Received: 21 July 2021 | Accepted: 3 March 2022

DOI: 10.1111/biom.13660

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Decomposition of variation of mixed variables by a latent mixed Gaussian copula model

Yutong Liu[1] | **Toni Darville**[2] | **Xiaojing Zheng**[1,2] | **Quefeng Li**[1]

[1]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

[2]Department of Pediatrics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

Correspondence
Quefeng Li, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.
Email: quefeng@email.unc.edu

Funding information
National Institutes of Health, Grant/Award Numbers: U19AI084024, U19AI144181; National Institute on Aging, Grant/Award Number: R01AG073259

Abstract

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 (LMGC) model that can quantify the correlations among binary, ordinal, 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 also 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.

KEYWORDS

high-dimensional matrix estimation, Kendall's τ, latent Gaussian copula model, variation decomposition

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 [SNP]), 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.

In previous studies, people aim to find the correlations of differentially expressed genes across groups of subjects with different phenotypes. A lot of methods have been

developed for such a differential gene coexpression analysis (van Dam *et al.*, 2018), which aims to reveal regulatory genes corresponding to different phenotypes. Watson (2006) proposed a hierarchical clustering method to identify groups of genes that are differentially expressed under different phenotypes. Choi and Kendziorski (2009) introduced a testing method to test whether some given gene sets are differentially expressed between different groups. Other related works include Tesson *et al.* (2010), Amar *et al.* (2013), and Rahmatallah *et al.* (2014). All these methods focus on finding the group-specific structure of different phenotypes but do not consider the shared information across groups. To account for the shared information, Alter *et al.* (2003) proposed to use the generalized singular value decomposition to decompose the gene expression data as a sum of the effects that are shared for both data sets and the effects that are unique for each data set. However, their method only applies to two data sets. Ponnapalli *et al.* (2011) further extended it to deal with three or more data sets. However, all of these methods only apply to continuous gene expression data and are hard to be extended to multi-omics data with mixed types of variables.

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 *et al.*, 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. Quan *et al.* (2018) and Feng and Ning (2019) extended the method to measure correlations among ordinal, binary, and continuous variables. Yoon *et al.* (2020) further extended it to incorporate truncated variables. However, all these works only considered one population. They did not consider decomposing the variation into common and group-specific components.

There are several works tackling the problem of variation decomposition. The problem 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 *et al.*, 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 angle-based joint and individual variation explained (AJIVE) method, where score subspaces were used to ensure an identifiable decomposition. Zhou *et al.* (2015) proposed common orthogonal basis extraction (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 *et al.* (2020) proposed D-CCA, a decomposition-based CCA 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, ordinal, or truncated variables. To carry out integrative analysis for data of different types and decompose the data into shared and individual structures, Li and Gaynanova (2018) developed the generalized association study (GAS), which uses the log-likelihood function to integrate different variables that follow exponential family distributions. Zhu *et al.* (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 blockwise 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.

To find the common and group-specific variation of mixed variables for multiple groups, we propose a two-step method. First, using the latent mixed Gaussian copula (LMGC) model, we measure correlations among binary, ordinal, continuous, and truncated variables under a unified framework. Compared with the existing works (Fan *et al.*, 2017; Quan *et al.*, 2018; Feng and Ning, 2019; Yoon *et al.*, 2020), we derive the bridge function for ordinal and truncated variables and prove that it is invertible. As pointed out by one reviewer, we acknowledge that such a bridge function was concurrently found by an independent work (Huang *et al.*, 2021). Using the LMGC model, we obtain 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 view the decomposition step as a denoising process that after removing the shared variation, the low-rank group-specific components can give a clearer view of the differences between groups.

The rest of this paper is organized as follows. In Section 2, we describe the formulation and solution of our proposed method in details. In Section 3, we carry out extensive simulation studies to compare our method with some competitive methods. In Section 4, we apply our method to a *C. trachomatis* genital tract infection study to demonstrate how it can be used to find useful biomarkers that differentiate subtypes of patients.

