}^{\Delta_{j2}}\int_{-\infty}^{0}\left(\frac{\partial\phi_3(x_1,x_2,x_3;\boldsymbol{R}_{3g})}{\partial\sigma_{13}(r)} - \frac{\partial\phi_3(x_1,x_2,x_3;\boldsymbol{R}_{3g})}{\partial\sigma_{23}(r)}\right)dx_1dx_2dx_3\right\} \\
&= \frac{1}{\sqrt{2}}\left\{\int_{-\infty}^{\Delta_{j1}}\int_{-\infty}^{\Delta_{j2}}\int_{-\infty}^{0}\left(\frac{\partial^2\phi_3(x_1,x_2,x_3;\boldsymbol{R}_{3g})}{\partial x_1\partial x_3} - \frac{\partial^2\phi_3(x_1,x_2,x_3;\boldsymbol{R}_{3g})}{\partial x_2\partial x_3}\right)dx_1dx_2dx_3\right\}
\end{aligned}
$$

$$= \frac{1}{\sqrt{2}}\Bigg\{ \int_{-\infty}^{\Delta_{j2}} \phi_3(\Delta_{j1}, x_2, 0; \boldsymbol{R}_{3g}) dx_2 - \int_{-\infty}^{\Delta_{j1}} \phi_3(x_1, \Delta_{j2}, 0; \boldsymbol{R}_{3g}) dx_1 \Bigg\}$$
$$= \frac{1}{\sqrt{2}}\Bigg\{ \int_{-\infty}^{\Delta_{j2}} \phi_3(\Delta_{j1}, x, 0; \boldsymbol{R}_{3g}) dx - \int_{-\infty}^{\Delta_{j1}} \phi_3(\Delta_{j2}, x, 0; \boldsymbol{R}_{3g}) dx \Bigg\}$$
$$= \frac{1}{\sqrt{2}}\Bigg\{ \phi_2(\Delta_{j1}, 0; \frac{R_{jk}}{\sqrt{2}}) \int_{-\infty}^{\Delta_{j2}} \phi(x) dx - \phi_2(\Delta_{j2}, 0; \frac{R_{jk}}{\sqrt{2}}) \int_{-\infty}^{\Delta_{j1}} \phi(x) dx \Bigg\}$$
$$= \frac{1}{\sqrt{2}}\Bigg\{ \phi_2(\Delta_{j1}, 0; \frac{R_{jk}}{\sqrt{2}}) \Phi(\Delta_{j2}) - \phi_2(\Delta_{j2}, 0; \frac{R_{jk}}{\sqrt{2}}) \Phi(\Delta_{j1}) \Bigg\}.$$

We have

$$\phi_2(x, y; \frac{R_{jk}}{\sqrt{2}}) = f_{X,Y}(x, y) = f_{X|Y}(x|y) f_Y(y) = \phi\Big(\frac{x}{\sqrt{1 - R_{jk}^2/2}}\Big) \phi(y).$$

Then,

$$\frac{\partial \Phi_3(\Delta_{j1}, \Delta_{j2}, 0; \boldsymbol{R}_{3g})}{\partial t} = \frac{1}{\sqrt{2}}\Bigg\{ \phi_2(\Delta_{j1}, 0; \frac{R_{jk}}{\sqrt{2}}) \Phi(\Delta_{j2}) - \phi_2(\Delta_{j2}, 0; \frac{R_{jk}}{\sqrt{2}}) \Phi(\Delta_{j1}) \Bigg\}$$
$$= \frac{1}{\sqrt{2}} \phi(0) \Bigg\{ \phi\Big(\frac{\Delta_{j1}}{\sqrt{1 - R_{jk}^2/2}}\Big) \Phi(\Delta_{j2}) - \phi\Big(\frac{\Delta_{j2}}{\sqrt{1 - R_{jk}^2/2}}\Big) \Phi(\Delta_{j1}) \Bigg\}$$

Therefore, we need to show that

$$\phi\Big(\frac{\Delta_{j1}}{\sqrt{1 - R_{jk}^2/2}}\Big) \Phi(\Delta_{j2}) > \phi\Big(\frac{\Delta_{j2}}{\sqrt{1 - R_{jk}^2/2}}\Big) \Phi(\Delta_{j1}),$$

which is equivalent to

$$\frac{\Phi(\Delta_{j2})}{\phi\Big(\frac{\Delta_{j2}}{s}\Big)} > \frac{\Phi(\Delta_{j1})}{\phi\Big(\frac{\Delta_{j1}}{s}\Big)},$$

where $s = \sqrt{1 - R_{jk}^2/2}$. Let $h(x) = \Phi(x)/\phi(x/s)$. Since $\Delta_{j2} > \Delta_{j1}$, we just need to show that $h(x)$ is an increasing function. We have

$$\frac{dh(x)}{dx} = \frac{\phi(x)\phi(x/s) + \frac{x}{s^2}\Phi(x)\phi(x/s)}{\phi^2(x/s)} = \frac{\phi(x)}{\phi(x/s)} + \frac{x}{s^2}\frac{\Phi(x)}{\phi(x/s)} = h(x)\Big(\frac{\phi(x)}{\Phi(x)} + \frac{x}{s^2}\Big).$$

When $x \geq 0$, it's obvious that $\phi(x)/\Phi(x) + x/s^2 \geq 0$. Since $x \sim N(0, s^2)$, from the property of Mill's ratio, we also know that for $x \geq 0$, $\phi(-x)/\Phi(-x) - x/s^2 \geq 0$, which means that when $x < 0$, it also holds that $\phi(x)/\Phi(x) + x/s^2 \geq 0$. Hence $h(x)$ is an increasing function.

(b)

$$\begin{aligned}\frac{\partial F(t;\Delta_{j1},\Delta_{j2},\Delta_k)}{\partial t} &= 2\left\{\frac{\partial \Phi_2(\Delta_{j2},\Delta_k;t)\big(1-\Phi_1(\Delta_{j1})\big)}{\partial t} + \frac{\partial \Phi_1(\Delta_{j2})\Phi_2(\Delta_{j1},\Delta_k;t)}{\partial t}\right\} \\ &= 2\left\{\big(1-\Phi_1(\Delta_{j1})\big)\frac{\partial \Phi_2(\Delta_{j2},\Delta_k;t)}{\partial t} + \Phi_1(\Delta_{j2})\frac{\partial \Phi_2(\Delta_{j1},\Delta_k;t)}{\partial t}\right\} \\ &> 0\end{aligned}$$

(c) Let

$$\boldsymbol{R}_{3i} = \begin{pmatrix} 1 & \frac{R_{jk}}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{R_{jk}}{\sqrt{2}} & 1 & R_{jk} \\ \frac{1}{\sqrt{2}} & R_{jk} & 1 \end{pmatrix},$$

$$\boldsymbol{R}_{4f} = \begin{pmatrix} 1 & \frac{R_{jk}}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{R_{jk}}{\sqrt{2}} & 1 & R_{jk} & 0 \\ \frac{1}{\sqrt{2}} & R_{jk} & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & 0 & 1 \end{pmatrix},$$

$$\boldsymbol{R}_{4g} = \begin{pmatrix} 1 & 0 & 0 & \frac{R_{jk}}{\sqrt{2}} \\ 0 & 1 & R_{jk} & \frac{R_{jk}}{\sqrt{2}} \\ 0 & R_{jk} & 1 & \frac{1}{\sqrt{2}} \\ \frac{R_{jk}}{\sqrt{2}} & \frac{R_{jk}}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 1 \end{pmatrix}.$$

$$\boldsymbol{R}_{4h} = \begin{pmatrix} 1 & \frac{R_{jk}}{\sqrt{2}} & \frac{1}{\sqrt{2}} & -\frac{R_{jk}}{\sqrt{2}} \\ \frac{R_{jk}}{\sqrt{2}} & 1 & R_{jk} & 0 \\ \frac{1}{\sqrt{2}} & R_{jk} & 1 & 0 \\ -\frac{R_{jk}}{\sqrt{2}} & 0 & 0 & 1 \end{pmatrix},$$

$$\boldsymbol{R}_{4i} = \begin{pmatrix} 1 & \frac{R_{jk}}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{R_{jk}}{\sqrt{2}} \\ \frac{R_{jk}}{\sqrt{2}} & 1 & R_{jk} & 0 \\ \frac{1}{\sqrt{2}} & R_{jk} & 1 & 0 \\ \frac{R_{jk}}{\sqrt{2}} & 0 & 0 & 1 \end{pmatrix}$$

We have

$$\Phi_3(0, \Delta_{j2}, \Delta_k; \boldsymbol{R}_{3f}) = \Phi_2(\Delta_{j2}, \Delta_k; R_{jk}) - \Phi_3(0, \Delta_{j2}, \Delta_k; \boldsymbol{R}_{3i}).$$

$$\Phi_4(0, \Delta_{j2}, \Delta_k, \Delta_k; \boldsymbol{R}_{4c}) = \Phi_1(\Delta_k)\Phi_2(\Delta_{j2}, \Delta_k; R_{jk}) - \Phi_3(0, \Delta_{j2}, \Delta_k; \boldsymbol{R}_{3i})$$

$$+ \Phi_4(0, \Delta_{j2}, \Delta_k, -\Delta_k; \boldsymbol{R}_{4f}).$$

$$\Phi_4(0, \Delta_{j2}, \Delta_{j1}\Delta_k; \boldsymbol{R}_{4d}) = \Phi_1(\Delta_{j2})\Phi_2(\Delta_{j1}, \Delta_k; R_{jk}) - \Phi_3(0, \Delta_{j1}, \Delta_k; \boldsymbol{R}_{3i})$$

$$+ \Phi_4(\Delta_{j2}, \Delta_{j1}, \Delta_k, 0; \boldsymbol{R}_{4g}).$$

$$\Phi_4(0, \Delta_{j2}, \Delta_{j1}, \Delta_k; \boldsymbol{R}_{4e}) = \Phi_4(0, \Delta_{j2}, \Delta_k, \Delta_{j1}; \boldsymbol{R}_{4h}) = \Phi_3(0, \Delta_{j2}, \Delta_k; \boldsymbol{R}_{3i})$$

$$- \Phi_4(0, \Delta_{j2}, \Delta_k, -\Delta_{j1}; \boldsymbol{R}_{4i}).$$

$$\Phi_4(0, \Delta_{j1}, \Delta_{j1}, \Delta_k; \boldsymbol{R}_{4e}) = \Phi_4(0, \Delta_{j1}, \Delta_k, \Delta_{j1}; \boldsymbol{R}_{4h}) = \Phi_3(0, \Delta_{j1}, \Delta_k; \boldsymbol{R}_{3i})$$

$$- \Phi_4(0, \Delta_{j1}, \Delta_k, -\Delta_{j1}; \boldsymbol{R}_{4i})$$

$$
\begin{aligned}
&\partial F(t;\Delta_{j1},\Delta_{j2},\Delta_k)/\partial t\\
&=2\partial\Big\{2\big(1-\Phi_1(\Delta_{j1})\big)\Phi_2(\Delta_{j2},\Delta_k;R_{jk})+\Phi_2(0,\Delta_{j2};-\frac{R_{jk}}{\sqrt{2}})-\Phi_1(\Delta_k)\Phi_2(\Delta_{j2},\Delta_k;R_{jk})\\
&-2\Phi_2(\Delta_{j1},\Delta_k;R_{jk})\Phi_2(\Delta_{j2},\Delta_k;R_{jk})+2\Phi_3(0,\Delta_{j2},\Delta_k;\boldsymbol{R}_{3e})-2\Phi_3(0,\Delta_{j2},\Delta_k;\boldsymbol{R}_{3f})\\
&+2\Phi_3(\Delta_{j2},\Delta_{j1},0;\boldsymbol{R}_{3d})+2\Phi_4(0,\Delta_{j2},\Delta_k,\Delta_k;\boldsymbol{R}_{4c})+2\Phi_4(0,\Delta_{j2},\Delta_{j1}\Delta_k;\boldsymbol{R}_{4d})\\
&+2\Phi_4(0,\Delta_{j2},\Delta_{j1},\Delta_k;\boldsymbol{R}_{4e})+2\Phi_4(0,\Delta_{j1},\Delta_{j1},\Delta_k;\boldsymbol{R}_{4e})\\
&-2\Phi_5(0,\Delta_{j1},\Delta_{j1},\Delta_k,\Delta_k;\boldsymbol{R}_5)+2\Phi_5(0,\Delta_{j2},\Delta_{j1},\Delta_k,\Delta_k;\boldsymbol{R}_5)\Big\}/\partial t\\
&=2\partial\Big\{2\big(1-\Phi_1(\Delta_{j1})\big)\Phi_2(\Delta_{j2},\Delta_k;R_{jk})+\Phi_1(\Delta_{j2})-\Phi_2(0,\Delta_{j2};R_{jk}/\sqrt{2})\\
&-2\Phi_2(\Delta_{j1},\Delta_k;R_{jk})\Phi_2(\Delta_{j2},\Delta_k;R_{jk})+2\Phi_3(0,\Delta_{j2},\Delta_k;\boldsymbol{R}_{3e})-2\Phi_2(\Delta_{j2},\Delta_k;R_{jk})\\
&+2\Phi_3(\Delta_{j2},\Delta_{j1},0;\boldsymbol{R}_{3d})+\Phi_1(\Delta_k)\Phi_2(\Delta_{j2},\Delta_k;R_{jk})+2\Phi_1(\Delta_{j2})\Phi_2(\Delta_{j1},\Delta_k;R_{jk})\\
&+2\Phi_4(0,\Delta_{j2},\Delta_k,-\Delta_k;\boldsymbol{R}_{4f})+2\Phi_4(\Delta_{j2},\Delta_{j1},\Delta_k,0;\boldsymbol{R}_{4g})\\
&+2\Phi_3(0,\Delta_{j2},\Delta_k;\boldsymbol{R}_{3i})-2\Phi_4(0,\Delta_{j2},\Delta_k,-\Delta_{j1};\boldsymbol{R}_{4i})\\
&-2\Phi_4(0,\Delta_{j1},\Delta_k,-\Delta_{j1};\boldsymbol{R}_{4i})\\
&-2\Phi_5(0,\Delta_{j1},\Delta_{j1},\Delta_k,\Delta_k;\boldsymbol{R}_5)+2\Phi_5(0,\Delta_{j2},\Delta_{j1},\Delta_k,\Delta_k;\boldsymbol{R}_5)\Big\}/\partial t\\
>&2\partial\big\{2\big(1-\Phi_1(\Delta_{j1})\big)\Phi_2(\Delta_{j2},\Delta_k;R_{jk})+\Phi_1(\Delta_{j2})-\Phi_2(0,\Delta_{j2};R_{jk}/\sqrt{2})\\
&-2\Phi_2(\Delta_{j1},\Delta_k;R_{jk})\Phi_2(\Delta_{j2},\Delta_k;R_{jk})+2\Phi_3(0,\Delta_{j2},\Delta_k;\boldsymbol{R}_{3e})-2\Phi_2(\Delta_{j2},\Delta_k;R_{jk})\\
&+2\Phi_3(\Delta_{j2},\Delta_{j1},0;\boldsymbol{R}_{3d})+\Phi_1(\Delta_k)\Phi_2(\Delta_{j2},\Delta_k;R_{jk})+2\Phi_1(\Delta_{j2})\Phi_2(\Delta_{j1},\Delta_k;R_{jk})\\
&+2\Phi_4(0,\Delta_{j2},\Delta_k,-\Delta_k;\boldsymbol{R}_{4f})+2\Phi_4(\Delta_{j2},\Delta_{j1},\Delta_k,0;\boldsymbol{R}_{4g})+2\Phi_3(0,\Delta_{j2},\Delta_k;\boldsymbol{R}_{3i})\\
&-2\Phi_4(0,\Delta_{j2},\Delta_k,-\Delta_{j1};\boldsymbol{R}_{4i})-2\Phi_4(0,\Delta_{j1},\Delta_k,-\Delta_{j1};\boldsymbol{R}_{4i})\big\}/\partial t\\
&>0
\end{aligned}
$$

(d) $\partial F(t;\Delta_{j1},\Delta_{j2},\Delta_{k1},\Delta_{k2})/\partial t$

$$
\begin{aligned}
&= 2\partial\Big\{ \Phi_2(\Delta_{j2},\Delta_{k2};t) - \Phi_2(\Delta_{j2},\Delta_{k2};t)\Phi_1(\Delta_{k1}) + \Phi_2(\Delta_{j2},\Delta_{k1};t)\Phi_1(\Delta_{k2}) \\
&\quad - \Phi_2(\Delta_{j2},\Delta_{k2};t)\Phi_1(\Delta_{j1}) + \Phi_2(\Delta_{j1},\Delta_{k2};t)\Phi_1(\Delta_{j2}) \\
&\quad + \Phi_2(\Delta_{j2},\Delta_{k2};t)\Phi_2(\Delta_{j1},\Delta_{k1};t) - \Phi_2(\Delta_{j2},\Delta_{k1};t)\Phi_2(\Delta_{j1},\Delta_{k2};t)\Big\}/\partial t \\
&= 2\partial\Big[\Phi_2(\Delta_{j2},\Delta_{k2};t)\Big\{1 - \Phi_1(\Delta_{k1}) - \Phi_1(\Delta_{j1}) + \Phi_2(\Delta_{j1},\Delta_{k1};t)\Big\} \\
&\quad + \Phi_2(\Delta_{j1},\Delta_{k2};t)\Phi_1(\Delta_{j2}) + \Phi_2(\Delta_{j2},\Delta_{k1};t)\big\{\Phi_1(\Delta_{k2}) - \Phi_2(\Delta_{j1},\Delta_{k2};t)\big\}\Big]/\partial t \\
&= 2\frac{\partial\big\{\Phi_2(\Delta_{j2},\Delta_{k2};t)\Phi_2(-\Delta_{j1},-\Delta_{k1};t)\big\}}{\partial t} + 2\Phi_1(\Delta_{j2})\frac{\partial\Phi_2(\Delta_{j1},\Delta_{k2};t)}{\partial t} \\
&\quad + 2\frac{\partial\Big[\Phi_2(\Delta_{j2},\Delta_{k1};t)\big\{\mathbf{P}(Z_k<\Delta_{k2}) - \mathbf{P}(Z_j<\Delta_{j1},Z_k<\Delta_{k2})\big\}\Big]}{\partial t} \\
&= 2\frac{\partial\big\{\Phi_2(\Delta_{j2},\Delta_{k2};t)\Phi_2(-\Delta_{j1},-\Delta_{k1};t)\big\}}{\partial t} + 2\Phi_1(\Delta_{j2})\frac{\partial\Phi_2(\Delta_{j1},\Delta_{k2};t)}{\partial t} \\
&\quad + 2\frac{\partial\Big[\Phi_2(\Delta_{j2},\Delta_{k1};t)\mathbf{P}(Z_j\geq\Delta_{j1},Z_k<\Delta_{k2})\Big]}{\partial t} \\
&> 0
\end{aligned}
$$

CHAPTER 5: IMPUTATION OF BLOCK-WISE MISSING VALUES IN MIXED MULTI-MODAL DATA

5.1 Introduction

In biomedical research, the block-wise missing structure is very common for high-dimensional multi-modal data. The development of technology provide researchers the opportunity to collect multi-modal data. However, the actual data collection process, especially for genetic data, is still expensive and the number of subjects available in each data modality will be limited due to various reasons. It is often commonly seen that not all subjects have information from every data modality, thus leading to a block-wise missing structure in the data. Another challenge for multi-modal data is that data from each modality might be of different types. An example is that in *Chlamydia trachomatis* genital tract infection researches, data collected have multi-source measurements, such as clinical variables, mRNA profiles, DNA genotypes, cervical microbiology, cervical cytokine profiles, endometrial microbiology and histology. The block of all cervical cytokine profiles in a subject could be missing due to the lack of cervical swap sample from that person. The block-wise missing values in the mixed-type data pose difficulty in the multi-modal data analysis, since most studies only utilize the completely observed data from a sample.