2 | METHODOLOGY

We consider two groups of subjects. For the gth 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, ordinal, 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

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

where $\mathbb{C}$, $\mathbb{B}$, $\mathbb{T}$, and $\mathbb{O}$ are the index sets for continuous, binary, truncated, and ordinal variables, respectively, and $\{C_{g,j}\}_{j \in \mathbb{B}}$, $\{D_{g,j}\}_{j \in \mathbb{T}}$, and $\{C_{g,j,l}\}_{j \in \mathbb{O}, 1 \le l \le L_j - 1}$ are the corresponding cutoffs. We assume that the latent $\boldsymbol{Y}_g$ follows a Gaussian copula model proposed by Liu *et al.* (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 $(f_{g,1}(Y_{g,1}), \ldots, f_{g,p}(Y_{g,p}))^T \sim N(\boldsymbol{0}, \boldsymbol{R}_g)$, where $\boldsymbol{R}_g$ is a correlation matrix. We call (1) as the LMGC model for mixed data. In the existing literature, Fan *et al.* (2017) studied the LMGC model for continuous and binary variables only. Yoon *et al.* (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 CCA. 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 their 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

$$\boldsymbol{f}_g(\boldsymbol{Y}_g) = \boldsymbol{\Lambda}_g \boldsymbol{F}_g + \boldsymbol{U}, \tag{2}$$

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 shared component, which is assumed to be uncorrelated with $\boldsymbol{F}_g$. To avoid the identifiability issue, we adopt the standard 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 (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 (2) that

$$\boldsymbol{R}_g = \boldsymbol{\Sigma}_g + \boldsymbol{\Sigma}_U = \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Sigma}_U. \tag{3}$$

Motivated by the factor model in (2), we assume the group-specific variation $\boldsymbol{\Sigma}_g$ is low rank, that is, r_g is small. In addition, we assume that the common variation $\boldsymbol{\Sigma}_U$ is sparse. In that way, we treat $\boldsymbol{\Sigma}_U$ as the background noise and (3) as a denosing process. That is, after removing the common $\boldsymbol{\Sigma}_U$, the group-specific variation are reflected by $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g'$. Since the signal in the background noise is usually small, we assume that $\boldsymbol{\Sigma}_U$ is sparse, meaning that variables are mostly not highly correlated in the background noise.

Next, we show in Proposition 1 that Equation (3) is a well-defined problem, in the sense that even if $\boldsymbol{f}_g(\boldsymbol{Y}_g)$ has different decomposition in Equation (2), the decomposition of $\boldsymbol{R}_g$ in Equation (3) is still unique. Furthermore, we demonstrate in Section 2.2 that the decomposition in Equation (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 $\boldsymbol{\Sigma}_U = \widetilde{\boldsymbol{\Sigma}}_U$.*

Finally, we remark that using factor models for variation decomposition has also been considered in some other works. De Vito *et al.* (2019) proposed a multistudy factor model that decomposes observed variables as common factors shared across multiple studies and study-specific factors. Ha *et al.* (2015) proposed a Gaussian Graphical Model based method to decompose the variation among multiple variables as a shared global component and a group-specific component, where they assume the group-specific component is driven by latent factors. However, these works only apply to continuous variables and require all variables to be observable. On the contrary, our method applies to mixed types of variables and allows $\boldsymbol{f}_g(\boldsymbol{Y}_g)$ to be latent.

2.1 | Rank-based latent correlation matrix estimator

In this section, we propose a rank-based estimator of $\boldsymbol{R}_g$ for mixed data. 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} being the covariance between the two variables.

Estimating the latent correlation matrix has been studied by Liu *et al.* (2009), Fan *et al.* (2017), Quan *et al.* (2018), and Yoon *et al.* (2020). 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 $\{(X_{ij}, X_{ik})\}_{i=1}^n$ 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

$$\widehat{\tau}_{jk} = \frac{2}{n(n-1)} \sum_{1 \le i < i' \le n} \text{sign}(X_{ij} - X_{i'j}) \, \text{sign}(X_{ik} - X_{i'k}). \tag{4}$$

Let $\tau_{jk} = \mathbb{E}(\widehat{\tau}_{jk})$ be the population Kendall's τ. Then, the latent correlation 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 formulas were derived in Liu *et al.* (2009), Fan *et al.* (2017), and Yoon *et al.* (2020).

Theorem 1. *((Liu* et al., *2009; Fan* et al., *2017; Yoon* et al., *2020))*

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

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

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

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

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

$$\boldsymbol{R}_{3b} = \begin{pmatrix} 1 & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 1 & -\frac{R_{jk}}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & -\frac{R_{jk}}{\sqrt{2}} & 1 \end{pmatrix}. \tag{5}$$

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

$$\boldsymbol{R}_{3c} = \begin{pmatrix} 1 & \frac{1}{\sqrt{2}} & \frac{R_{jk}}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 1 & R_{jk} \\ \frac{R_{jk}}{\sqrt{2}} & R_{jk} & 1 \end{pmatrix}. \tag{6}$$

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

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

$$\boldsymbol{R}_{4b} = \begin{pmatrix} 1 & R_{jk} & \frac{1}{\sqrt{2}} & \frac{R_{jk}}{\sqrt{2}} \\ R_{jk} & 1 & \frac{R_{jk}}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{R_{jk}}{\sqrt{2}} & 1 & R_{jk} \\ \frac{R_{jk}}{\sqrt{2}} & \frac{1}{\sqrt{2}} & R_{jk} & 1 \end{pmatrix}. \tag{7}$$

Liu *et al.* (2009), Fan *et al.* (2017), and Yoon *et al.* (2020) proved 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 $\widehat{R}_{jk} = F_{jk}^{-1}(\widehat{\tau}_{jk})$, where F_{jk}^{-1} is the inverse of the bridge function. The inversion can be done by solving $F_{jk}(x) = \widehat{\tau}_{jk}$ using the Newton–Raphson algorithm. For a binary or truncated variable, $\Delta_k = f_k(C_k)$ is unknown. We follow Fan *et al.* (2017) and use the plug-in estimator $\widehat{\Delta}_k = \Phi^{-1}\{\sum_{i=1}^n I(X_{ik} \neq 0)/n\}$ to estimate it.

Next, we derive the formulas of bridge functions for the latent correlations between three-level ordinal variables and continuous/binary/truncated/three-level ordinal variables. We also prove that all these bridge functions are monotone so that they are invertible.

Theorem 2. *The following results hold.*

(a) For $j \in \mathbb{O}$ *and* $k \in \mathbb{C}$, $F_{jk}(R_{jk}; \Delta_{j1}, \Delta_{j2}) = 2\Phi_2(\Delta_{j2}, 0; \frac{R_{jk}}{\sqrt{2}}) - 2\Phi_2(\Delta_{j2}, 0; -\frac{R_{jk}}{\sqrt{2}}) - 2\Phi_3(\Delta_{j1}, \Delta_{j2}, 0; \boldsymbol{R}_{3d}) + 2\Phi_3(\Delta_{j2}, \Delta_{j1}, 0; \boldsymbol{R}_{3d})$, *where* $\Delta_{j1} = f_j(C_{j1}), \Delta_{j2} = f_j(C_{j2})$, *and*

$$\boldsymbol{R}_{3d} = \begin{pmatrix} 1 & 0 & -\frac{R_{jk}}{\sqrt{2}} \\ 0 & 1 & \frac{R_{jk}}{\sqrt{2}} \\ -\frac{R_{jk}}{\sqrt{2}} & \frac{R_{jk}}{\sqrt{2}} & 1 \end{pmatrix}. \tag{8}$$

(b) For $j \in \mathbb{O}$ *and* $k \in \mathbb{B}$, $F_{jk}(R_{jk}) = 2\Phi_2(\Delta_{j2}, \Delta_k; R_{jk}) - 2\Phi_1(\Delta_{j2})\Phi_1(\Delta_k) - 2\Phi_1(\Delta_{j1})\Phi_2(\Delta_{j2}, \Delta_k; R_{jk}) + 2\Phi_1(\Delta_{j2})\Phi_2(\Delta_{j1}, \Delta_k; R_{jk})$, *where* $\Delta_{j1} = f_j(C_{j1}), \Delta_{j2} = f_j(C_{j2})$, *and* $\Delta_k = f_k(C_k)$.