We first consider the approaches for solving the block-wise missing data problem. There are several popular approaches when a block-wise missing structure is presented, including complete-case analysis, use all data available, or impute missing values. The most straightforward way is to only use data with complete observations and remove those with any missing values (Liu et al., 2022). However, in many cases, only a small fraction of subjects have data from all modalities. Removing observations with any missing values, will lead to loss of much information. Another approach is to use all available data, without any deletion or imputation. Yuan et al. (2012) developed the iMSF method, a multi-task sparse learning framework for classification of patients' Alzheimer's disease (AD) progression, where observations with data from at least one modality can all be included. Xiang et al. (2014) extended the iMSF and proposed a bi-level model, where they performed covariate-level and modality-level analysis at the same time. Yu et al. (2020)

introduced the DISCOM method, where coefficients in the optimal linear prediction were estimated using an extended Lasso-type estimator, based on estimates for covariance matrices among covariates and between the response and covariates. The above methods deal with block missing values for supervised problems. Finally, for unsupervised problems, there are also works aim to impute block-wise missing data. From a matrix completion perspective, Cai, Cai and Zhang (2016) proposed a structured matrix completion (SMC) method based on singular value decomposition (SVD). They showed that the data matrix can be recovered if the matrix is exactly or approximately low-rank. However, SMC requires certain rows and columns to have complete data. Also, SMC can only be applied to Gaussian variables, and can not handle more than two modalities. A more recent work by Zhou, Cai and Lu (2021) proposed another matrix completion method (BONMI) that can complete multiple missing blocks. The BONMI exploits the orthogonal Procrustes problem and impute the block-wise missing data by the inner product of the low-rank components. However, their method is also based on SVD, therefore not applicable for binary, ordinal or truncated data. Zhang, Tang and Qu (2020) considered a factor model approach for imputing the missing blocks. The method does not rely on any specific missing mechanism. However, this approach can only be used for continuous data. To handle mixed variables, Xue and Qu (2020) proposed a multiple block-wise imputation (MBI) approach. Zhu, Li and Lock (2020) developed GIPCA, a low rank approach. For these two methods, since they used parametric models, covariates should follow distributions from the exponential family. For the *Chlamydia trachomatis* genital tract infection study, none of the methods mentioned above is ideal.

The second problem we face in the *Chlamydia trachomatis* genital tract infection study is that data from different modality are of mixed types. For example, the cytokine data are continuous, and microbiology data are binary or ordinal. In order to handle mixed variables, Fan et al. (2017) proposed a latent Gaussian copula model to measure the correlations between binary and continuous variables. The latent Gaussian copula model assumes that the observed continuous and binary variables are driven by some latent variables that follow the nonparanormal distribution (Liu, Lafferty and Wasserman, 2009), where Kendall's τ, a semiparametric rank-based correlation estimator was used to measure the latent correlations between them. They also derived formulae of bridge functions that maps the observed variables to the latent ones. Their works were then extended by Quan, Booth and Wells (2018); Yoon, Carroll and Gaynanova (2020); Liu et al. (2022) to incorporate ordinal and truncated variables. Quan, Booth and Wells (2018) provided formulae for measuring correlations among ordinal, continuous and binary variables. Yoon, Carroll and Gaynanova (2020) provided formulae for measuring correlations among truncated, continuous and binary variables. Liu

et al. (2022) provided formulae for measuring correlations between ordinal and truncated variables. However, these works do not allow missing values.

To efficiently utilize multi-modal data with mixed variables, we propose to impute the block-wise missing values using the Latent Mixed Gaussian Copula (LMGC) model. Using the LMGC model, we first impute the latent variables with missing values based on the observed variables without missing values. Then, we estimate the tranformation function to map the imputed latent variables to the mixed variables at their original scales.

The rest of this chapter is organized as follows. In Section 5.2, we describe the formulation and solution of our proposed method in details. In Section 5.3, we carry out extensive simulation studies to compare our method with some competitive methods. In Section 5.4, we apply our method to data collected from a *Chlamydia trachomatis* genital tract infection study.

5.2 Methodology

Assume that there are n subjects and for each subject, we observe a p-dimensional vector $\boldsymbol{X} = (X_1, \ldots, X_p)'$ containing variables of mixed types, such as continuous, binary, ordinal, or truncated variables. We use $\mathcal{C}$, $\mathcal{B}$, $\mathcal{G}$, $\mathcal{T}$ to denote continuous, binary, ordinal and truncated variables, respectively. We use $\mathcal{O}$ and $\mathcal{M}$ to denote observed and missing variables. Let $\mathcal{A} = \mathcal{B} \cup \mathcal{G} \cup \mathcal{T}$. We denote the set of variables other than continuous variables without any missing values by $\mathcal{A} \cap \mathcal{O}$.

We aim to impute the block missing data in $\mathcal{M}$ with variables in $\mathcal{C} \cap \mathcal{O}$ using the Latent Mixed Gaussian Copula (LMGC) model. The LMGC model assume that $\boldsymbol{X}$ is derived from some latent variable $\boldsymbol{Z}$. For any subject with missing values, write $\boldsymbol{X} = (\boldsymbol{X}'_{\mathcal{C}\cap\mathcal{O}}, \boldsymbol{X}'_{\mathcal{A}\cap\mathcal{O}}, \boldsymbol{X}'_{\mathcal{M}})'$ and $\boldsymbol{Z} = (\boldsymbol{Z}'_{\mathcal{C}\cap\mathcal{O}}, \boldsymbol{Z}'_{\mathcal{A}\cap\mathcal{O}}, \boldsymbol{Z}'_{\mathcal{M}})'$. Therefore, to impute $\boldsymbol{X}_{\mathcal{M}}$, we proposed to first estimate $\boldsymbol{Z}_{\mathcal{C}\cap\mathcal{O}}$. Then, we estimate $\boldsymbol{Z}_{\mathcal{M}}$ using $\boldsymbol{Z}_{\mathcal{C}\cap\mathcal{O}}$. Finally, we transform the estimated $\boldsymbol{Z}_{\mathcal{M}}$ to $\boldsymbol{X}_{\mathcal{M}}$.

In this section, we first introduce some basics of the LMGC model for complete data in Section 5.2.1 and then present our method in Section 5.2.2.

5.2.1 Latent Mixed Gaussian Copula (LMGC) model for mixed data

In this section, we introduce the Latent Mixed Gaussian Copula (LMGC) model for mixed data when there is no missing data presented.

Assume that we have a p-dimensional vector $\boldsymbol{X} = (X_1, \ldots, X_p)'$, containing variables of mixed types, such as continuous, binary, ordinal, or truncated variables. We assume that $\boldsymbol{X}$ is derived from latent continuous variables $\boldsymbol{Z} = (Z_1, \ldots, Z_p)'$ by the monotonically increasing transformation functions $\boldsymbol{f} = (f_1, \ldots, f_p)$ such that $\boldsymbol{Z} \sim N_p(\boldsymbol{0}, \boldsymbol{R})$, where $\boldsymbol{R}$ is a correlation matrix. We have

$$X_j = \begin{cases} f_j^{-1}(Z_j), & \text{if } j \in \mathcal{C}; \\ I(Z_j > \Delta_j), & \text{if } j \in \mathcal{B}; \\ I(Z_j > \Delta_j) f_j^{-1}(Z_j), & \text{if } j \in \mathcal{T}; \\ \sum_{l=1}^{L_j - 1} I(Z_j > \Delta_{j,l}), & \text{if } j \in \mathcal{G}; \end{cases} \tag{5.1}$$

where $\mathcal{C}$, $\mathcal{B}$, $\mathcal{T}$, and $\mathcal{G}$ are the index sets for continuous, binary, truncated, and ordinal variables respectively, and $\{\Delta_j\}_{j \in \mathcal{B}}$, $\{\Delta_j\}_{j \in \mathcal{T}}$ and $\{\Delta_{j,l}\}_{j \in \mathcal{G}, 1 \leq l \leq L_j - 1}$ are the corresponding cut-offs. We call (5.1) as the Latent Mixed Gaussian Copula (LMGC) model for mixed data. In the existing literature, Fan et al. (2017) studied the LMGC model for continuous and binary variables only. The idea was then extended by Quan, Booth and Wells (2018); Yoon, Carroll and Gaynanova (2020); Liu et al. (2022) to incorporate ordinal and truncated data type.

In all these works, the authors developed consistent estimators of the latent correlation matrix $\boldsymbol{R}$. They proposed to calculate the Kendall's τ correlations of observed variables and relate them to the correlations of latent variables via some bridge functions. In particular, let $\left\{(X_{ij}, X_{ik})\right\}_{i=1}^{n}$ be the realizations of the observed variables X_j and X_k, the Kendall's τ between X_j and X_k is defined as

$$\widehat{\tau}_{jk} = \frac{2}{n(n-1)} \sum_{1 \leq i < i' \leq n} \operatorname{sign}(X_{ij} - X_{i'j}) \operatorname{sign}(X_{ik} - X_{i'k}). \tag{5.2}$$

Let $\tau_{jk} = \mathbb{E}(\widehat{\tau}_{jk})$ be the population Kendall's τ. Then, the latent correlation between Z_j and Z_k is $R_{jk} = F_{jk}^{-1}(\tau_{jk})$, where $F_{jk}(\cdot)$ is a bridge function. The explicit form of the bridge functions $F_{jk}(\cdot)$ for the correlations between continuous, binary, 3 level ordinal and truncated variables are provided in the Supplementary materials. These formulae were derived in Fan et al. (2017), Quan, Booth and Wells (2018), Yoon, Carroll and Gaynanova (2020) and Liu et al. (2022).

Fan et al. (2017), Quan, Booth and Wells (2018), Yoon, Carroll and Gaynanova (2020) and Liu et al. (2022) proved that all bridge functions are strictly increasing for any $R_{jk} \in (-1, 1)$. Thus, they are invertible.

In practice, we estimate R_{jk} by $\widehat{R}_{jk} = F_{jk}^{-1}(\widehat{\tau}_{jk})$. For a binary or truncated variable, $\Delta_k = f_k(C_k)$ is unknown. We follow Fan et al. (2017) and estimate it with the plug-in estimator $\widehat{\Delta}_k = \Phi^{-1}\big\{\sum_{i=1}^n I(X_{ik} \neq 0)/n\big\}$. For a 3-level ordinal variable, we estimate Δ_{j1} and Δ_{j2} by the moment estimators as $\widehat{\Delta}_{j1} = \Phi^{-1}\big\{\sum_{i=1}^n I(X_{ij} = 0)/n\big\}$, and $\widehat{\Delta}_{j2} = \Phi^{-1}\big\{1 - \sum_{i=1}^n I(X_{ij} = 2)/n\big\}$.

For $j \in \mathcal{C}$ and $\mathcal{T}$, we will need to estimate the transformation function f_j. To estimate f_j, Liu, Lafferty and Wasserman (2009) proposed to use a Winsorized empirical C.D.F. estimator, which is defined as

$$\widehat{F}_j(t; \delta_n) := T_{\delta_n}\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{I}(X_{ij} \leq t)\right) \tag{5.3}$$

and

$$T_{\delta_n}(a) := \begin{cases} \delta_n, & \text{if } a < \delta_n, \\ a, & \text{if } \delta_n \leq a \leq 1 - \delta_n, \\ 1 - \delta_n, & \text{if } a > 1 - \delta_n. \end{cases}$$

Define

$$\widehat{f}_j(t) = \Phi^{-1}(\widehat{F}_j(t)) \tag{5.4}$$

and use the truncation level $\delta_n = \frac{1}{4n^{1/4}\sqrt{\pi \log n}}$, Han, Zhao and Liu (2013) proved that $\widehat{f}_j(t)$ converges to f_j uniformly over an expanding interval with high probability.

5.2.2 Imputation of missing values based on the conditional multivariate normal distribution

For any subject with missing values, let $\boldsymbol{X} = (\boldsymbol{X}'_{\mathcal{C}\cap\mathcal{O}}, \boldsymbol{X}'_{\mathcal{A}\cap\mathcal{O}}, \boldsymbol{X}'_{\mathcal{M}})'$ and $\boldsymbol{Z} = (\boldsymbol{Z}'_{\mathcal{C}\cap\mathcal{O}}, \boldsymbol{Z}'_{\mathcal{A}\cap\mathcal{O}}, \boldsymbol{Z}'_{\mathcal{M}})'$. Since variables $\mathcal{A} \cap \mathcal{O}$ were not used in the imputation process, for simplicity, we use $\mathcal{O}$ to represent $\mathcal{C} \cap \mathcal{O}$ in the rest of this section. To impute $\boldsymbol{X}_{\mathcal{M}}$, we propose to first estimate $\boldsymbol{Z}_{\mathcal{O}}$. Then, we estimate $\boldsymbol{Z}_{\mathcal{M}}$ using $\boldsymbol{Z}_{\mathcal{O}}$ and the correlations between $\boldsymbol{Z}_{\mathcal{O}}$ and $\boldsymbol{Z}_{\mathcal{M}}$. Finally, we transform the estimated $\boldsymbol{Z}_{\mathcal{M}}$ to $\boldsymbol{X}_{\mathcal{M}}$.

By definition, we have $(\boldsymbol{Z}'_{\mathcal{O}}, \boldsymbol{Z}'_{\mathcal{M}})' \sim N_p(\boldsymbol{0}, \boldsymbol{R})$, with

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{R}_{\mathcal{O}} & \boldsymbol{R}_{\mathcal{O}\mathcal{M}} \\ \boldsymbol{R}_{\mathcal{M}\mathcal{O}} & \boldsymbol{R}_{\mathcal{M}} \end{pmatrix},$$

where $\boldsymbol{R}_{\mathcal{O}}$, $\boldsymbol{R}_{\mathcal{M}}$, and $\boldsymbol{R}_{\mathcal{O}\mathcal{M}}$ are the correlation matrices of $\boldsymbol{Z}_{\mathcal{O}}$, $\boldsymbol{Z}_{\mathcal{M}}$ and between them.

By the conditional distribution of multivariate normal distribution, we have

$$\boldsymbol{Z}_{\mathcal{M}}|\boldsymbol{Z}_{\mathcal{O}} = \boldsymbol{z}_{\mathcal{O}} \sim N(\boldsymbol{R}'_{\mathcal{O}\mathcal{M}}\boldsymbol{R}_{\mathcal{O}}^{-1}\boldsymbol{z}_{\mathcal{O}}, \boldsymbol{R}_{\mathcal{M}} - \boldsymbol{R}'_{\mathcal{O}\mathcal{M}}\boldsymbol{R}_{\mathcal{O}}^{-1}\boldsymbol{R}_{\mathcal{O}\mathcal{M}}).$$

Therefore, we can estimate $\boldsymbol{Z}_{\mathcal{M}}$ by

$$\widehat{\boldsymbol{Z}}_{\mathcal{M}} = \widehat{\boldsymbol{R}}'_{\mathcal{O}\mathcal{M}}\widehat{\boldsymbol{R}}_{\mathcal{O}}^{-1}\widehat{\boldsymbol{Z}}_{\mathcal{O}}. \tag{5.5}$$

Define $S_{\mathcal{O}} = \{i : \boldsymbol{X}_i \text{ does not contain any missing values}\}$ and number of elements in $S_{\mathcal{O}}$ as $n_{\mathcal{O}}$. For $1 \leq j \leq p$, define $S_j = \{i : X_{ij} \text{ is not missing}\}$ and number of elements in S_j as n_j. For $1 \leq j, k \leq p$, define $S_{jk} = \{i : X_{ij} \text{ and } X_{ik} \text{ are not missing}\}$ and number of elements in S_{jk} as n_{jk}. To use (5.5), we need to estimate $\boldsymbol{R}$. By the LMGC model described in Section 5.2.1, we can estimate each element R_{jk} using the bridge function $F_{jk}(\cdot)$ and subjects in S_{jk}. When there are missing values, the Kendall's τ between X_j and X_k is defined as

$$\widehat{\tau}_{jk} = \frac{2}{n_{jk}(n_{jk}-1)} \sum_{i,i' \in S_{jk}} \operatorname{sign}(X_{ij} - X_{i'j}) \operatorname{sign}(X_{ik} - X_{i'k}). \tag{5.6}$$

The form of bridge functions $F_{jk}(\cdot)$ remains the same. For a binary or truncated variable, the estimator for $\Delta_k = f_k(C_k)$ becomes

$$\widehat{\Delta}_k = \Phi^{-1}\Big\{ \sum_{i \in S_k} I(X_{ik} \neq 0)/n_k \Big\}. \tag{5.7}$$

For a three-level ordinal variable, the estimators for Δ_{j1} and Δ_{j2} becomes

$$\begin{cases} \widehat{\Delta}_{j1} = \Phi^{-1}\big\{ \sum_{i \in S_j} I(X_{ij} = 0)/n_j \big\}, \\ \widehat{\Delta}_{j2} = \Phi^{-1}\big\{ 1 - \sum_{i \in S_j} I(X_{ij} = 2)/n_j \big\}. \end{cases} \tag{5.8}$$

For $j \in \mathcal{C}$ and $\mathcal{T}$, the Winsorized empirical C.D.F. estimator becomes

$$\widehat{F}_j(t; \delta_{n_j}) := T_{\delta_{n_j}}\left(\frac{1}{n_j} \sum_{i \in S_j} I(X_{ij} \leq t) \right) \tag{5.9}$$

where

$$T_{\delta_{n_j}}(a) := \begin{cases} \delta_{n_j}, & \text{if } a < \delta_{n_j} \\ a, & \text{if } \delta_{n_j} \leq a \leq 1-\delta_{n_j} \\ 1-\delta_{n_j}, & \text{if } a > 1-\delta_{n_j} \end{cases}, \text{ and } \delta_{n_j} = \frac{1}{4{n_j}^{1/4}\sqrt{\pi \log n_j}}.$$

We define

$$\widehat{f}_j(t) = \Phi^{-1}(\widehat{F}_j(t)). \tag{5.10}$$

Therefore, by (5.10), we could estimate $\boldsymbol{Z}_{\mathcal{O}}$ by

$$\widehat{Z}_j = \widehat{f}_j(X_j),\ \forall j \in \mathcal{O}. \tag{5.11}$$

Our imputation method is based on the fact that variables in $\boldsymbol{X}_{\mathcal{O}}$ are correlated with variables in $\boldsymbol{X}_{\mathcal{M}}$, and hence by considering the latent correlations between the observed and the missing data, the observed data could provide information for the imputation of the missing data. Let $\widetilde{\boldsymbol{R}}$ be the estimator of the correlation matrix whose elements are obtained by (5.6). When p is large, not all of observed variables are informative with the missing values and $\widetilde{\boldsymbol{R}}$ might not be a good estimator for the latent correlation matrix. If the actual correlations between the observed variables and the missing variables are small, including those variables would not help with the accuracy of imputation. Instead, those variables will introduce noise to the imputation process. When the extra error is larger than the extra information brought by these variables, thresholding and conditioning only on part of the variables for imputation will improve the accuracy. The sparsity assumption for the covariance matrix is frequently made to balance between biases and variances (Huang et al., 2006; d'Aspremont, Banerjee and El Ghaoui, 2008; Bickel and Levina, 2008; Rothman et al., 2008). We assume $\boldsymbol{R}$ has a sparse structure that is comprised of mostly zero values and $(\log p)/n \to 0$. Following Bickel and Levina (2008), we apply element-wise hard-thresholding to $\widetilde{\boldsymbol{R}}$. We denote the hard-thresholding results for $\widetilde{\boldsymbol{R}}$ as $\widehat{\boldsymbol{R}}$, which is given by

$$\widehat{R}_{jk} = \widetilde{R}_{jk} \cdot I(|\widetilde{R}_{jk}| \geq \lambda_{jk}) \text{ and } \widehat{\boldsymbol{R}} = \begin{pmatrix} \widehat{\boldsymbol{R}}_{\mathcal{O}} & \widehat{\boldsymbol{R}}_{\mathcal{O}\mathcal{M}} \\ \widehat{\boldsymbol{R}}_{\mathcal{M}\mathcal{O}} & \widehat{\boldsymbol{R}}_{\mathcal{M}} \end{pmatrix}, \tag{5.12}$$

where $\lambda_{jk} = c\sqrt{\frac{\log p}{n_{jk}}}$ is the hard-thresholding level and c is the tuning parameter. Therefore, we update (5.5) and estimate $\boldsymbol{Z}_{\mathcal{M}}$ by

$$\widehat{\boldsymbol{Z}}_{\mathcal{M}} = \widehat{\boldsymbol{R}}'_{\mathcal{O}\mathcal{M}}\{\widehat{\boldsymbol{R}}_{\mathcal{O}}\}^{-1}\widehat{\boldsymbol{Z}}'_{\mathcal{O}}. \tag{5.13}$$

We choose the optimal tuning parameter c by performing a grid search via cross-validation similar to that proposed in Bickel and Levina (2008). For a given c, we split the sample into two sets where one fifth of the subjects is reserved for testing and the rest is for training. We repeat this process T times. We split the data such that subjects in $S_{\mathcal{O}}$ are presented in both the training set and the testing set. For the t-th splitting process, we use the subjects in the testing set to calculate $\widetilde{\boldsymbol{R}}^{(t)}_{\text{test}}$ and use the training set to calculate $\widetilde{\boldsymbol{R}}^{(t)}_{\text{train}}$. Denote $\widehat{\boldsymbol{R}}^{(t)}_{\text{train}}$ as the hard-thresholding result using parameter c of $\widetilde{\boldsymbol{R}}^{(t)}_{\text{train}}$, we choose the optimal c that minimizes $\sum_{t=1}^{T} \|\widetilde{\boldsymbol{R}}^{(t)}_{\text{test}} - \widehat{\boldsymbol{R}}^{(t)}_{\text{train}}\|_F^2$. For the numerical studies, we set $T = 10$.