(c) For $j \in \mathbb{O}$ *and* $k \in \mathbb{T}$, $F_{jk}(R_{jk}; \Delta_{j1}, \Delta_{j2}, \Delta_k) = 2\{-2\Phi_1(\Delta_k)\Phi_1(\Delta_{j2}) - \Phi_1(\Delta_{j2}) + 2\Phi_2(\Delta_{j2}, \Delta_k; R_{jk}) - \Phi_1(\Delta_{j1})\Phi_1(\Delta_{j2}) + \Phi_2(0, \Delta_{j2}; -\frac{R_{jk}}{\sqrt{2}}) - 2\Phi_1(\Delta_{j1})\Phi_2(\Delta_{j2}, \Delta_k; R_{jk}) - \Phi_1(\Delta_k)\Phi_2(\Delta_{j2}, \Delta_k; R_{jk}) + 2\Phi_3(0, \Delta_{j2}, \Delta_k; \boldsymbol{R}_{3e}) - 2\Phi_3(0, \Delta_{j2}, \Delta_k; \boldsymbol{R}_{3f}) + 2\Phi_3(\Delta_{j2}, \Delta_{j1}, 0; \boldsymbol{R}_{3d}) + 2\Phi_4(0, \Delta_{j2}, \Delta_k, \Delta_k; \boldsymbol{R}_{4c}) + 2\Phi_4(0, \Delta_{j2}, \Delta_{j1}, \Delta_k; \boldsymbol{R}_{4d}) + 2\Phi_4(0, \Delta_{j2}, \Delta_{j1}, \Delta_k; \boldsymbol{R}_{4e}) + 2\Phi_4(0, \Delta_{j1}, \Delta_{j1}, \Delta_k; \boldsymbol{R}_{4e}) - 2\Phi_2(\Delta_{j1}, \Delta_k; R_{jk})\Phi_2(\Delta_{j2}, \Delta_k; R_{jk}) - 2\Phi_5(0, \Delta_{j1}, \Delta_{j1}, \Delta_k, \Delta_k; \boldsymbol{R}_5) + 2\Phi_5(0, \Delta_{j2}, \Delta_{j1}, \Delta_k, \Delta_k; \boldsymbol{R}_5)\}$, *where* $\Delta_{j1} = f_j(C_{j1}), \Delta_{j2} = f_j(C_{j2}), \Delta_k = f_k(C_k)$,

$$\boldsymbol{R}_{3e} = \begin{pmatrix} 1 & \frac{R_{jk}}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{R_{jk}}{\sqrt{2}} & 1 & 0 \\ -\frac{1}{\sqrt{2}} & 0 & 1 \end{pmatrix}, \boldsymbol{R}_{3f} = \begin{pmatrix} 1 & -\frac{R_{jk}}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ -\frac{R_{jk}}{\sqrt{2}} & 1 & R_{jk} \\ -\frac{1}{\sqrt{2}} & R_{jk} & 1 \end{pmatrix}, \tag{9}$$

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

$$\boldsymbol{R}_{4d} = \begin{pmatrix} 1 & \frac{R_{jk}}{\sqrt{2}} & -\frac{R_{jk}}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{R_{jk}}{\sqrt{2}} & 1 & 0 & 0 \\ -\frac{R_{jk}}{\sqrt{2}} & 0 & 1 & R_{jk} \\ -\frac{1}{\sqrt{2}} & 0 & R_{jk} & 1 \end{pmatrix}, \tag{10}$$

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

$$\boldsymbol{R}_5 = \begin{pmatrix} 1 & -\frac{R_{jk}}{\sqrt{2}} & \frac{R_{jk}}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{R_{jk}}{\sqrt{2}} & 1 & 0 & R_{jk} & 0 \\ \frac{R_{jk}}{\sqrt{2}} & 0 & 1 & 0 & R_{jk} \\ -\frac{1}{\sqrt{2}} & R_{jk} & 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & R_{jk} & 0 & 1 \end{pmatrix}. \tag{11}$$

(d) For $j \in \mathbb{O}$ *and* $k \in \mathbb{O}$, $F(R_{jk}) = 2\Phi_2(\Delta_{j2}, \Delta_{k2}; R_{jk}) - 2\Phi_1(\Delta_{j2})\Phi_1(\Delta_{k2}) - 4\Phi_2(\Delta_{k2}, \Delta_{j2}; R_{jk})\Phi_1(\Delta_{j1}) + 4\Phi_2(\Delta_{j1}, \Delta_{k2}; R_{jk})\Phi_1(\Delta_{j2}) + 2\Phi_2(\Delta_{j2}, \Delta_{k2}; R_{jk})\Phi_2(\Delta_{j1}, \Delta_{k1}; R_{jk}) - 2\Phi_2(\Delta_{j2}, \Delta_{k1}; R_{jk})\Phi_2(\Delta_{j1}, \Delta_{k2}; R_{jk})$, *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 ordinal variable, $\Delta_{j1} = f_j(C_{j1})$ and $\Delta_{j2} = f_j(C_{j2})$ are unknown in practice. We observe that

$$\begin{aligned} \mathbb{E}\{I(X_{ij} = 2)\} &= \mathbb{P}(X_{ij} = 2) = \mathbb{P}(f_j(C_{j2}) > \Delta_{j2}) \\ &= 1 - \Phi_1(\Delta_{j2}), \\ \mathbb{E}\{I(X_{ij} = 0)\} &= \mathbb{P}(X_{ij} = 0) = \mathbb{P}(f_j(C_{j1}) < \Delta_{j1}) \\ &= \Phi_1(\Delta_{j1}). \end{aligned} \tag{12}$$

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 $n_0 + n_1 + n_2 = n$.

Given the bridge functions derived in Theorem 2, we can use the same procedure described below Theorem 1 to obtain the latent correlations between ordinal and other variables. In this way, we can estimate each element of $\boldsymbol{R}$. 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 $\text{argmin}_{\boldsymbol{A} \geq 0} \|\widehat{\boldsymbol{R}} - \boldsymbol{A}\|_F$, where $\boldsymbol{A} \geq 0$ means $\boldsymbol{A}$ is positive semidefinite. Such a problem can be solved by Zhao *et al.* (2014). With a slight abuse of notation, we still denote the solution as $\widehat{\boldsymbol{R}}$. In the Supporting Information, we provide the R code to obtain $\widehat{\boldsymbol{R}}$ for the four types of variables. In our code, we used the code

in the "mixedCCA" package (Yoon and Gaynanova, 2021) to compute correlations of variables other than ordinal variables and used our own code to compute correlations of ordinal variables with other variables. As mentioned by one reviewer, a recent "latentcor" package (Huang *et al.*, 2021) can compute $\widehat{\boldsymbol{R}}$ more efficiently for these four types of variables.

2.2 | Decomposition of the latent correlation matrices

In this section, we describe how to solve the decomposition problem (3) and obtain estimators for low-rank $\boldsymbol{\Sigma}_g$ and sparse $\boldsymbol{\Sigma}_U$ after we obtain an estimator of $\boldsymbol{R}_g$ using the LMGC model. Given an estimator $\widehat{\boldsymbol{R}}_g$ of $\boldsymbol{R}_g$, 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

$$\left(\widehat{\boldsymbol{\Sigma}}_1, \widehat{\boldsymbol{\Sigma}}_2, \widehat{\boldsymbol{\Sigma}}_U\right) = \underset{\boldsymbol{\Sigma}_1 \geq 0, \boldsymbol{\Sigma}_2 \geq 0, \boldsymbol{\Sigma}_U \geq 0}{\operatorname{argmin}} \{\ell(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_U) + \nu_1\|\boldsymbol{\Sigma}_1\|_* + \nu_2\|\boldsymbol{\Sigma}_2\|_* + \nu_3\|\boldsymbol{\Sigma}_U\|_1\}, \quad (13)$$

where ν_1, ν_2, and ν_3 are all nonnegative 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 a matrix $\boldsymbol{M} = (M_{ij}) \in \mathbb{R}^{n\times p}$, 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 kth largest eigenvalue of $\boldsymbol{M}$. In Equation (13), 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 *et al.*, 2010). The L_1-penalty is a well-known penalty function to render a sparse solution (Tibshirani, 1996). In Equation (13), we choose optimal tuning parameters by performing a grid search via fivefold cross-validation. For each combination of these tuning parameters, we reserve one-fifth samples from each group for testing and use the rest for training. For the kth fold, we use the test set to calculate the latent correlation matrix $\widehat{\boldsymbol{R}}_g^{(k)}$, solve (13) using the training set, and denote the solutions as $\widehat{\boldsymbol{\Sigma}}_g^{(-k)}$ and $\widehat{\boldsymbol{\Sigma}}_U^{(-k)}$. We choose the optimal (ν_1, ν_2, ν_3) that minimizes $\sum_{k=1}^5 \|\widehat{\boldsymbol{\Sigma}}_1^{(-k)} + \widehat{\boldsymbol{\Sigma}}_2^{(-k)} + 2\widehat{\boldsymbol{\Sigma}}_U^{(-k)} - \widehat{\boldsymbol{R}}_1^{(k)} - \widehat{\boldsymbol{R}}_2^{(k)}\|_F^2$.