However, the resulting estimator $\widehat{\boldsymbol{R}}$ is not guaranteed to be positive semidefinite. In that case, we project it to the nearest positive semidefinite matrix by solving $\operatorname{argmin}_{\boldsymbol{A}\geq 0} \|\widehat{\boldsymbol{R}} - \boldsymbol{A}\|_F$, where $\boldsymbol{A} \geq 0$ means $\boldsymbol{A}$ is positive semidefinite. Such a problem can be solved by Zhao, Roeder and Liu (2014). With a slight abuse of notation, we still denote the solution as $\widehat{\boldsymbol{R}}$. Finally, by partitioning $\widehat{\boldsymbol{R}}$, we obtain estimates for $\widehat{\boldsymbol{R}}_{\mathcal{O}}$ and $\widehat{\boldsymbol{R}}_{\mathcal{O}\mathcal{M}}$ that can be used in (5.13).

In conclusion, and any variable j with missing values, our method for imputing different types of missing data can be implemented using the following formula:

$$\widehat{X}_j = \begin{cases} \widehat{f}_j^{-1}(\widehat{Z}_j), & \text{if } j \in \mathcal{C}; \\ I(\widehat{Z}_j > \widehat{\Delta}_j), & \text{if } j \in \mathcal{B}; \\ I(\widehat{Z}_j > \widehat{\Delta}_j)\widehat{f}_j^{-1}(\widehat{Z}_j), & \text{if } j \in \mathcal{T}; \\ \sum_{l=1}^{L_j-1} I(\widehat{Z}_j > \widehat{\Delta}_{j,l}), & \text{if } j \in \mathcal{G}; \end{cases} \tag{5.14}$$

where $\widehat{f}_j^{-1}$ can be estimated using (5.9) and (5.10), $\widehat{Z}_j$ can be estimated using (5.13) and $\widehat{\Delta}_j$ can be estimated by (5.7) or (5.8) based on the data type of j. In the following Algorithm 1, we give an algorithm describing steps of the imputation procedure.

5.2.3 Existing methods for block-wise imputation

We introduce three methods in existing literature for missing value imputation in this section.

Input: Data $\boldsymbol{X}$, consists of data from multiple modalities of different types, with missing values.
Output: Complete data $\widehat{\boldsymbol{X}}$, where no missing value is presented.
begin
With the completely observed continuous data $\boldsymbol{X}_{\mathcal{C}\cap\mathcal{O}}$ in $\boldsymbol{X}$, estimate its corresponding latent data $\widehat{\boldsymbol{Z}}_{\mathcal{C}\cap\mathcal{O}}$ using equations (5.3), (5.4) and (5.11)
With $\widehat{\boldsymbol{Z}}_{\mathcal{C}\cap\mathcal{O}}$, estimate missing latent data $\widehat{\boldsymbol{Z}}_{\mathcal{M}}$ that corresponding to the data with missing $\boldsymbol{X}_{\mathcal{M}}$ in $\boldsymbol{X}$ using the bridge function and (5.12), (5.13)
Estimate $\widehat{\boldsymbol{X}}_{\mathcal{M}}$ using the estimated latent data $\widehat{\boldsymbol{Z}}_{\mathcal{M}}$ with equations (5.7), (5.8) and (5.14)
Complete the missing entries in $\widehat{\boldsymbol{X}}$ with the estimated data $\widehat{\boldsymbol{X}}_{\mathcal{M}}$.
end

Algorithm 2: Algorithm for imputing missing data

The first is *Complete Case Analysis* method, which is not designed specifically for block-wise missing structure. For each $i \in S_{\mathcal{O}}$ and $j \in \mathcal{M}$, it builds a generalized linear model (GLM) McCullagh and Nelder, 1989 with $\boldsymbol{X}_j = (X_{j1}, X_{j2}, \ldots, X_{jn_{\mathcal{O}}})'$ as the response and variables in $\mathcal{O}$ as covariates, and obtain estimates for the regression coefficients $\widehat{\boldsymbol{\beta}}_j$. Let $p_{\mathcal{O}}$ be the number of variables in $\mathcal{O}$. To estimate $\widehat{\boldsymbol{\beta}}_j$, write the likelihood function as $\ell(\boldsymbol{\beta}_j)$, and we solve the following maximization problem

$$\widehat{\boldsymbol{\beta}}_j = \begin{cases} \operatorname{argmax}_{\boldsymbol{\beta}_j} \ell(\boldsymbol{\beta}_j), & \text{if } p_{\mathcal{O}} < n_{\mathcal{O}}, \\ \operatorname{argmax}_{\boldsymbol{\beta}_j} \Big\{ \ell(\boldsymbol{\beta}_j) + \lambda \|\boldsymbol{\beta}_j\|_1 \Big\}, & \text{otherwise.} \end{cases} \tag{5.15}$$

where $\|\boldsymbol{\beta}_j\|_1 = \sum_{k=1}^{p_{\mathcal{O}}} |\beta_{jk}|$ is the L_1-penalty, and λ is the tuning parameter which can be chosen by cross-validation. When $p_{\mathcal{O}} < n_{\mathcal{O}}$, (5.15) is a standard GLM problem. When $p_{\mathcal{O}} \geq n_{\mathcal{O}}$, (5.15) is the standard lasso problem (Tibshirani, 1996). After obtaining $\widehat{\boldsymbol{\beta}}_j$, predict variable j by calculating $\widehat{\boldsymbol{X}}_j = g^{-1}(\boldsymbol{X}'_{\mathcal{O}}\widehat{\boldsymbol{\beta}}_j)$, where g is the link function corresponding to the distribution of variable j. For this method, we impute one variable j each time.

The second is *Multiple Block-Wise Imputation* (MBI, Section 3.1, Xue and Qu, 2021) method. For this method, block-wise imputations were carried out multiple times based on each missing pattern and results are then integrated together for imputation of missing data. Based on the missing patterns across all data sources, divide n samples into G disjoint groups. For $g = 1, \ldots, G$, let $g = 1$ be the group with complete data, and for $g = 2, \ldots, G$, let $\mathcal{O}_g$ and $\mathcal{M}_g$ be the index sets of the observed covariates and missing covariates for group g. Let p_g be the number of elements in $\mathcal{O}_g$. Let N_g be the index set of subjects in group g and n_g be the number of subjects in group g. In addition, let $\mathcal{U}(g)$ be the index set of the groups g where missing variables in $\mathcal{M}_g$ and variables in at least one of the other sources are observed. For the complete group $g = 1$,

let $\mathcal{U}(1) = 1$. Denote the number of elements in $\mathcal{U}(g)$ as B_g. For group g with missing values in $\mathcal{M}_g$, each of the groups $u \in \mathcal{U}(g)$ contains both observed values corresponding to missing variables in $\mathcal{M}_g$, and observed values corresponding to a subset of observed variables $\mathcal{O}_g$. Let $\mathcal{O}_{gu}$ denote the index set of covariates which are observed in Groups g and u. To impute the missing values for each $j \in \mathcal{M}_g$, we build a GLM model for each u using for subjects in N_1, N_u with X_j as the response, $\boldsymbol{X}_{\mathcal{O}_{gu}}$ as the covariates, and obtain the estimated coefficients as $\widehat{\boldsymbol{\beta}}_{gu}$. Then, predict X_j for subjects in N_g using the GLM model with $\widehat{\boldsymbol{\beta}}_{gu}$ and $\boldsymbol{X}_{\mathcal{O}_{gu}}$. To estimate $\widehat{\boldsymbol{\beta}}_{gu}$, write the likelihood function as $\ell(\boldsymbol{\beta}_{gu})$, and we solve the following maximization problem

$$\widehat{\boldsymbol{\beta}}_{gu} = \begin{cases} \text{argmax}_{\boldsymbol{\beta}_{gu}} \, \ell(\boldsymbol{\beta}_{gu}), & \text{if } p_g < n_u, \\ \text{argmax}_{\boldsymbol{\beta}_{gu}} \Big\{ \ell(\boldsymbol{\beta}_{gu}) + \lambda \|\boldsymbol{\beta}_{gu}\|_1 \Big\}, & \text{otherwise.} \end{cases} \tag{5.16}$$

where $\|\boldsymbol{\beta}_{gu}\|_1 = \sum_{k=1}^{p_g} |\beta_{gu[k]}|$ is the L_1-penalty, and λ is the tuning parameter which can be chosen by cross-validation. When $p_g < n_u$, (5.16) is a standard GLM problem. When $p_g \geq n_u$, (5.16) is the standard lasso problem (Tibshirani, 1996). After imputing B_g times of X_j for each $j \in \mathcal{M}_g$, mean for the results of the B_g imputations was used as the imputed value for continuous variables, and mode for the results of the B_g imputations was used as the imputed value for discrete variables. This method incorporates more information compared to complete case analysis since data with only partially observed values were also included in the imputation process.

The third is *Structured Matrix Completion* (SMC, Cai, Cai and Zhang, 2016): This singular value decomposition (SVD) based method consider the imputation as an approximate low-rank matrix recovery problem and aims to recover the full matrix based on a subset of fully observed rows and columns. However, it can only handle continuous data and imputes one block at a time. It can not be applied on binary data or categorical data since its imputation rely on SVD. Assume that the observed data can be written in the following block form:

$$\begin{array}{c} \\ \boldsymbol{X} = \\ \\ \end{array} \begin{array}{c} \begin{array}{cc} p_1 & p - p_1 \end{array} \\ \begin{bmatrix} \boldsymbol{X}_{\mathcal{O}\mathcal{O}} & \boldsymbol{X}_{\mathcal{O}\mathcal{M}} \\ \boldsymbol{X}_{\mathcal{M}\mathcal{O}} & \boldsymbol{X}_{\mathcal{M}\mathcal{M}} \end{bmatrix} \\ \end{array} \begin{array}{c} \\ n_1 \\ n - n_1 \end{array}$$

The SMC method aims to recover the missing block $\boldsymbol{X}_{\mathcal{M}\mathcal{M}}$ based on the observed blocks $\boldsymbol{X}_{\mathcal{O}\mathcal{O}}, \boldsymbol{X}_{\mathcal{O}\mathcal{M}}$ and $\boldsymbol{X}_{\mathcal{M}\mathcal{O}}$. Let $\boldsymbol{X}_{[,\mathcal{O}]} = [\boldsymbol{X}_{\mathcal{O}\mathcal{O}}^\top, \boldsymbol{X}_{\mathcal{M}\mathcal{O}}^\top]^\top$ and $\boldsymbol{X}_{[\mathcal{O},]} = [\boldsymbol{X}_{\mathcal{O}\mathcal{O}}, \boldsymbol{X}_{\mathcal{O}\mathcal{M}}]$. Define $\boldsymbol{X}_{[:,1:r]}$ to be the submatrix

consists of the first r columns of matrix $\boldsymbol{X}$, $\boldsymbol{X}_{[1:r,:]}$ to be the submatrix consists of the first r rows of matrix $\boldsymbol{X}$, and $\boldsymbol{X}_{[1:r,1:r]}$ to be the submatrix consists of the first r columns and the first r rows of matrix $\boldsymbol{X}$.

The first step of SMC is to use SVD and move the significant factors of $\boldsymbol{X}_{[,\mathcal{O}]}$ and $\boldsymbol{X}_{[\mathcal{O},]}$ to the front. We calculate $\boldsymbol{X}_{[,\mathcal{O}]} = \boldsymbol{U}^{(1)}\boldsymbol{\Sigma}^{(1)}\boldsymbol{V}^{(1)\top}$, $\boldsymbol{X}_{[\mathcal{O},]} = \boldsymbol{U}^{(2)}\boldsymbol{\Sigma}^{(2)}\boldsymbol{V}^{(2)\top}$, and that $\widetilde{\boldsymbol{X}}_{\mathcal{O}\mathcal{O}} = \boldsymbol{U}^{(2)\top}\boldsymbol{X}_{\mathcal{O}\mathcal{O}}\boldsymbol{V}^{(1)}$, $\widetilde{\boldsymbol{X}}_{\mathcal{O}\mathcal{M}} = \boldsymbol{U}^{(2)\top}\boldsymbol{X}_{\mathcal{O}\mathcal{M}}$, $\widetilde{\boldsymbol{X}}_{\mathcal{M}\mathcal{O}} = \boldsymbol{X}_{\mathcal{M}\mathcal{O}}\boldsymbol{V}^{(1)}$. Next, to obtain a good estimate $\widehat{r}$ for the rank r of $\boldsymbol{X}$, which is the largest $\widehat{r}$ under the condition that $\widetilde{\boldsymbol{X}}_{\mathcal{O}\mathcal{O},[1:\widehat{r},1:\widehat{r}]}$ is nonsingular and that $\|\widetilde{\boldsymbol{X}}_{\mathcal{M}\mathcal{O},[1:\widehat{r},1:\widehat{r}]}\widetilde{\boldsymbol{X}}^{-1}_{\mathcal{O}\mathcal{O},[1:\widehat{r},1:\widehat{r}]}\| \leq 2\sqrt{n/n_1}$. Finally, the missing block $\boldsymbol{X}_{\mathcal{M}\mathcal{M}}$ can be estimated by $\widehat{\boldsymbol{X}}_{\mathcal{M}\mathcal{M}} = \widetilde{\boldsymbol{X}}_{\mathcal{M}\mathcal{O},[:,1:\widehat{r}]}\widetilde{\boldsymbol{X}}^{-1}_{\mathcal{O}\mathcal{O},[1:\widehat{r},1:\widehat{r}]}\widetilde{\boldsymbol{X}}_{\mathcal{O}\mathcal{M},[1:\widehat{r},:]}$.

5.3 Simulation

To evaluate the numerical performance of our method, we carry out simulation experiments and compare our method with three competitors introduced in the previous section, namely the Complete Case Analysis method, the Multiple Block-Wise Imputation (MBI) method, and the Structured Matrix Completion (SMC) method.

In the simulation experiments, Scenarios 1 to 5 are low-dimensional Scenarios where $n > p$, and Scenarios 6 to 8 are high-dimensional Scenarios where $n < p$. In Section 5.3.1, we use correlation matrices of the same structure but different in magnitude of the true correlation to examine the effect of true correlation on the performance of the imputation methods. In Section 5.3.2, we use the same latent correlation matrix, set up the missing so that unbiased estimators can be obtained, and compare the performance of the imputation methods on mixed-type data under different missing mechanisms in both the low- and high-dimensional settings. In addition, we also examined the effect of different transformation function on the imputation performance for continuous data modality. For the binary data modality, we consider the case where the two categories are either balanced or unbalanced in terms of sizes. When the missing mechanism is MNAR, we examine the imputation performance in Scenarios where unbiased estimator can not be obtained. These effects were checked in both the low- and the high-dimensional settings.

5.3.1 The impact of latent correlations on imputation

For the following two Scenarios, we focus on effect of the magnitude of true latent correlation on the performance of the different imputation methods.

Scenario 1: Consider a correlation matrix $\boldsymbol{S}_1$ for $p = 9$ with a block structure, where

$$\boldsymbol{S}_1 = \begin{pmatrix} \boldsymbol{S}_0 & \mathbf{0}_3 & \mathbf{0}_3 \\ \mathbf{0}_3 & \boldsymbol{S}_0 & \mathbf{0}_3 \\ \mathbf{0}_3 & \mathbf{0}_3 & \boldsymbol{S}_0 \end{pmatrix}, \quad \boldsymbol{S}_0 = \begin{pmatrix} 1 & r & r \\ r & 1 & r \\ r & r & 1 \end{pmatrix}, \quad \text{and} \quad \mathbf{0}_3 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

We set $r = 0.9$, use $n = 500$ and generate $\boldsymbol{Z} \sim N_9(\mathbf{0}, \boldsymbol{S}_1)$. For $j \in \{1, 2, 3, \ldots, 9\}$, set $\mathcal{O} = \{1, 4, 7\}$, $\mathcal{C} = \{2, 5, 8\}$ and $\mathcal{B} = \{3, 6, 9\}$. For $j \in \{\mathcal{C}, \mathcal{B}\}$, we generate the missing labels for each modality following a Bernoulli distribution with success probability equals to 0.2. Therefore, the missing mechanism for this Scenario is MCAR. We apply identity, cubic, and exponential transformation on the latent layer data $\boldsymbol{Z}$ to generate $\boldsymbol{X}$ for $j \in \{\mathcal{O}, \mathcal{C}\}$. For $j \in \mathcal{B}$, we set $\Delta_j = 0$ or -0.2 when the latent transformation for $j \in \{\mathcal{O}, \mathcal{C}\}$ is identity, 0 or $-\sqrt[3]{0.2}$ when the transformation is cubic, and 1 or $\log(0.6)$ when the transformation is exponential.

Scenario 2: The settings are the same as Scenario 1 except that we set $r = 0.6$.

After we generated the block-wise missing data for each Scenario, we apply our method and the three other competitors on the data to impute the missing values in each modality. We note again that SMC can only handle continuous data while our method and the two other competitors can handle both continuous data and binary data. Denote the continuous missing block in the second data modality as $\boldsymbol{X}_c$ and its estimates as $\widehat{\boldsymbol{X}}_c$. We evaluate the imputation performance for the continuous missing data using each method by calculating the Frobenius norm of the difference between the truth and the estimates, given by $\|\widehat{\boldsymbol{X}}_c - \boldsymbol{X}_c\|_F$. We evaluate the imputation performance for the binary missing data using each method by calculating the sensitivity, specificity, and overall accuracy for each method. The sensitivity is defined as the percentage of ones estimated correctly as ones and the specificity is defined as the percentage of zeros estimated correctly zeros for the missing binary block. The overall accuracy is defined as the percentage of entries that has their values correctly estimated.

We repeated the simulation 100 times under each Scenario and the results for Scenarios 1 and 2 are shown in Figures 5.1 and 5.2. In Figures 5.1 and 5.2, panel (a) shows the error of imputed data in $\mathcal{C}$. Panels (b), (c) and (d) shows the sensitivity, specificity and overall accuracy of the imputed data in $\mathcal{B}$. We consider two cases where the proportion for the two classes in binary data are balanced or unbalanced. When the two classes in binary data is unbalanced, there are more class one in the truth.

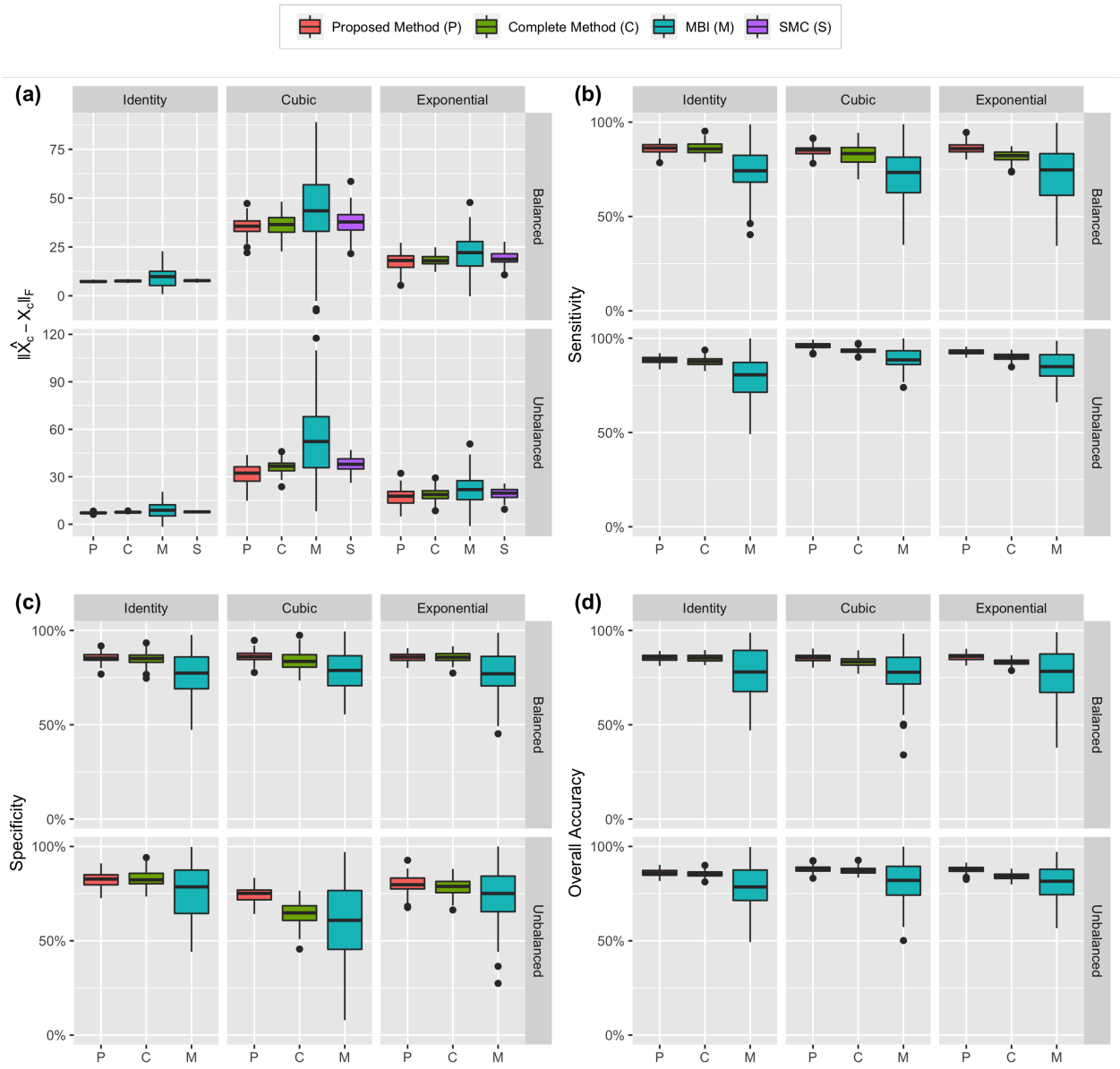


Figure 5.1: Simulation results for Scenario 1 when $r = 0.9$. Panel (a) shows the error of imputed data in $\mathcal{C}$. Panels (b), (c) and (d) shows the sensitivity, specificity and overall accuracy of the imputed data in $\mathcal{B}$.

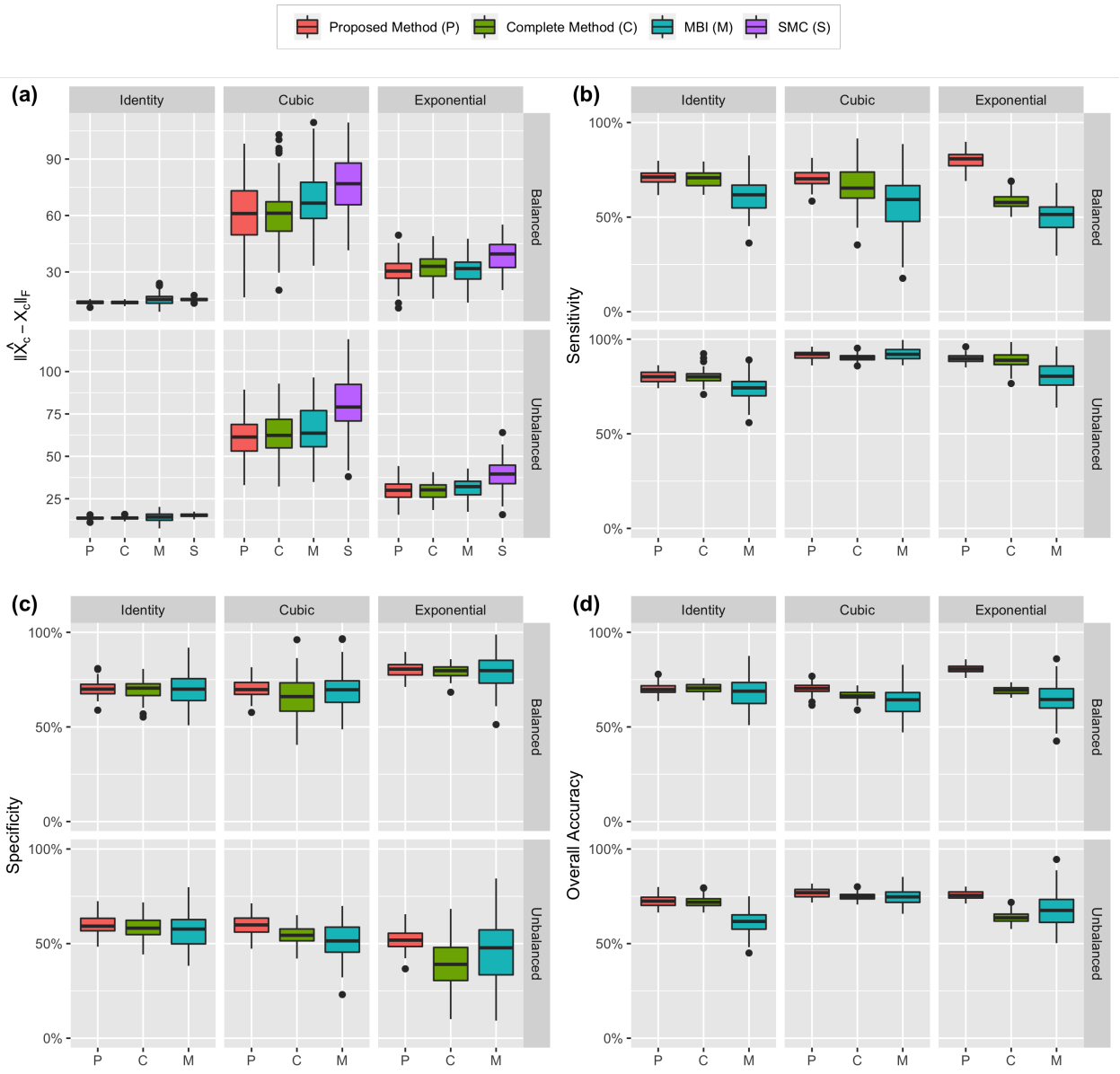


Figure 5.2: Simulation results for Scenario 2 when $r = 0.6$. Panel (a) shows the error of imputed data in $\mathcal{C}$. Panels (b), (c) and (d) shows the sensitivity, specificity and overall accuracy of the imputed data in $\mathcal{B}$.

From panels (a) in Figures 5.1 and 5.2, we could see that under both Scenarios, our proposed method have the smallest error under Frobenius norm for imputing the continuous missing block. It also have the highest overall accuracy for imputing the binary missing block. Overall, when the latent correlation between variables is stronger (Scenario 1), the performance of all methods will be better compared to when the latent correlation between variables is small (Scenario 2). When the latent correlation is high and the latent transformation function is an identity function, all methods have comparable performance for imputing the continuous block. The advantage of our method for imputing the continuous missing block gets bigger when there is a non-identity latent transformation and when the latent correlation is smaller. For the binary blocks, our method have a much higher accuracy for the classifying the smaller class when the latent transformation is non-identity and the two binary classes are not balanced.

5.3.2 The impact of missing mechanism on imputation

For the following six Scenarios, we focus on effect of different missing mechanism on the performance of the different imputation methods under low- and high-dimensional settings.

Scenario 3: Consider a correlation matrix $\boldsymbol{S}_2$ for $p = 16$ where

$$\boldsymbol{S}_{2[j_1,j_2]} = \begin{cases} 0.3^{|j_1-j_2|}, & \text{if } |j_1 - j_2| \leq 1 \\ 0.4, & \text{if } |j_1 - j_2| = 8 \\ 0, & \text{otherwise} \end{cases}.$$

We use $n = 500$ and generate $\boldsymbol{Z} \sim N_p(\boldsymbol{0}, \boldsymbol{S}_2)$. For $j \in \{1, 2, \ldots, 16\}$, set $\mathcal{O} = \{1, 2, \ldots, 8\}$, $\mathcal{C} = \{9, 10\}$, $\mathcal{B} = \{11, 12\}$, $\mathcal{G} = \{13, 14\}$ and $\mathcal{T} = \{15, 16\}$. For $j \in \{\mathcal{C}, \mathcal{B}, \mathcal{G}, \mathcal{T}\}$, we generate the missing labels for each modality following a Bernoulli distribution with success probability equals to 0.25. Therefore, the missing mechanism for this Scenario is MCAR. We apply exponential transformation functions, and set $\Delta_j = 0.6$ for $j \in \mathcal{B}$, $\Delta_{j1} = 0.6, \Delta_{j2} = 1.2$ for $j \in \mathcal{G}$ and $\Delta_j = 0.5$ for $j \in \mathcal{T}$.

Scenario 4: The same set-up is used as in Scenario 3 except for the missing labels. For $j \in \mathcal{C}$ and subject i, calculate $P_{\mathcal{C}i} = \text{logit}^{-1}(\gamma_{10} + \gamma_{11} * Z_{i1})$, where $\gamma_{10} = 0.2, \gamma_{11} = 0.3$. If $P_{\mathcal{C}i} \leq 0.5$, then set X_{ij} as missing. For $j \in \mathcal{B}$ and subject i, calculate $P_{\mathcal{B}i} = \text{logit}^{-1}(\gamma_{20} + \gamma_{21} * Z_{i3})$ where $\gamma_{20} = 0.15, \gamma_{21} = 0.35$. If $P_{\mathcal{B}i} \leq 0.5$, then set X_{ij} as missing. For $j \in \mathcal{G}$ and subject i, calculate $P_{\mathcal{G}i} = \text{logit}^{-1}(\gamma_{30} + \gamma_{31} * Z_{i5})$,

where $\gamma_{30} = 0.25, \gamma_{11} = 0.3$. If $P_{\mathcal{G}i} \leq 0.5$, then set X_{ij} as missing. For $j \in \mathcal{T}$ and subject i, calculate $P_{\mathcal{T}i} = \text{logit}^{-1}(\gamma_{40} + \gamma_{41} * Z_{i7})$ where $\gamma_{40} = 0.1, \gamma_{41} = 0.25$. If $P_{\mathcal{T}i} \leq 0.5$, then set X_{ij} as missing.

Scenario 5: The same set-up is used as in Scenario 3 except for the missing labels. For $j \in \mathcal{C}$ and subject i, calculate $P_{\mathcal{C}i} = \text{logit}^{-1}(\gamma_{10} + \gamma_{11} * Z_{i9})$, where $\gamma_{10} = 0.2, \gamma_{11} = 0.3$. If $P_{\mathcal{C}i} \leq 0.5$, then set X_{ij} as missing. For $j \in \mathcal{B}$ and subject i, calculate $P_{\mathcal{B}i} = \text{logit}^{-1}(\gamma_{20} + \gamma_{21} * Z_{i11})$ where $\gamma_{20} = 0.15, \gamma_{21} = 0.25$. If $P_{\mathcal{B}i} \leq 0.5$, then set X_{ij} as missing. For $j \in \mathcal{G}$ and subject i, calculate $P_{\mathcal{G}i} = \text{logit}^{-1}(\gamma_{30} + \gamma_{31} * Z_{i13})$, where $\gamma_{30} = 0.25, \gamma_{11} = 0.3$. If $P_{\mathcal{G}i} \leq 0.5$, then set X_{ij} as missing. For $j \in \mathcal{T}$ and subject i, calculate $P_{\mathcal{T}i} = \text{logit}^{-1}(\gamma_{40} + \gamma_{41} * Z_{i15})$ where $\gamma_{40} = 0.1, \gamma_{41} = 0.25$. If $P_{\mathcal{T}i} \leq 0.5$, then set X_{ij} as missing.

Scenario 6: Consider a correlation matrix $\boldsymbol{S}_3$ for $p = 200$ where

$$\boldsymbol{S}_{3[j_1,j_2]} = \begin{cases} 0.3^{|j_1 - j_2|}, & \text{if } |j_1 - j_2| \leq 1 \\ 0.4, & \text{if } |j_1 - j_2| = 120 \cdot \\ 0, & \text{otherwise} \end{cases}$$

We use $n = 100$ and generate $\boldsymbol{Z} \sim N_p(\boldsymbol{0}, \boldsymbol{S}_3)$. For $j \in \{1, 2, \ldots, 200\}$, set $\mathcal{O} = \{1, 2, \ldots, 120\}$, $\mathcal{C} = \{121, 122, \ldots, 140\}$, $\mathcal{B} = \{141, 142, \ldots, 160\}$, $\mathcal{G} = \{161, 162, \ldots, 180\}$, and $\mathcal{T} = \{181, 182, \ldots, 200\}$. For $j \in \{\mathcal{C}, \mathcal{B}, \mathcal{G}, \mathcal{T}\}$, we generate the missing labels for each modality following a Bernoulli distribution with success probability equals to 0.25. Therefore, the missing mechanism for this Scenario is MCAR. We apply exponential transformation functions for $j \in \{\mathcal{O}, \mathcal{C}\}$ and the same Δ_j for $j \in \{\mathcal{B}, \mathcal{G}, \mathcal{T}\}$ as in Scenario 3.

Scenario 7: The same set-up is used as in Scenario 6 except for the missing labels. For $j \in \mathcal{C}$ and subject i, calculate $P_{\mathcal{C}i} = \text{logit}^{-1}(\gamma_{10} + \gamma_{11} * Z_{i1})$, where $\gamma_{10} = 0.2, \gamma_{11} = 0.3$. If $P_{\mathcal{C}i} \leq 0.5$, then set X_{ij} as missing. For $j \in \mathcal{B}$ and subject i, calculate $P_{\mathcal{B}i} = \text{logit}^{-1}(\gamma_{20} + \gamma_{21} * Z_{i21})$ where $\gamma_{20} = 0.15, \gamma_{21} = 0.17$. If $P_{\mathcal{B}i} \leq 0.5$, then set X_{ij} as missing. For $j \in \mathcal{G}$ and subject i, calculate $P_{\mathcal{G}i} = \text{logit}^{-1}(\gamma_{30} + \gamma_{31} * Z_{i41})$, where $\gamma_{30} = 0.25, \gamma_{11} = 0.45$. If $P_{\mathcal{G}i} \leq 0.5$, then set X_{ij} as missing. For $j \in \mathcal{T}$ and subject i, calculate $P_{\mathcal{T}i} = \text{logit}^{-1}(\gamma_{40} + \gamma_{41} * Z_{i61})$ where $\gamma_{40} = 0.1, \gamma_{41} = 0.15$. If $P_{\mathcal{T}i} \leq 0.5$, then set X_{ij} as missing.

Scenario 8: The same set-up is used as in Scenario 6 except for the missing labels. For $j \in \mathcal{C}$ and subject i, calculate $P_{\mathcal{C}i} = \text{logit}^{-1}(\gamma_{10} + \gamma_{11} * Z_{i121})$, where $\gamma_{10} = 0.2, \gamma_{11} = 0.3$. If $P_{\mathcal{C}i} \leq 0.5$, then set X_{ij} as missing. For $j \in \mathcal{B}$ and subject i, calculate $P_{\mathcal{B}i} = \text{logit}^{-1}(\gamma_{20} + \gamma_{21} * Z_{i141})$ where $\gamma_{20} = 0.15, \gamma_{21} = 0.35$. If $P_{\mathcal{B}i} \leq 0.5$, then set X_{ij} as missing. For $j \in \mathcal{G}$ and subject i, calculate $P_{\mathcal{G}i} = \text{logit}^{-1}(\gamma_{30} + \gamma_{31} * Z_{i161})$,

where $\gamma_{30} = 0.25, \gamma_{11} = 0.3$. If $P_{\mathcal{G}i} \leq 0.5$, then set X_{ij} as missing. For $j \in \mathcal{T}$ and subject i, calculate $P_{\mathcal{T}i} = \text{logit}^{-1}(\gamma_{40} + \gamma_{41} * Z_{i181})$ where $\gamma_{40} = 0.3, \gamma_{41} = 0.15$. If $P_{\mathcal{T}i} \leq 0.5$, then set X_{ij} as missing.

We repeated the simulation 100 times under each Scenario. For ordinal and truncated variables, we evaluate the imputation performance for missing data by calculating the Kendall's τ between the truth and the estimated data. Results for Scenarios 3, 4 and 5 are displayed in Figure 5.3, and results for Scenarios 6, 7 and 8 are displayed in Figure 5.4. In both figures, panel (a) shows the error of imputed data in $\mathcal{C}$. Panel (b) shows the sensitivity, specificity and overall accuracy of the imputed data in $\mathcal{B}$. Panel (c) is the Kendall's τ between the imputed data and truth for the missing block in $\mathcal{G}$. Panels (d) is the Kendall's τ between the imputed data and truth for the missing block in $\mathcal{T}$. For the low-dimensional settings, we can see that our method has the smallest error for data in $\mathcal{C}$ among the four methods. For data in $\mathcal{B}$, since the two classes are not balanced, where there are more ones in the truth, we can see that all three methods have high and comparable sensitivity. However, the specificity for our method is much higher than the other two methods, suggesting that for the other two methods, a lot of zeros were misclassified as ones. Nevertheless, our method has the highest overall accuracy regardless of the missing mechanism compared to the other two methods when imputing binary data. As for ordinal data and truncated data, we can see from panels (c) and (d) that our method has a much higher $\widehat{\tau}$ under all Scenarios. A similar pattern can be observed in the high-dimensional settings as well, that our method has better performance in terms of imputing continuous, binary, ordinal and truncated data regardless of the missing mechanism as long as an unbiased estimator can be obtained.

5.3.3 Additional simulation results

We consider the following additional Scenarios for the simulation experiments. Scenarios 1* to 4* are low-dimensional Scenarios where $n > p$, and Scenarios 5* to 8* are high-dimensional Scenarios where $n < p$. We consider three different transformation functions for the continuous data modality: identity, cubic and exponential. For the binary data modality, we consider the case where the two categories are either balanced or unbalanced in terms of sizes. In Scenario 4*, we set up the missing so that unbiased estimators can not be obtained, and compare with Scenario 3* to examine the effect of unbiased estimator on the performance of the imputation methods. Similar Scenarios are considered in the high-dimensional settings as well, where in Scenario 8*, we set up the missing so that unbiased estimators can not be obtained, and compare with Scenario 7* to examine the effect of unbiased estimator. In conclusion, we aim to examine

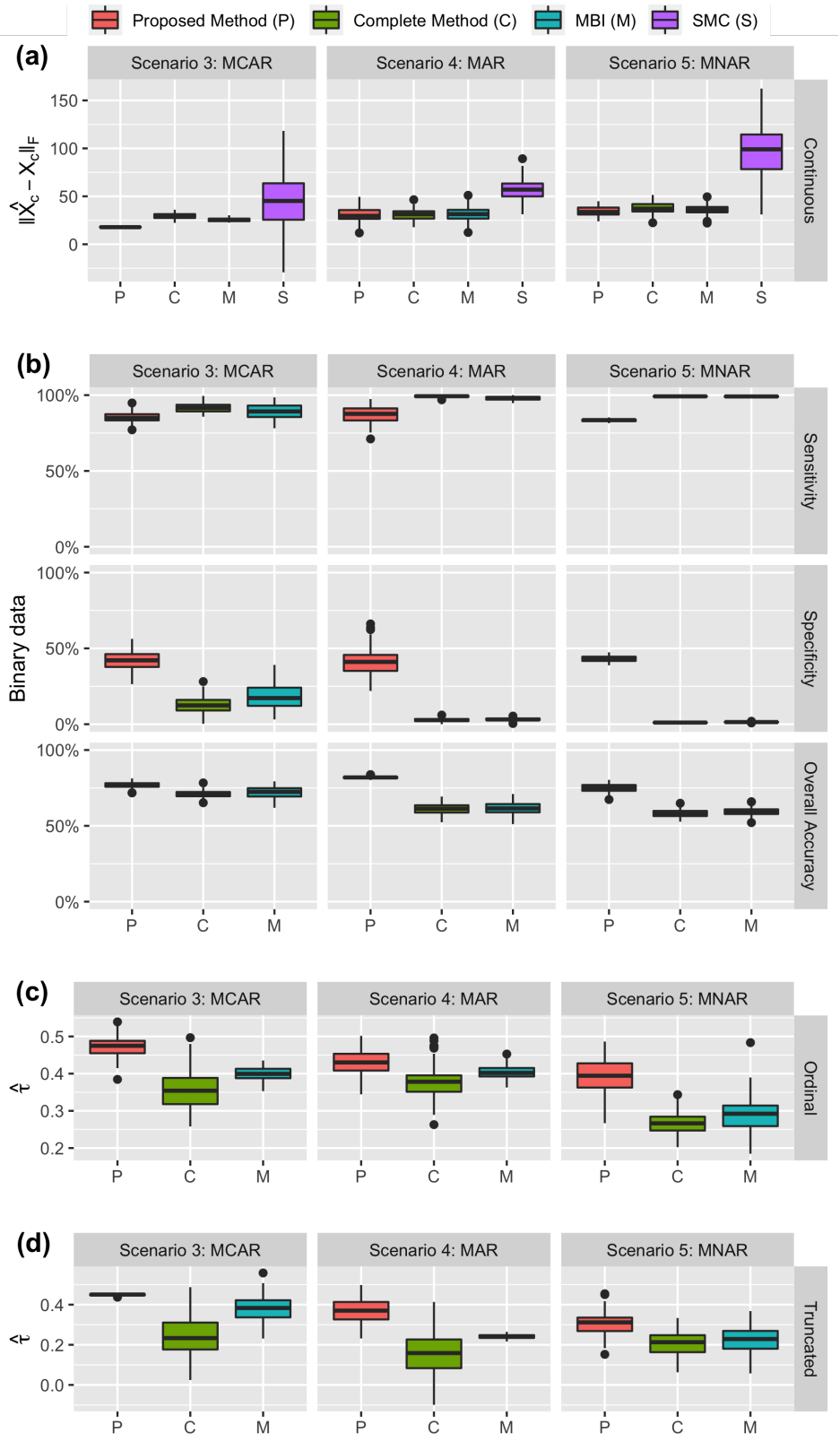


Figure 5.3: Imputation performance of four methods when imputing continuous data: panel (a), binary data: panel (b), 3-level ordinal data: panel (c) and truncated data: panel (d) under Scenarios 3, 4 and 5

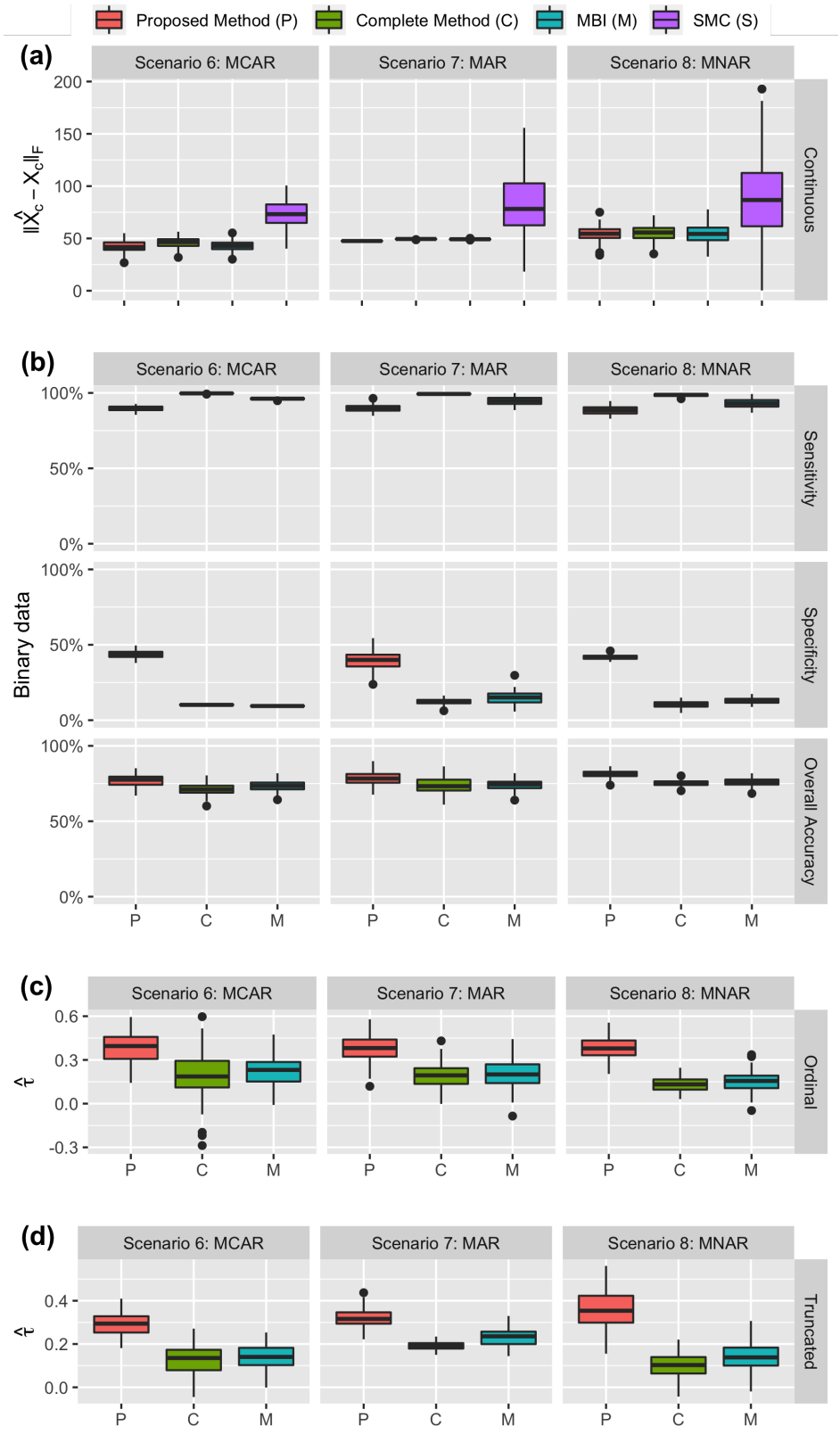


Figure 5.4: Imputation performance of four methods when imputing continuous data: panel (a), binary data: panel (b), 3-level ordinal data: panel (c) and truncated data: panel (d) under Scenarios 6, 7 and 8

the effect of missing mechanism, the mixed data type, and transformation function on the performance of imputation methods through the following simulation experiments.

Scenario 1*: Consider a correlation matrix $\boldsymbol{S}_4$ for $p = 10$ with a banded structure, where

$$\boldsymbol{S}_{4[j_1,j_2]} = \begin{cases} 0.6^{|j_1-j_2|}, & \text{if } |j_1 - j_2| \leq 2 \\ 0, & \text{otherwise} \end{cases}.$$

We use $n = 500$ and generate $\boldsymbol{Z} \sim N_{10}(\boldsymbol{0}, \boldsymbol{S}_4)$. For $j \in \{1, 2, 3, \ldots, 10\}$, set $\mathcal{O} = \{1, 2, 5, 8, 9, 10\}$, $\mathcal{C} = \{3, 4\}$ and $\mathcal{B} = \{6, 7\}$. For $j \in \{\mathcal{C}, \mathcal{B}\}$, we generate the missing labels for each modality following a Bernoulli distribution with success probability equals to 0.2. Therefore, the missing mechanism for this Scenario is MCAR. We apply identity, cubic, and exponential transformation on the latent layer data $\boldsymbol{Z}$ to generate $\boldsymbol{X}$ for $j \in \{\mathcal{O}, \mathcal{C}\}$. For $j \in \mathcal{B}$, we set $\Delta_j = 0$ or -0.2 when the latent transformation for $j \in \{\mathcal{O}, \mathcal{C}\}$ is identity, 0 or $-\sqrt[3]{0.2}$ when the transformation is cubic, and 1 or $\log(0.6)$ when the transformation is exponential.

Scenario 2*: The same set-up is used as in Scenario 1* except for the missing labels. Denote the $q\%$ quantile for $\boldsymbol{Z}_j$ by $\boldsymbol{Z}_{j;q\%}$. We set Z_{i3} and Z_{i4} to be missing if $Z_{i2} < \boldsymbol{Z}_{2;10\%}$ or $Z_{i2} > \boldsymbol{Z}_{2;90\%}$. We set Z_{i6} and Z_{i7} to be missing if $Z_{i8} < \boldsymbol{Z}_{8;10\%}$ or $Z_{i8} > \boldsymbol{Z}_{8;90\%}$. The missing mechanism for this Scenario is MAR.

Scenario 3*: The same set-up is used as in Scenario 1* except for the missing labels. We set Z_{i3} and Z_{i4} to be missing if $Z_{i4} < \boldsymbol{Z}_{4;10\%}$ or $Z_{i4} > \boldsymbol{Z}_{4;90\%}$. We set Z_{i6} and Z_{i7} to be missing if $Z_{i6} < \boldsymbol{Z}_{6;10\%}$ or $Z_{i6} > \boldsymbol{Z}_{6;90\%}$. The missing mechanism for this Scenario is MNAR.

Scenario 4*: The same set-up is used as in Scenario 1* except for the missing labels. We set Z_{i3} and Z_{i4} to be missing if $Z_{i4} < \boldsymbol{Z}_{4;20\%}$. We set Z_{i6} and Z_{i7} to be missing if $Z_{i6} < \boldsymbol{Z}_{6;20\%}$. The missing mechanism for this Scenario is MNAR.

Scenario 5*: We use $n = 100$ and generate $\boldsymbol{Z} \sim N_p(\boldsymbol{0}, \boldsymbol{S}_3)$. For $j \in \{1, 2, \ldots, 200\}$, set $\mathcal{C} = \{1, 2, \ldots, 160\}$ and $\mathcal{B} = \{161, 162, \ldots, 200\}$. Also, set $\mathcal{O} = \{1, 2, \ldots, 120\}$ and $\mathcal{M} = \{121, 122, \ldots, 200\}$. For $j \in \mathcal{M}$, we generate the missing labels for each modality following a Bernoulli distribution with success probability equals to 0.25. Therefore, the missing mechanism for this Scenario is MCAR. We apply the same transformation functions for and the same Δ_j for $j \in \mathcal{B}$ as in Scenario 1.

Scenario 6*: The same set-up is used as in Scenario 5* except for the missing labels. For $j \in \mathcal{C} \cap \mathcal{M}$, we set Z_{ij} to be missing if $Z_{i1} < \boldsymbol{Z}_{1;10\%}$ or $Z_{i1} > \boldsymbol{Z}_{1;90\%}$. For $j \in \mathcal{B} \cap \mathcal{M}$, we set Z_{ij} to be missing if $Z_{i41} < \boldsymbol{Z}_{41;10\%}$ or $Z_{i41} > \boldsymbol{Z}_{41;90\%}$. The missing mechanism for this Scenario is MAR.

Scenario 7*: The same set-up is used as in Scenario 5* except for the missing labels. For $j \in \mathcal{C} \cap \mathcal{M}$, we set Z_{ij} to be missing if $Z_{i121} < \boldsymbol{Z}_{121;10\%}$ or $Z_{i121} > \boldsymbol{Z}_{121;90\%}$. For $j \in \mathcal{B} \cap \mathcal{M}$, we set Z_{ij} to be missing if $Z_{i161} < \boldsymbol{Z}_{161;10\%}$ or $Z_{i161} > \boldsymbol{Z}_{161;90\%}$. The missing mechanism for this Scenario is MNAR.

Scenario 8*: The same set-up is used as in Scenario 5* except for the missing labels. For $j \in \mathcal{C} \cap \mathcal{M}$, we set Z_{ij} to be missing if $Z_{i121} < \boldsymbol{Z}_{121;20\%}$. For $j \in \mathcal{B} \cap \mathcal{M}$, we set Z_{ij} to be missing if $Z_{i161} < \boldsymbol{Z}_{161;20\%}$. The missing mechanism for this Scenario is MNAR.

The results for Scenarios 1* - 8*, are displayed in Figures 5.5 to 5.12. We repeated the simulation 100 times under each Scenario. From Figures 5.5 to 5.7, we can see that our proposed method have the smallest error under Frobenius norm for imputing the continuous missing block under these Scenarios. It also have the highest overall accuracy for imputing the binary missing block. When there is no transformation, the complete data method is comparable to our method in terms of error for both the continuous block and the binary block. However, when there is a non-identity latent transformation, our method performs much better for the binary block. Further more, we also considered the effect of latent binary cut-off value on the imputation. Take the cubic transformation for example, when using the -0.2 cut-off, the resulting data includes more 1s than 0s. Since 0 is the smaller class, it is harder to identify correctly. The specificity of our method under that setting is much higher than the other two competitors, meaning that the method performs well even when the data is not balanced and have better performance in classifying the smaller class correctly. We can observe a similar pattern under the exponential transformation when using the 0.6 cut-off. For the binary blocks, the specificity of our method is much higher than the other two competitors and remains so when the missing mechanism is MAR and MNAR. Under Scenario 4, the missing values for both blocks are all values less than the 20% quantile and therefore, the moment estimator based on the observed data is biased. Under this Scenario, we could see from Figure 5.8 that our proposed method still have the smallest error under frobenius norm for imputing the continuous missing block compared to the other methods. And the sensitivity for imputing the binary missing block is the highest under all transformations and different binary cut-offs. Under identity transformation and exponential transformation using 1 as the binary cut-off, the MBI method has the highest specificity and the highest overall accuracy, but its sensitivity is also the lowest among all methods, while our method have a much higher sensitivity and a relatively satisfying specificity and the

highest overall accuracy in general. Therefore, as long as we can obtain unbiased moment estimators from the observed data, our method has very desirable performance for both continuous and binary data even when the missing mechanism is not MCAR. Scenarios 5* to 8* are the high-dimensional cases compared with Scenarios 1* to 4*. From Figures 5.9 to 5.12, we can see a similar pattern as in the low-dimensional settings, where our method out-performs all other methods when imputing both continuous and binary variables under various missing mechanisms.

5.4 Multi-Modal Data from *Chlamydia trachomatis* Genital Tract Infection Study

We assessed our method and the alternatives in real data from the T cell Response Against Chlamydia (TRAC) study (Russell et al., 2016) in this section. The Institutional Review Boards for Human Subject Research at the University of Pittsburgh and the University of North Carolina approved the study and all participants provided written informed consent prior to inclusion. *Chlamydia trachomatis* (*C. trachomatis*) can ascend from the lower genital tract (e.g., vagina, cervix) to the upper genital tract (e.g., uterus and fallopian tubes) in some women, potentially resulting in severe reproductive disease sequelae. Infection is often asymptomatic. Diagnosis of endometrial histology and infection is critically needed but frequently missing due to challenges in obtaining sufficient endometrial biopsy samples. Another challenge is the high proportion of missing cytokine values in cervical secretion, which significantly influences the interpretation of final data. The major goal of this real data analysis is to impute the missing endometrial histologic diagnosis and *C. trachomatis* infection status, as well as cervical cytokine expressions. Data from five different modalities with 240 women in TRAC cohort at enrollment were used in this study (Table 5.1). These data were collected from vaginal samples, cervical samples (including *C. trachomatis* burden, microbiology and protein expression of cytokines) and endometrial biopsies (including *C. trachomatis* infection status and histologic evaluation), and contained both block-wise and element-wise missing.

5.4.1 Data preprocess

There are three continuous variables with complete observations in the vaginal and cervical microbiology data, including Cervical *C. trachomatis* burden, Nugent score and count of cervical white B cells. Cervical *C. trachomatis* burden was log10 transformed.

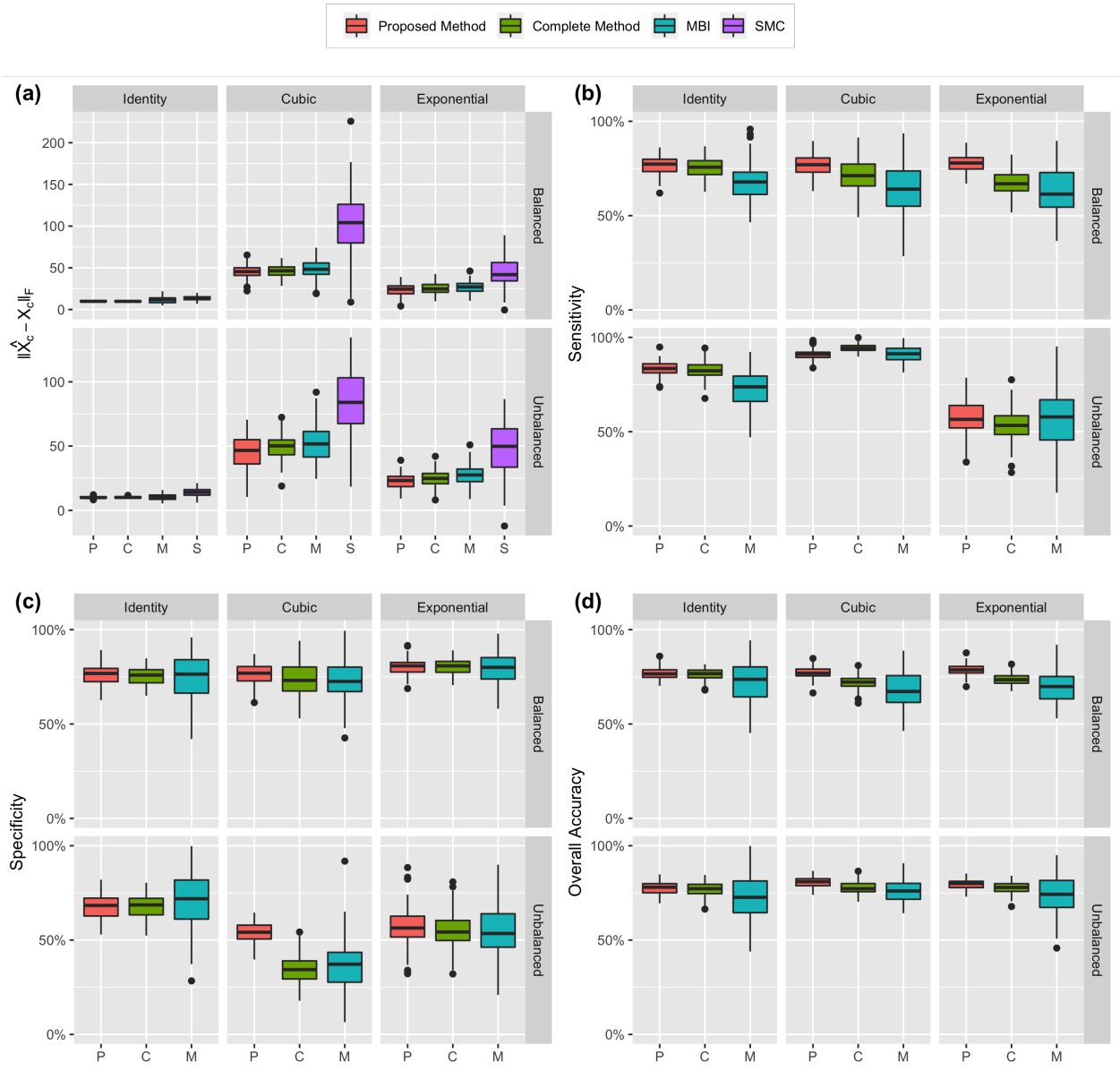


Figure 5.5: Simulation results under Scenario 1* (MCAR, 20% missing)

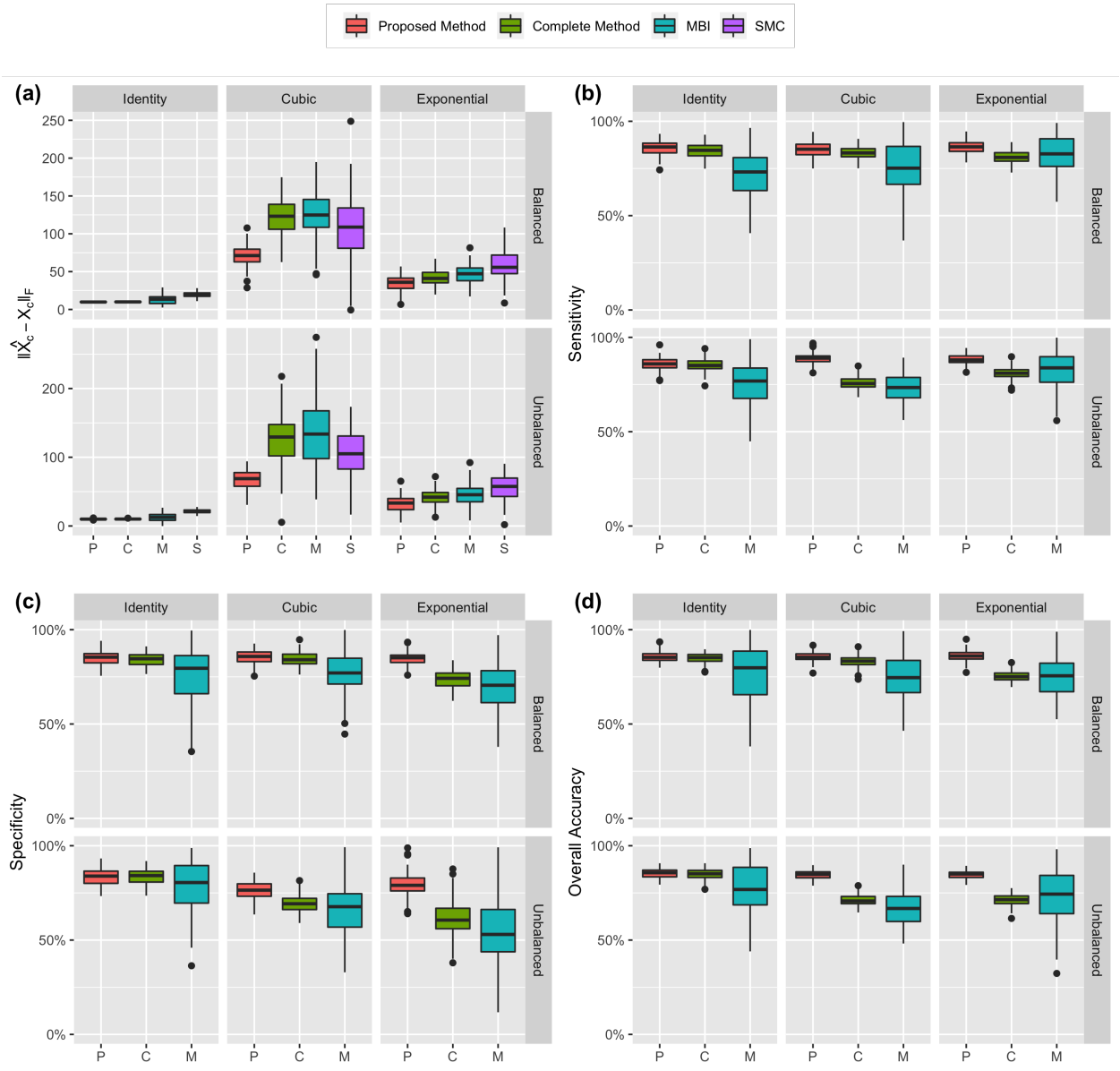


Figure 5.6: Simulation results under Scenario 2* (MAR, 20% missing)

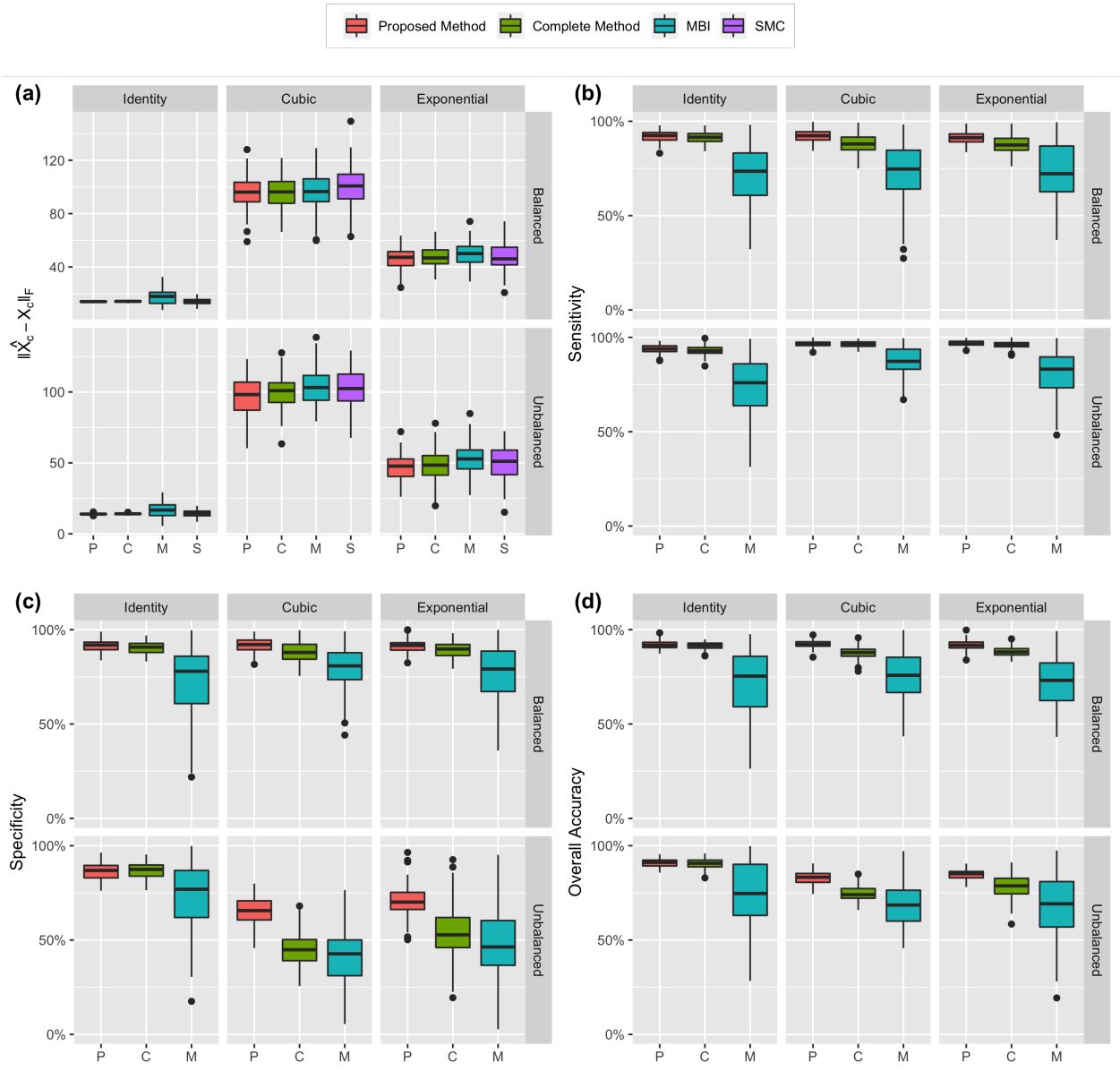


Figure 5.7: Simulation results under Scenario 3* (MNAR, 20% missing)

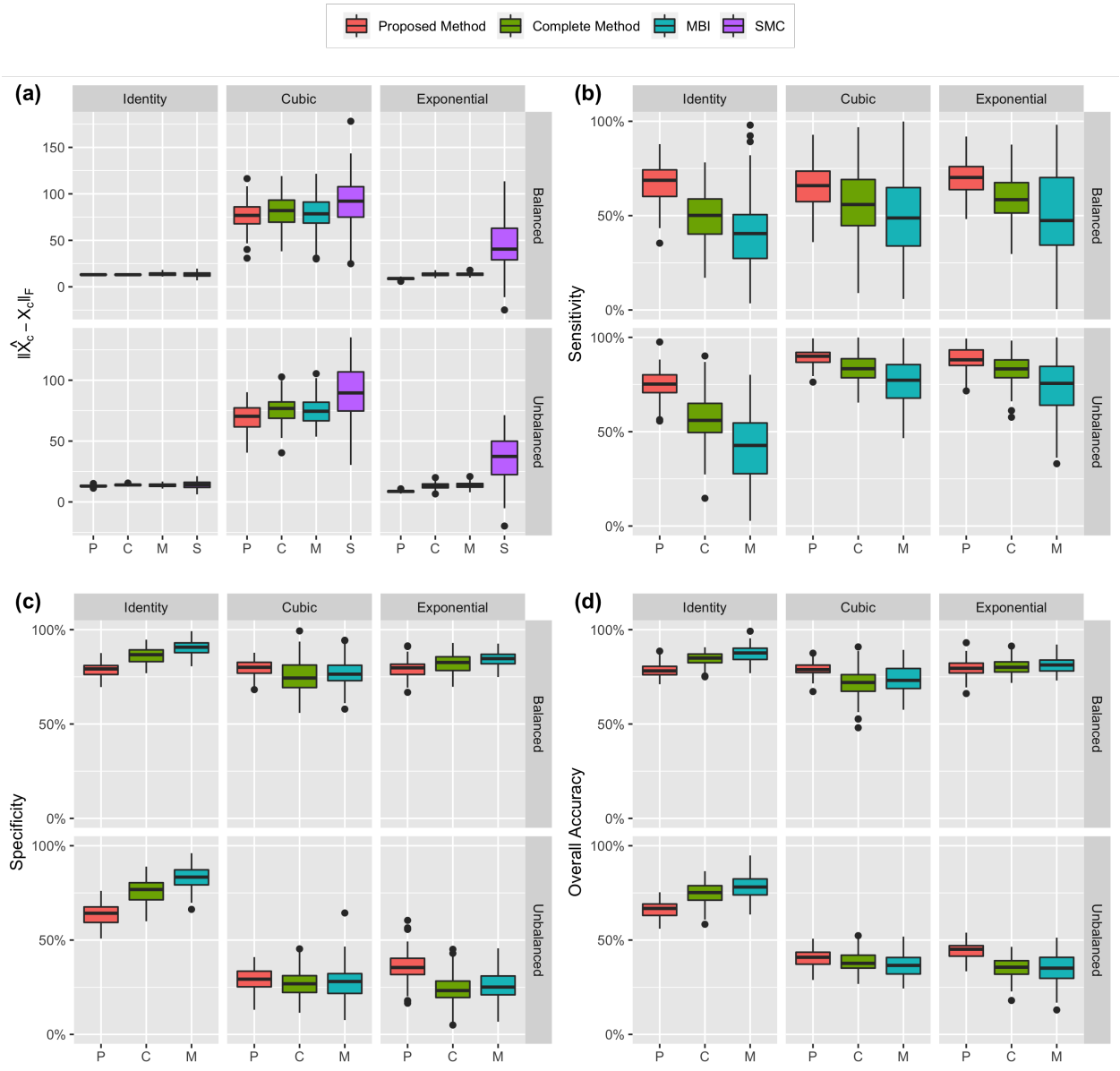


Figure 5.8: Simulation results under Scenario 4* (MNAR, 20% missing)

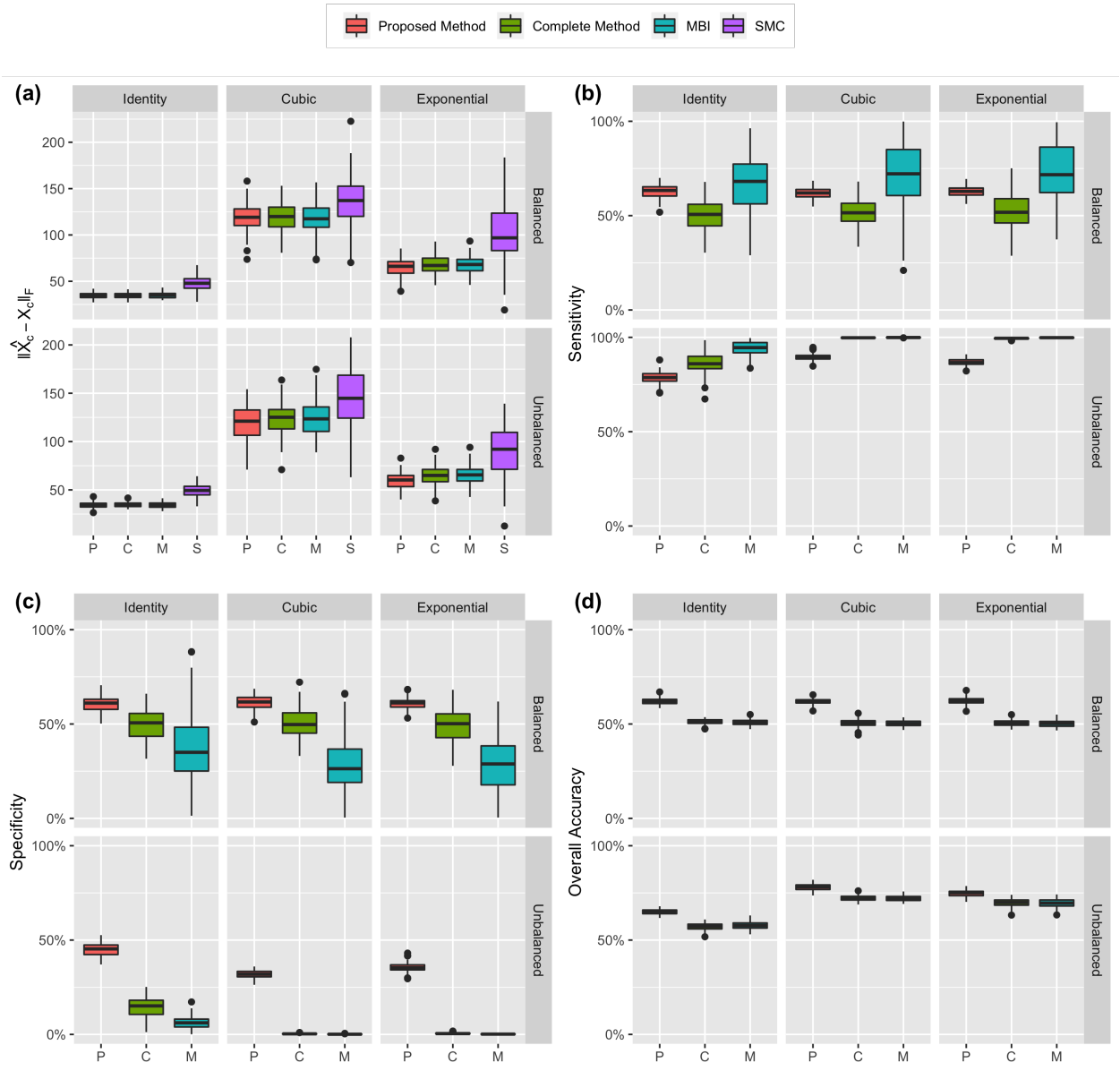


Figure 5.9: Simulation results under Scenario 5* (MCAR, 25% missing)

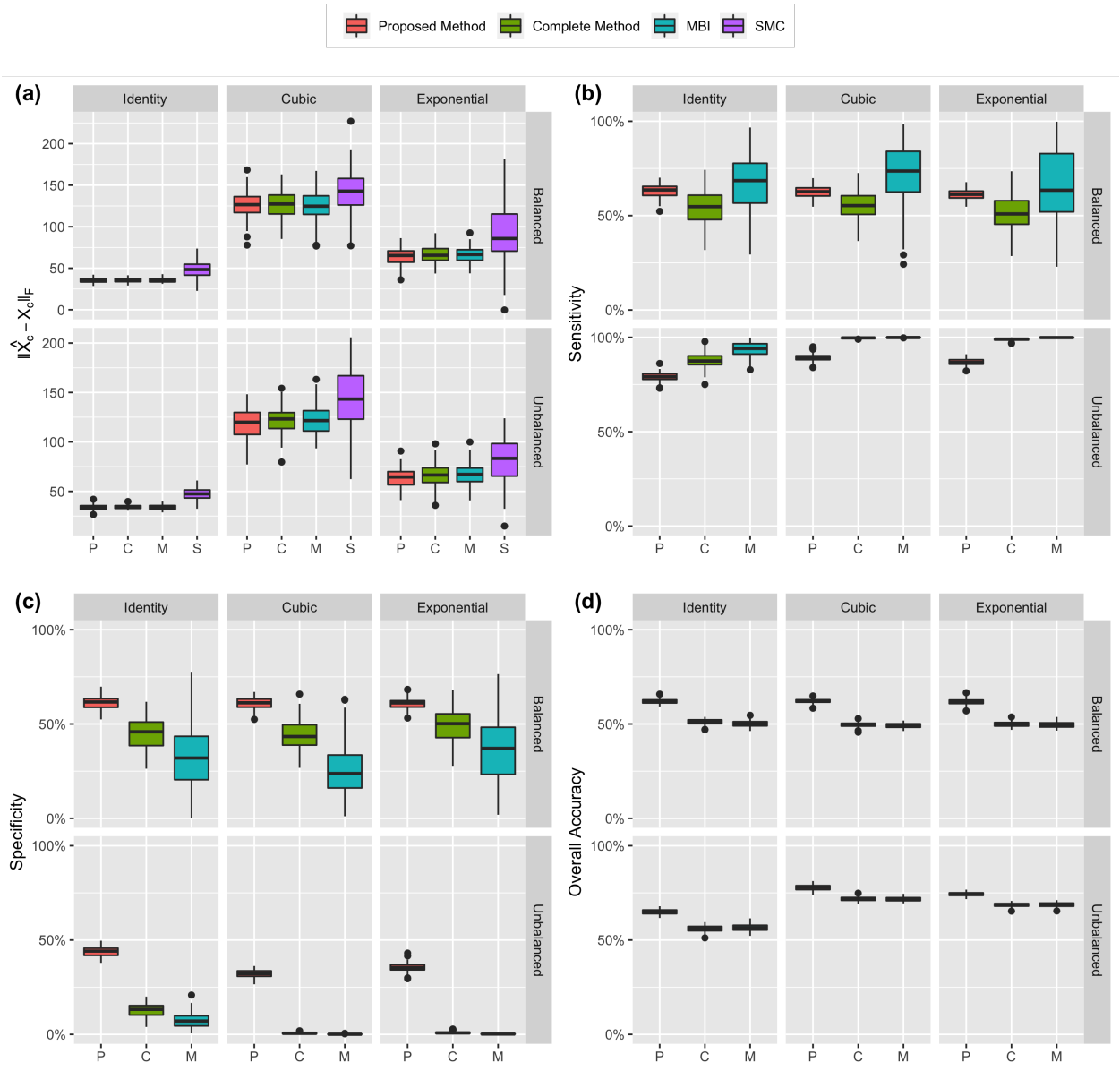


Figure 5.10: Simulation results under Scenario 6* (MAR, 25% missing)

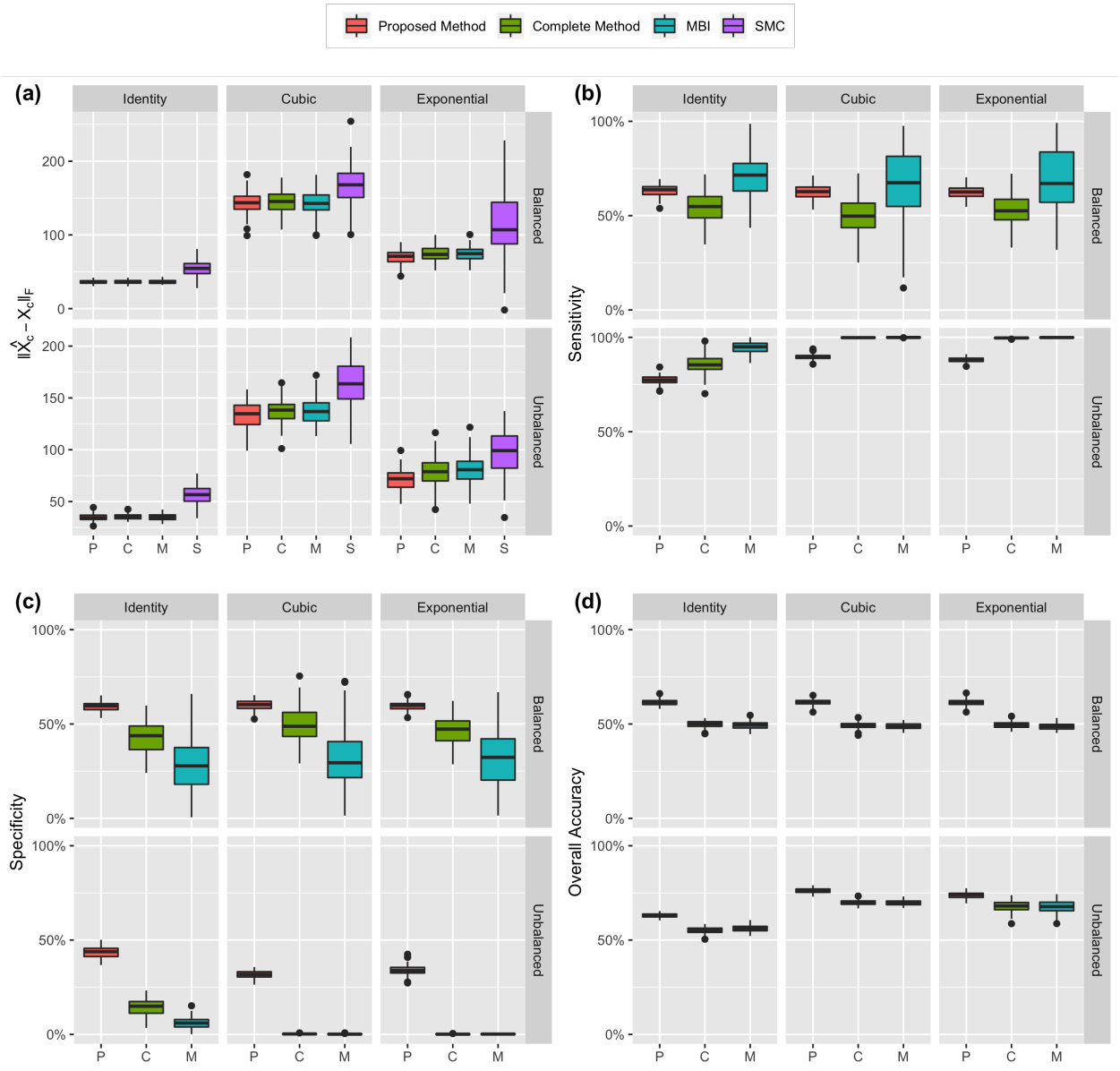


Figure 5.11: Simulation results under Scenario 7* (MNAR, 25% missing)

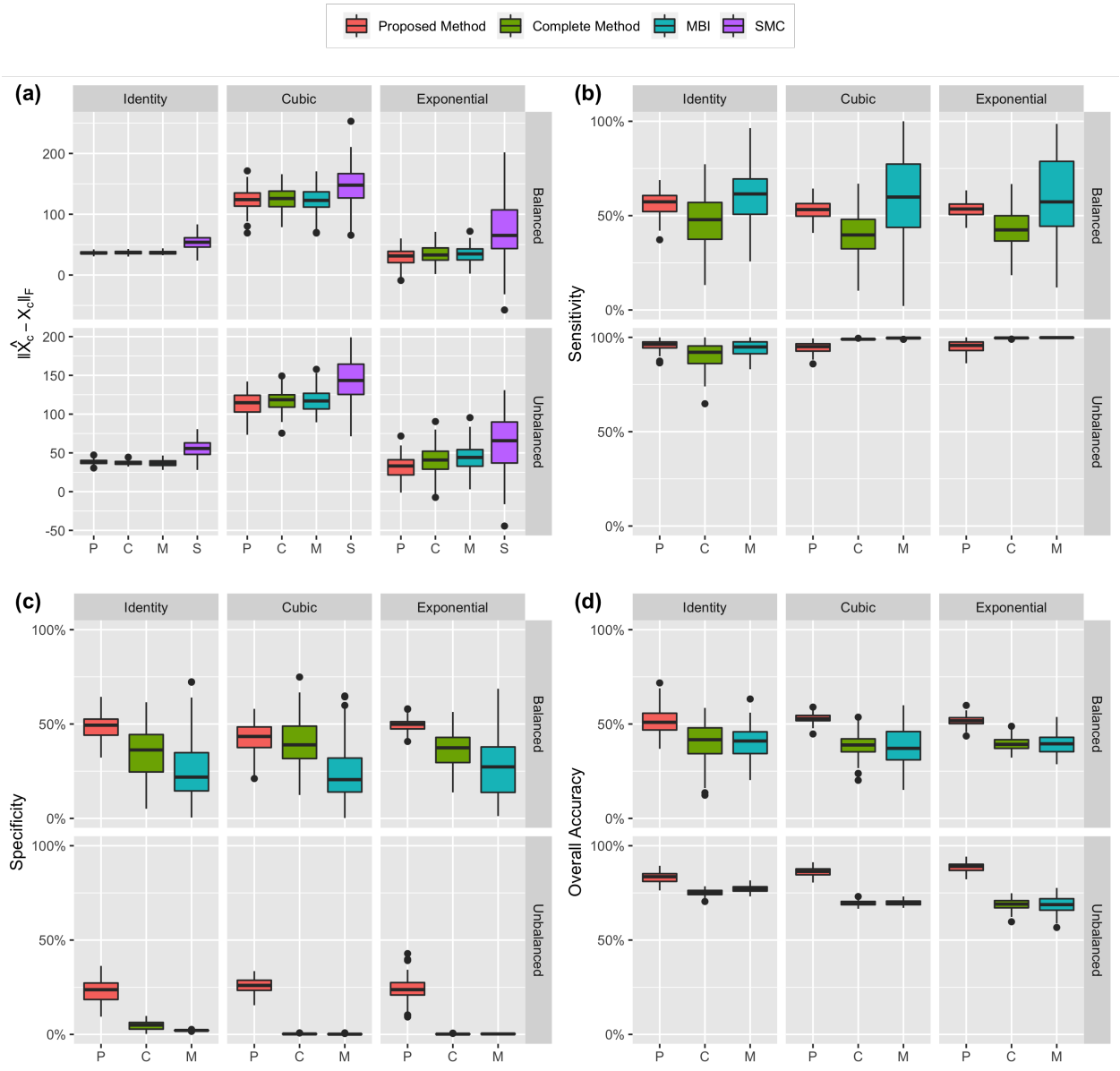


Figure 5.12: Simulation results under Scenario 8* (MNAR, 25% missing)

Table 5.1: Block-wise missingness of multi-modalities data from TRAC cohort

Missing Pattern	Vaginal and Cervical microbiology continuous (3)*	Cervical cytokine continuous (13)	Cervical cytokine truncated (12)	Endometrial *C. trachomatis* infection binary (1)	Endometrial histology binary (1)	# of subjects
1	Observed**	Observed	Observed	Observed	Observed	127
2	Observed	Observed	Observed	Observed	-	92
3	Observed	Observed	Observed	- ***	Observed	3
4	Observed	Observed	Observed	-	-	18
# of subjects	240	240	240	219	130	240

* The number in parentheses is number of measured variables in each data modality.
** "Observed" indicates variables in the corresponding data modality were completely observed.
*** Dash line "-" indicates block-wise missing.

The levels of 96 cytokines in cervical sponge eluates were measured using multiplex assays. Cytokine levels below the lower limit of quantification (LLOQ) were set to 0. Cytokines with more than 80% missing values or value of 0 were filtered, and there are 56 cytokines after the initial filtering. Among the 56 remaining cytokines, 25 of them have complete data and 31 have element-wise missing data. Although we could impute the element-wise missing data with our method for those 31 cytokines, since there is no golden standard to evaluate the performance of the imputation, we exclude those 31 cytokines with element-wise missing data from the following analysis. The 25 cytokines with complete observations were divided into 2 data modalities: 13 cytokines with less than or equal to 20% values of 0 were treated as continuous variables, and the remaining 12 cytokines which contained more than 20% values of 0 were treated as truncated variables. All non-zero values of cytokines were log2 transformed.

Endometrial infection status of *C. trachomatis* was treated as binary variable and had missing values. Histological evaluation of endometritis using the endometrial biopsies were independently provided by three physicians. We scored "Endometritis Negative" as 0, "Endometritis Positive" as 1 and "insufficient for diagnosis" as missing. The final diagnosis score would be the consensus diagnosis among at least two of them, and otherwise would be treated as missing. There were 130 participants with final diagnosis score, of which 86 were endometritis negative and 44 were endometritis positive.

5.4.2 Imputation of endometrial *C. trachomatis* infection status

We imputed the missing endometrial *C. trachomatis* infection status leveraging the 16 completely observed continuous variables (including 13 cervical cytokines and three variables from vaginal and cervical microbiology). For the alternative method, we first trained the model by regressing endometrial *C. trachomatis*

infection status on those 16 completely observed continuous variables among the 219 subjects with observed endometrial *C. trachomatis* infection status by logistic regression. Using the estimated coefficients from the trained model, we then predicted the endometrial *C. trachomatis* infection status of the remaining 21 subjects with missing infection status. It is noted that when only the 16 completely observed continuous variables and the infection status variable were included in the analysis, there was only one missing pattern. Thus, the MBI method reduced to the standard complete case analysis method.

We used the 219 subjects with observed infection status (31% had positive infection and 69% had negative infection) to compare the performance of imputation between our method and the alternative method. We randomly split the 219 subjects into a 80% training (n=175) set and a 20% testing set (n=44), and repeated this process 500 times. Each time, we assumed that the infection status for subjects in the testing set was missing, and impute the infection status using the training set with our method and the alternative method respectively. We then compared the imputed infection status with the true status in the testing set to obtain the sensitivity, specificity, overall accuracy and Kendall's τ for each method. The sensitivity is defined as the percentage of true "Positive Infections" and the specificity is defined as the percentage of true "Negative Infections". The overall accuracy is defined as the percentage of both true positive and true negative infection status. Kendall's τ is a similarity score between the imputed status and the true infection status.

The mean and standard error of sensitivity, specificity, overall accuracy and Kendall's τ estimated using our and alternative method over 500 splitting process are demonstrated in boxplots (Figure 5.13) and listed in Table 5.2. Our method has a clear advantage on the estimation of sensitivity, specificity, overall accuracy and Kendall's τ over the alternative method. Our method achieved an average sensitivity and specificity of 52% and 92% respectively, compared to 48% and 88% by the alternative method. The overall accuracy of our method (79%) also beat the alternative method (75%). In addition, our method had a much higher similarity to the truth with Kendall's τ of 0.49, compared to the alternative method with Kendall's τ of 0.39.

Table 5.2: Imputed endometrial *C. trachomatis* infection status result (average of 500 splitting process, mean and SE)

	Sensitivity	Specificity	Overall accuracy	Kendall's τ
Our method	52.31% (0.56%)	92.05% (0.19%)	79.45% (0.25%)	0.49 (0.01)
Alternative	48.83% (0.6%)	87.58% (0.27%)	75.41% (0.26%)	0.39 (0.01)

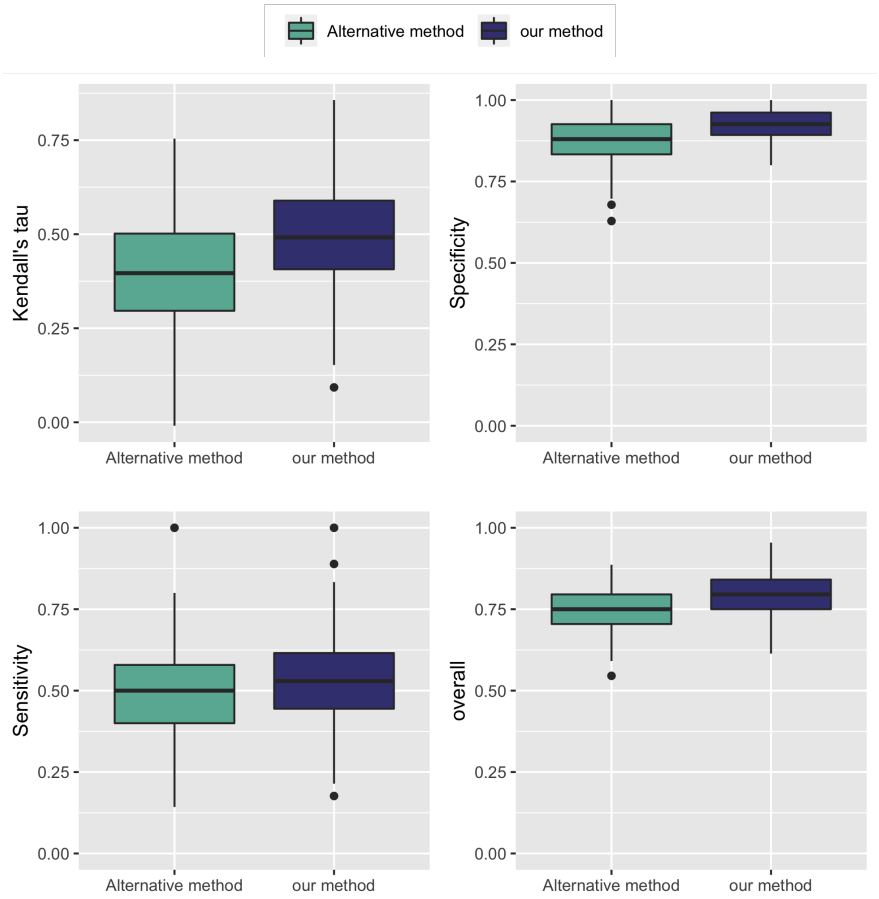


Figure 5.13: Boxplots for imputing endometrial *C. trachomatis* infection status over 500 splitting processes using our and the alternative method

5.4.3 Imputation of final histological diagnosis

We next imputed the missing final diagnosis scores leveraging the 16 completely observed continuous variables. For the alternative method, we first trained the model by regressing the final diagnosis score on those 16 completely observed continuous variables among the 130 subjects with consensus final diagnosis by logistic regression. Using the estimated coefficients from the trained model, we then predicted the diagnosis score of the remaining 110 subjects.

We used the 130 subjects with consensus final diagnosis (34% had positive endometritis and 66% had negative endometritis) to compare the performance of imputation between our method and the alternative method. We randomly split the 130 subjects into a 80% training (n=104) set and a 20% testing set (n=26), and repeated this process 500 times. Each time, we assumed that the diagnosis score for subjects in the testing set was missing, and impute the score using the training set with our method and the alternative method respectively. We then compared the imputed scores with the true scores in the testing set to obtain the sensitivity, specificity, overall accuracy and Kendall's τ for each method. The sensitivity is defined as the percentage of true "Endometritis Positive" and the specificity is defined as the percentage of true "Endometritis Negative". The overall accuracy is defined as the percentage of both true positive and true negative diagnosis. Kendall's τ is a similarity score between the imputed result and the true diagnosis result.

The mean and standard error of sensitivity, specificity, overall accuracy and Kendall's τ estimated using our and alternative methods over 500 splitting process are demonstrated in boxplots (Figure 5.14) and listed in Table 5.3. Our method has a clear advantage on the estimation of sensitivity, specificity, overall accuracy and Kendall's τ over the alternative method. Our method achieved an average sensitivity and specificity of 62% and 83% respectively, compared to 47% and 77% by the alternative method. The overall accuracy of our method (75%) also beat the alternative method (67%). In addition, our method had a much higher similarity to the truth with Kendall's τ of 0.45, compared to the alternative method with Kendall's τ of 0.25.

Table 5.3: Imputed diagnosis result (average of 500 splitting process, mean and SE)

	Sensitivity	Specificity	Overall accuracy	Kendall's τ
Our method	65.66% (0.67%)	83.22% (0.39%)	76.95% (0.32%)	0.49 (0.01)
Alternative	47.12% (0.74%)	77.49% (0.48%)	66.88% (0.39%)	0.25 (0.01)

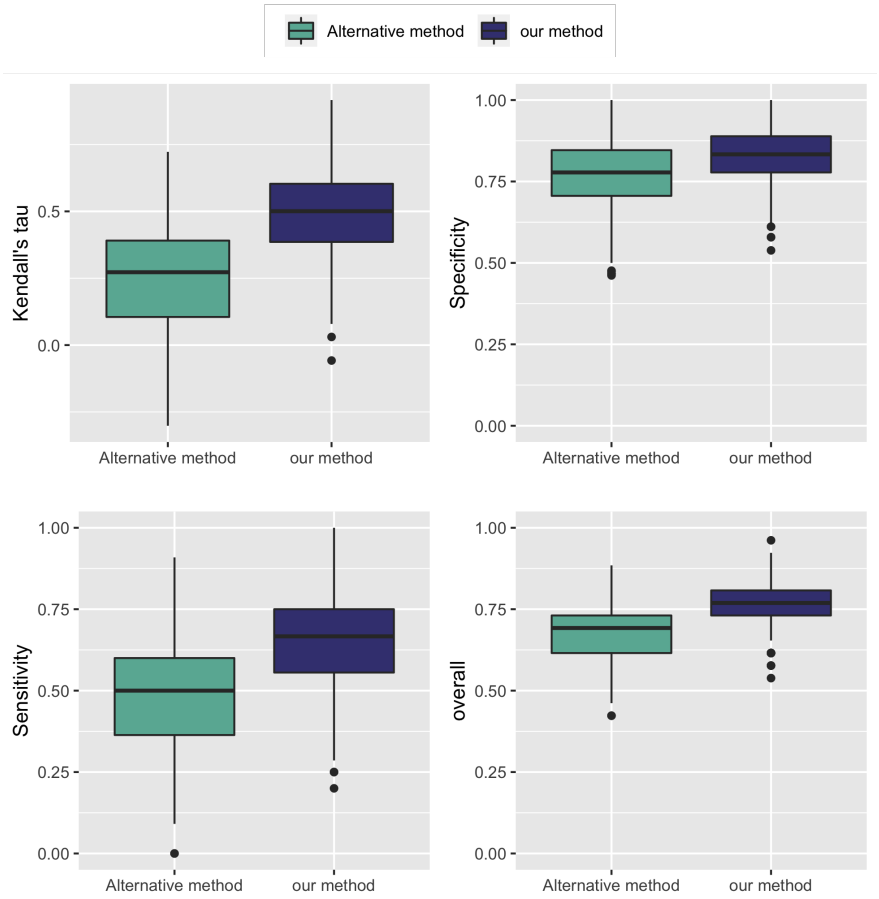


Figure 5.14: Boxplots for imputing diagnosis result over 500 splitting processes using our and the alternative method

5.4.4 Imputation of truncated cytokine variables

We finally compared the performance of our method and the alternative method for imputing the truncated cytokine variables. We random split the 240 subjects into a 80% training set (n=192) and a 20% testing set (n=48), and repeat this process 500 times. Each time, we assumed that the values of these 12 truncated cytokine variables in the test set were all missing, and imputed the score using the training set with our method and the alternative method. For variable j, let $\boldsymbol{X}_{[\text{test},j]}$ denote the truth of the values in the testing set, $\widehat{\boldsymbol{X}}_{[\text{test},j]}$ denote the estimate given by our method and $\widetilde{\boldsymbol{X}}_{[\text{test},j]}$ denote the estimate given by the alternative method. We then calculated the similarity score, measured by Kendall's tau as $\widehat{\tau}_j$ between $\widehat{\boldsymbol{X}}_{[\text{test},j]}$ and $\boldsymbol{X}_{[\text{test},j]}$ and $\widetilde{\tau}_j$ between $\widetilde{\boldsymbol{X}}_{[\text{test},j]}$ and $\boldsymbol{X}_{[\text{test},j]}$.

Let $\widehat{\boldsymbol{\tau}} = (\widehat{\tau}_1, \widehat{\tau}_2, \ldots, \widehat{\tau}_{12})$ and $\widetilde{\boldsymbol{\tau}} = (\widetilde{\tau}_1, \widetilde{\tau}_2, \ldots, \widetilde{\tau}_{12})$. Finally, we calculated the Euclidean norm of $\widehat{\boldsymbol{\tau}}$ and $\widetilde{\boldsymbol{\tau}}$ and use the Euclidean norm to measure how well each method performed when imputing the truncated cytokine variables. For the alternative method, since the truncated variables targeted for imputation had zero inflation, we employed the Gamma Hurdle Model (GHM) to impute the truncated element-wise missing variables.

A boxplot of $\|\widehat{\boldsymbol{\tau}}\|_E$ and $\|\widetilde{\boldsymbol{\tau}}\|_E$ over 500 splitting process using two methods is presented in Figure 5.15. $\|\widehat{\boldsymbol{\tau}}\|_E$ has a mean of 2.07 with standard error of 0.58 and $\|\widetilde{\boldsymbol{\tau}}\|_E$ has a mean of 1.77 with standard error of 0.01 over the 500 splitting process. We could see from Figure 5.15 that our method have a clear advantage when imputing the truncated variables compared to the alternative method.

5.5 Discussion

In this chapter, we proposed a novel method that can perform imputation for both block-wise and element-wise missings in multimodal data. Our method can handle various data types, including continuous, binary, 3-level ordinal and truncated data. In the simulation experiments, the method presents superior performance regardless of the missing mechanism (MCAR, MAR and MNAR) compared to other popular methods in the current literature under both low- and high-dimensional settings. We applied the method to the multimodal data collect from a *Chlamydia trachomatis* genital tract infection study to impute patients' missing histological diagnosis results, endometrial *C. trachomatis* infection status and missing truncated cytokine variables.

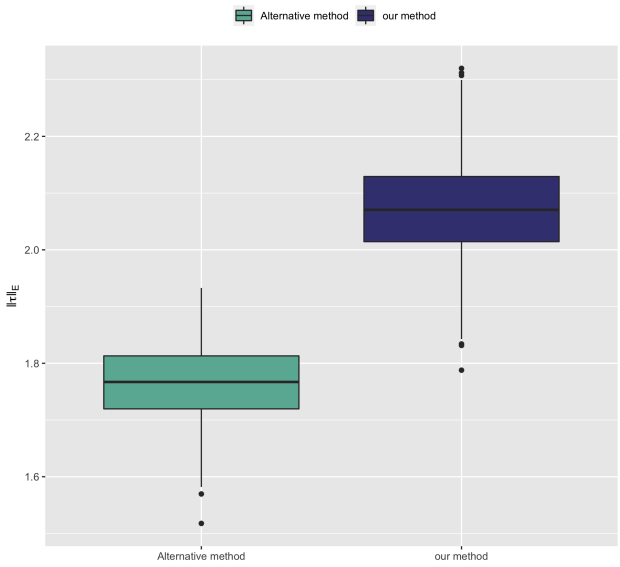


Figure 5.15: A boxplot of $\|\widehat{\boldsymbol{\tau}}\|_E$ and $\|\widetilde{\boldsymbol{\tau}}\|_E$ for imputing truncated variables over 500 splitting processes using our and the alternative method

We want to point out that our method of imputing missing data based on the observed data highly rely on the correlation between the observed data and the missing data. As demonstrated in the simulation experiments, our method have a better performance when the true latent correlation between the observed data and missing data is higher. Moreover, the missing mechanism does not affect the performance of our method as long as we could obtain accurate moment estimator from the observed data.

We also want to mentioned that our methods focus on imputing the block-wise missing structure brought by mulit-modal data. However, if there is element-wise missingness within each data modality, our method can also be used for the element-wise imputation. In this case, the missing block will only contain one variable which is the variable with element-wise missingness.

BIBLIOGRAPHY

Alter, Orly, Patrick O Brown and David Botstein. 2003. "Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms." *Proceedings of the National Academy of Sciences* 100(6):3351–3356.

Amar, David, Hershel Safer and Ron Shamir. 2013. "Dissection of regulatory networks that are altered in disease via differential co-expression." *PLoS Comput Biol* 9(3):e1002955.

Andrew, Dean W, Melanie Cochrane, Justin H Schripsema, Kyle H Ramsey, Samantha J Dando, Connor P O'Meara, Peter Timms and Kenneth W Beagley. 2013. "The duration of Chlamydia muridarum genital tract infection and associated chronic pathological changes are reduced in IL-17 knockout mice but protection is not increased further by immunization." *PloS one* 8(9):e76664.

Baird, J Kevin. 2013. "Evidence and implications of mortality associated with acute Plasmodium vivax malaria." *Clinical microbiology reviews* 26(1):36–57.

Bhattacharya, Palash, Muthusamy Thiruppathi, Hatem A Elshabrawy, Khaled Alharshawi, Prabhakaran Kumar and Bellur S Prabhakar. 2015. "GM-CSF: An immune modulatory cytokine that can suppress autoimmunity." *Cytokine* 75(2):261–271.

Bickel, Peter J and Elizaveta Levina. 2008. "Covariance regularization by thresholding." *The Annals of Statistics* 36(6):2577–2604.

Bureau, Alexandre, Stephen Shiboski and James P Hughes. 2003. "Applications of continuous time hidden Markov models to the study of misclassified disease outcomes." *Statistics in Medicine* 22(3):441–462.

Cai, Jian-Feng, Emmanuel J Candès and Zuowei Shen. 2010. "A singular value thresholding algorithm for matrix completion." *SIAM Journal on optimization* 20(4):1956–1982.

Cai, Tianxi, T Tony Cai and Anru Zhang. 2016. "Structured matrix completion with applications to genomic data integration." *Journal of the American Statistical Association* 111(514):621–633.

Candès, Emmanuel J and Benjamin Recht. 2009. "Exact matrix completion via convex optimization." *Foundations of Computational mathematics* 9(6):717.

Candès, Emmanuel J and Terence Tao. 2010. "The power of convex relaxation: Near-optimal matrix completion." *IEEE Transactions on Information Theory* 56(5):2053–2080.

Choi, YounJeong and Christina Kendziorski. 2009. "Statistical methods for gene set co-expression analysis." *Bioinformatics* 25(21):2780–2786.

Chu, Cindy S and Nicholas J White. 2016. "Management of relapsing Plasmodium vivax malaria." *Expert review of anti-infective therapy* 14(10):885–900.

Darville, Toni, Hannah L Albritton, Wujuan Zhong, Li Dong, Catherine M O'Connell, Taylor B Poston, Alison J Quayle, Nilu Goonetilleke, Harold C Wiesenfeld, Sharon L Hillier et al. 2019. "Anti-chlamydia IgG and IgA are insufficient to prevent endometrial chlamydia infection in women, and increased anti-chlamydia IgG is associated with enhanced risk for incident infection." *American Journal of Reproductive Immunology* 81(5):e13103.

d'Aspremont, Alexandre, Onureena Banerjee and Laurent El Ghaoui. 2008. "First-order methods for sparse covariance selection." *SIAM Journal on Matrix Analysis and Applications* 30(1):56–66.

Dini, Saber, Nicholas M Douglas, Jeanne Rini Poespoprodjo, Enny Kenangalem, Paulus Sugiarto, Ian D Plumb, Ric N Price and Julie A Simpson. 2020. "The risk of morbidity and mortality following recurrent malaria in Papua, Indonesia: a retrospective cohort study." *BMC medicine* 18(1):1–12.

Dinse, Gregg E. 1982. "Nonparametric estimation for partially-complete time and type of failure data." *Biometrics* pp. 417–431.

Edlund, Christopher K., Malin Anker, Fredrick R. Schumacher, W. James Gauderman and David V. Conti. 2016.
URL: *http://prioritypruner.sourceforge.net/index.html*

Effraimidis, Georgios and Christian M Dahl. 2014. "Nonparametric estimation of cumulative incidence functions for competing risks data with missing cause of failure." *Statistics & Probability Letters* 89:1–7.

Fan, Jianqing, Han Liu, Weichen Wang and Ziwei Zhu. 2018. "Heterogeneity adjustment with applications to graphical model inference." *Electronic journal of statistics* 12(2):3908.

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

Fan, Jianqing and Jinchi Lv. 2011. "Nonconcave penalized likelihood with NP-dimensionality." *IEEE Transactions on Information Theory* 57(8):5467–5484.

Feng, Qing, Meilei Jiang, Jan Hannig and JS Marron. 2018. "Angle-based joint and individual variation explained." *Journal of multivariate analysis* 166:241–265.

Ferreira, Marcelo U, Tais Nobrega de Sousa, Gabriel W Rangel, Igor C Johansen, Rodrigo M Corder, Simone Ladeia-Andrade and José Pedro Gil. 2020. "Monitoring Plasmodium vivax resistance to antimalarials: Persisting challenges and future directions." *International Journal for Parasitology: Drugs and Drug Resistance* 15:9.

Friedrich, Lindsey R, Jean Popovici, Saorin Kim, Lek Dysoley, Peter A Zimmerman, Didier Menard and David Serre. 2016. "Complexity of Infection and Genetic Diversity in Cambodian Plasmodium vivax." *PLoS Neglected Tropical Diseases* 10(3):e0004526.

Goetghebeur, Els and Louise Ryan. 1995. "Analysis of competing risks survival data when some failure types are missing." *Biometrika* 82(4):821–833.

Gouskova, Natalia A, Feng-Chang Lin and Jason P Fine. 2017. "Nonparametric analysis of competing risks data with event category missing at random." *Biometrics* 73(1):104–113.

Gross, David. 2011. "Recovering low-rank matrices from few coefficients in any basis." *IEEE Transactions on Information Theory* 57(3):1548–1566.

Ha, Min Jin, Veerabhadran Baladandayuthapani and Kim-Anh Do. 2015. "DINGO: differential network analysis in genomics." *Bioinformatics* 31(21):3413–3420.

Han, Fang, Tuo Zhao and Han Liu. 2013. "CODA: High dimensional copula discriminant analysis." *Journal of Machine Learning Research* 14(Feb):629–671.

Hastie, Trevor, Rahul Mazumder, Jason D Lee and Reza Zadeh. 2015. "Matrix completion and low-rank SVD via fast alternating least squares." *The Journal of Machine Learning Research* 16(1):3367–3402.

Hastie, Trevor, Robert Tibshirani and Martin Wainwright. 2019. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.

Hathaway, Nicholas J, Christian M Parobek, Jonathan J Juliano and Jeffrey A Bailey. 2018. "SeekDeep: single-base resolution de novo clustering for amplicon deep sequencing." *Nucleic Acids Research* 46(4):e21.

Howes, Rosalind E, Katherine E Battle, Kamini N Mendis, David L Smith, Richard E Cibulskis, J Kevin Baird and Simon I Hay. 2016. "Global epidemiology of Plasmodium vivax." *The American Journal of Tropical Medicine and Hygiene* 95(6_Suppl):15–34.

Huang, Jianhua Z, Naiping Liu, Mohsen Pourahmadi and Linxu Liu. 2006. "Covariance matrix selection and estimation via penalised normal likelihood." *Biometrika* 93(1):85–98.

Iwakura, Yoichiro, Harumichi Ishigame et al. 2006. "The IL-23/IL-17 axis in inflammation." *The Journal of clinical investigation* 116(5):1218–1222.

Juraska, Michal and Peter B Gilbert. 2016. "Mark-specific hazard ratio model with missing multivariate marks." *Lifetime data analysis* 22(4):606–625.

Kalbfleisch, John D and Ross L Prentice. 2002. *The statistical analysis of failure time data*. Vol. 360 John Wiley & Sons.

Keates, Sarah, Xinbing Han, Ciarán P Kelly and Andrew C Keates. 2007. "Macrophage-inflammatory protein-3α mediates epidermal growth factor receptor transactivation and ERK1/2 MAPK signaling in Caco-2 colonic epithelial cells via metalloproteinase-dependent release of amphiregulin." *The Journal of Immunology* 178(12):8013–8021.

Keshavan, Raghunandan H, Andrea Montanari and Sewoong Oh. 2010. "Matrix completion from a few entries." *IEEE transactions on information theory* 56(6):2980–2998.

Kiviat, NB, P Wolner-Hanssen, DA Eschenbach, JN Wasserheit, JA Paavonen, TA Bell, CW Critchlow, WE Stamm, DE Moore and KK Holmes. 1990. "Endometrial Histopathology in Patients With Culture-Proved Upper Genital Tract Infection and Laparoscopically Diagnosed Acute Salpingitis." *The American journal of surgical pathology* 14(2):167–175.

Langfelder, Peter and Steve Horvath. 2008. "WGCNA: an R package for weighted correlation network analysis." *BMC bioinformatics* 9(1):1–13.

Laurent, Monique. 2001. *Matrix completion problems*. Boston, MA: Springer US pp. 1311–1319.

Li, Gen, Irina Gaynanova et al. 2018. "A general framework for association analysis of heterogeneous data." *The Annals of Applied Statistics* 12(3):1700–1726.

Lijek, Rebeccah S, Jennifer D Helble, Andrew J Olive, Kyra W Seiger and Michael N Starnbach. 2018. "Pathology after Chlamydia trachomatis infection is driven by nonprotective immune cells that are distinct from protective populations." *Proceedings of the National Academy of Sciences* 115(9):2216–2221.

Lin, Danyu Y, Lee-Jen Wei and Zhiliang Ying. 1993. "Checking the Cox model with cumulative sums of martingale-based residuals." *Biometrika* 80(3):557–572.

Lin, Feng-Chang, Quefeng Li and Jessica T Lin. 2020. "Relapse or reinfection: Classification of malaria infection using transition likelihoods." *Biometrics* 76(4):1351–1363.

Lin, Jessica T, Jaymin C Patel, Oksana Kharabora, Jetsumon Sattabongkot, Sinuon Muth, Ratawan Ubalee, Anthony L Schuster, William O Rogers, Chansuda Wongsrichanalai and Jonathan J Juliano. 2013. "Plasmodium vivax isolates from Cambodia and Thailand show high genetic complexity and distinct patterns of P. vivax multidrug resistance gene 1 (pvmdr1) polymorphisms." *The American Journal of Tropical Medicine and Hygiene* 88(6):1116–1123.

Lin, Jessica T, Nicholas J Hathaway, David L Saunders, Chanthap Lon, Sujata Balasubramanian, Oksana Kharabora, Panita Gosi, Sabaithip Sriwichai, Laurel Kartchner, Char Meng Chuor et al. 2015. "Using amplicon deep sequencing to detect genetic signatures of Plasmodium vivax relapse." *The Journal of Infectious Diseases* 212(6):999–1008.

Liu, Han, John Lafferty and Larry Wasserman. 2009. "The nonparanormal: Semiparametric estimation of high dimensional undirected graphs." *Journal of Machine Learning Research* 10(10).

Liu, Yutong, Toni Darville, Xiaojing Zheng and Quefeng Li. 2022. "Decomposition of variation of mixed variables by a latent mixed Gaussian copula model." *Biometrics* pp. 1–14.

Lock, Eric F, Katherine A Hoadley, James Stephen Marron and Andrew B Nobel. 2013. "Joint and individual variation explained (JIVE) for integrated analysis of multiple data types." *The annals of applied statistics* 7(1):523.

Löfstedt, Tommy and Johan Trygg. 2011. "OnPLS—a novel multiblock method for the modelling of predictive and orthogonal variation." *Journal of Chemometrics* 25(8):441–455.

Lon, Chanthap, Jessica E Manning, Pattaraporn Vanachayangkul, Mary So, Darapiseth Sea, Youry Se, Panita Gosi, Charlotte Lanteri, Suwanna Chaorattanakawee, Sabaithip Sriwichai et al. 2014. "Efficacy of two versus three-day regimens of dihydroartemisinin-piperaquine for uncomplicated malaria in military personnel in northern Cambodia: an open-label randomized trial." *PLoS One* 9(3):e93138.

Lu, Kaifeng and Anastasios A Tsiatis. 2001. "Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure." *Biometrics* 57(4):1191–1197.

Marino, Julieta, Verónica A Furmento, Elsa Zotta and Leonor P Roguin. 2009. "Peritumoral administration of granulocyte colony-stimulating factor induces an apoptotic response on a murine mammary adenocarcinoma." *Cancer biology & therapy* 8(18):1737–1743.

Mazumder, Rahul, Trevor Hastie and Robert Tibshirani. 2010. "Spectral regularization algorithms for learning large incomplete matrices." *The Journal of Machine Learning Research* 11:2287–2322.

McCullagh, P and JA Nelder. 1989. *Generalized linear models*. Chapman and Hill.

Nastase, Madalina Viviana, Jinyang Zeng-Brouwers, Janet Beckmann, Claudia Tredup, Urs Christen, Heinfried H Radeke, Malgorzata Wygrecka and Liliana Schaefer. 2018. "Biglycan, a novel trigger of Th1 and Th17 cell recruitment into the kidney." *Matrix Biology* 68:293–317.

Neafsey, Daniel E, Kevin Galinsky, Rays HY Jiang, Lauren Young, Sean M Sykes, Sakina Saif, Sharvari Gujja, Jonathan M Goldberg, Sarah Young, Qiandong Zeng et al. 2012. "The malaria parasite Plasmodium vivax exhibits greater genetic diversity than Plasmodium falciparum." *Nature Genetics* 44(9):1046–1050.

Parikh, Neal and Stephen Boyd. 2014. "Proximal algorithms." *Foundations and Trends in Optimization* 1(3):127–239.

Parobek, Christian M, Jeffrey A Bailey, Nicholas J Hathaway, Duong Socheat, William O Rogers and Jonathan J Juliano. 2014. "Differing patterns of selection and geospatial genetic diversity within two leading Plasmodium vivax candidate vaccine antigens." *PLoS Negl Trop Dis* 8(4):e2796.

Plackett, Robin L. 1954. "A reduction formula for normal multivariate integrals." *Biometrika* 41(3/4):351–360.

Ponnapalli, Sri Priya, Michael A Saunders, Charles F Van Loan and Orly Alter. 2011. "A higher-order generalized singular value decomposition for comparison of global mRNA expression from multiple organisms." *PloS one* 6(12):e28072.

Poston, Taylor B, De'Ashia E Lee, Toni Darville, Wujuan Zhong, Li Dong, Catherine M O'Connell, Harold C Wiesenfeld, Sharon L Hillier, Gregory D Sempowski and Xiaojing Zheng. 2019. "Cervical cytokines associated with Chlamydia trachomatis susceptibility and protection." *The Journal of infectious diseases* 220(2):330–339.

Qin, Jing. 1998. "Inferences for case-control and semiparametric two-sample density ratio models." *Biometrika* 85(3):619–630.

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

Rahmatallah, Yasir, Frank Emmert-Streib and Galina Glazko. 2014. "Gene Sets Net Correlations Analysis (GSNCA): a multivariate differential coexpression test for gene sets." *Bioinformatics* 30(3):360–368.

Rangel-Moreno, Javier, Damian M Carragher, Maria de la Luz Garcia-Hernandez, Ji Young Hwang, Kim Kusser, Louise Hartson, Jay K Kolls, Shabaana A Khader and Troy D Randall. 2011. "The development of inducible bronchus-associated lymphoid tissue depends on IL-17." *Nature immunology* 12(7):639–646.

Rennie, Jasson DM and Nathan Srebro. 2005. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on Machine learning*. pp. 713–719.

Robinson, Leanne J, Rahel Wampfler, Inoni Betuela, Stephan Karl, Michael T White, Connie SN Li Wai Suen, Natalie E Hofmann, Benson Kinboro, Andreea Waltmann, Jessica Brewster et al. 2015. "Strategies for understanding and reducing the Plasmodium vivax and Plasmodium ovale hypnozoite reservoir in Papua New Guinean children: a randomised placebo-controlled trial and mathematical model." *PLoS Med* 12(10):e1001891.

Rothman, Adam J, Peter J Bickel, Elizaveta Levina and Ji Zhu. 2008. "Sparse permutation invariant covariance estimation." *Electronic Journal of Statistics* 2:494–515.

Rubin, Donald B. 1976. "Inference and missing data." *Biometrika* 63(3):581–592.

Russell, Ali N, Xiaojing Zheng, Catherine M O'Connell, Brandie D Taylor, Harold C Wiesenfeld, Sharon L Hillier, Wujuan Zhong and Toni Darville. 2016. "Analysis of factors driving incident and ascending infection and the role of serum antibody in Chlamydia trachomatis genital tract infection." *The Journal of infectious diseases* 213(4):523–531.

Schwarz, Gideon et al. 1978. "Estimating the dimension of a model." *The Annals of Statistics* 6(2):461–464.

Shabalin, Andrey A. 2012. "Matrix eQTL: ultra fast eQTL analysis via large matrix operations." *Bioinformatics* 28(10):1353–1358.

Shin, Soon Young, Jishin Lee Da Hyun Lee, Chan Choi, Ji-Young Kim, Jeong-Seok Nam, Yoongho Lim and Young Han Lee. 2017. "CC motif chemokine receptor 1 (CCR1) is a target of the EGF-AKT-mTOR-STAT3 signaling axis in breast cancer cells." *Oncotarget* 8(55):94591.

Shu, Hai, Xiao Wang and Hongtu Zhu. 2020. "D-CCA: A decomposition-based canonical correlation analysis for high-dimensional datasets." *Journal of the American Statistical Association* 115(529):292–306.

Starkey, MR, DH Nguyen, AT Essilfie, RY Kim, LM Hatchwell, AM Collison, H Yagita, PS Foster, JC Horvat, J Mattes et al. 2014. "Tumor necrosis factor-related apoptosis-inducing ligand translates neonatal respiratory infection into chronic lung disease." *Mucosal immunology* 7(3):478–488.

Sun, Yanqing and Peter B Gilbert. 2012. "Estimation of stratified mark-specific proportional hazards models with missing marks." *Scandinavian Journal of Statistics* 39(1):34–52.

Taylor, Aimee R, James A Watson, Cindy S Chu, Kanokpich Puaprasert, Jureeporn Duanguppama, Nicholas PJ Day, Francois Nosten, Daniel E Neafsey, Caroline O Buckee, Mallika Imwong et al. 2019. "Resolving the cause of recurrent Plasmodium vivax malaria probabilistically." *Nature Communications* 10(1):1–11.

Tesson, Bruno M, Rainer Breitling and Ritsert C Jansen. 2010. "DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules." *BMC bioinformatics* 11(1):1–9.

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

van Dam, Sipko, Urmo Vosa, Adriaan van der Graaf, Lude Franke and Joao Pedro de Magalhaes. 2018. "Gene co-expression analysis for functional classification and gene–disease predictions." *Briefings in bioinformatics* 19(4):575–592.

Watson, Michael. 2006. "CoXpress: differential co-expression in gene expression data." *BMC bioinformatics* 7(1):1–12.

WHO. 2019*a*. *World Health Statistics 2019. Monitoring health for the SDGs.* World Health Organization.

WHO. 2019*b*. *World Malaria Report 2019*. World Health Organization.

Xiang, Shuo, Lei Yuan, Wei Fan, Yalin Wang, Paul M Thompson, Jieping Ye, Alzheimer's Disease Neuroimaging Initiative et al. 2014. "Bi-level multi-source learning for heterogeneous block-wise missing data." *NeuroImage* 102:192–206.

Xue, Fei and Annie Qu. 2020. "Integrating Multisource Block-Wise Missing Data in Model Selection." *Journal of the American Statistical Association* pp. 1–14.

Xue, Fei and Annie Qu. 2021. "Integrating multisource block-wise missing data in model selection." *Journal of the American Statistical Association* 116(536):1914–1927.

Yoon, Grace, Raymond J Carroll and Irina Gaynanova. 2020. "Sparse semiparametric canonical correlation analysis for data of mixed types." *Biometrika* 107(3):609–625.

Yu, Guan, Quefeng Li, Dinggang Shen and Yufeng Liu. 2020. "Optimal sparse linear prediction for block-missing multi-modality data without imputation." *Journal of the American Statistical Association* 115(531):1406–1419.

Yuan, Lei, Yalin Wang, Paul M Thompson, Vaibhav A Narayan, Jieping Ye, Alzheimer's Disease Neuroimaging Initiative et al. 2012. "Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data." *NeuroImage* 61(3):622–632.

Zhang, Yan-Qing, Nian-Sheng Tang and Annie Qu. 2020. "Imputed Factor Regression for High-dimensional Block-wise Missing Data." *Statistica Sinica* .

Zhao, Tuo, Kathryn Roeder and Han Liu. 2014. "Positive semidefinite rank-based correlation matrix estimation with application to semiparametric graph estimation." *Journal of Computational and Graphical Statistics* 23(4):895–922.

Zhong, Wujuan, Li Dong, Taylor B Poston, Toni Darville, Cassandra N Spracklen, Di Wu, Karen L Mohlke, Yun Li, Quefeng Li and Xiaojing Zheng. 2020. "Inferring regulatory networks from mixed observational data using directed acyclic graphs." *Frontiers in genetics* 11:8.

Zhong, Wujuan, Toni Darville, Xiaojing Zheng, Jason Fine and Yun Li. 2019. "Generalized multi-SNP mediation intersection-union test." *Biometrics* .

Zhou, Doudou, Tianxi Cai and Junwei Lu. 2021. "Multi-source Learning via Completion of Block-wise Overlapping Noisy Matrices." *arXiv preprint arXiv:2105.10360* .

Zhou, Guoxu, Andrzej Cichocki, Yu Zhang and Danilo P Mandic. 2015. "Group component analysis for multiblock data: Common and individual feature extraction." *IEEE transactions on neural networks and learning systems* 27(11):2426–2439.

Zhu, Huichen, Gen Li and Eric F Lock. 2020. "Generalized integrative principal component analysis for multi-type data with block-wise missing structure." *Biostatistics* 21(2):302–318.

Zou, Hui and Trevor Hastie. 2005. "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society: Series B (statistical methodology)* 67(2):301–320.