Next, we discuss how to solve (13). First, we obtain an initial estimator of r_g by letting $\widehat{r}_g = \operatorname{argmax}_{j\leq \min\{(n_g,p)\}} \lambda_{j-1}(\widehat{\boldsymbol{R}}_g)/\lambda_j(\widehat{\boldsymbol{R}}_g)$, where $\lambda_j(\widehat{\boldsymbol{R}}_g)$ is the jth largest eigenvalue of $\widehat{\boldsymbol{R}}_g$. Such a rank estimator is commonly used in factor analysis literature (Lam and Yao, 2012; Ahn and Horenstein, 2013). We remark that $\widehat{r}_g$ is allowed to be larger than r_g as the nuclear norm penalty in Equation (13) can further shrink the rank estimator (see Algorithm 1). Let $\boldsymbol{\Theta} = (\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_U)$. We propose to iteratively fix two components in $\boldsymbol{\Theta}$ and solve for the other. Below we provide an algorithm for solving Equation (13). More details about the derivation and numerical convergence of the algorithm, and the rank estimation are provided in Web Appendices B, C, and D of the Supporting Information. The corresponding R code is also included in the Supporting Information.

ALGORITHM 1 The Proximal Gradient Descent Algorithm for solving (13)

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 $\widehat{r}_g = \operatorname{argmax}_{j\leq \min\{(n_g,p)\}} \lambda_{j-1}(\widehat{\boldsymbol{R}}_g)/\lambda_j(\widehat{\boldsymbol{R}}_g)$ and $\widehat{\boldsymbol{\Sigma}}_g^{(0)} = \widehat{\boldsymbol{V}}_g \widehat{\boldsymbol{D}}_g \widehat{\boldsymbol{V}}_g^T$, where $\widehat{\boldsymbol{D}}_g = \operatorname{diag}\{\lambda_1(\widehat{\boldsymbol{R}}_g), \ldots, \lambda_{\widehat{r}_g}(\widehat{\boldsymbol{R}}_g)\}$, $\lambda_j(\widehat{\boldsymbol{R}}_g)$ is the jth 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 $\widehat{\boldsymbol{\Sigma}}_U^{(0)} = (\widehat{\boldsymbol{R}} - \widehat{\boldsymbol{\Sigma}}_1^{(0)} - \widehat{\boldsymbol{\Sigma}}_2^{(0)})/2$. 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 $\boldsymbol{S}_{\nu_1}(\boldsymbol{D}_{\widehat{r}_1}^{(h)}) = \operatorname{diag}\{(\lambda_1(\boldsymbol{D}_{\widehat{r}_1}^{(h)}) - \nu_1)_+, \ldots, (\lambda_{\widehat{r}_1}(\boldsymbol{D}_{\widehat{r}_1}^{(h)}) - \nu_1)_+\}$.

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 $\boldsymbol{A} = \operatorname{argmin}_{\lambda_{\min}(\boldsymbol{A})\geq 0} \|\widehat{\boldsymbol{\Sigma}}_1^{(h+1)} - \boldsymbol{A}\|_F$.

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 $\boldsymbol{S}_{\nu_2}(\boldsymbol{D}_{\widehat{r}_2}^{(h)}) = \operatorname{diag}\{(\lambda_1(\boldsymbol{D}_{\widehat{r}_2}^{(h)}) - \nu_2)_+, \ldots, (\lambda_{\widehat{r}_2}(\boldsymbol{D}_{\widehat{r}_2}^{(h)}) - \nu_2)_+\}$.

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 $\boldsymbol{A} = \operatorname{argmin}_{\lambda_{\min}(\boldsymbol{A})\geq 0} \|\widehat{\boldsymbol{\Sigma}}_2^{(h+1)} - \boldsymbol{A}\|_F$.

Let $\widehat{\boldsymbol{\Sigma}}_U^{(h+1)} = s(\widehat{\boldsymbol{\Sigma}}_U^{(h)} - d\nabla_{\boldsymbol{\Sigma}_U} \ell(\widehat{\boldsymbol{\Sigma}}_1^{(h+1)}, \widehat{\boldsymbol{\Sigma}}_2^{(h+1)}, \widehat{\boldsymbol{\Sigma}}_U^{(h)}), \nu_3 d)$, where $s(\boldsymbol{x}, \pi)_{i,j} = \operatorname{sign}(x_{i,j})(|x_{i,j}| - \pi)_+$ and

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

If $\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}_U^{(h+1)}) \geq 0$ then go to the next step, else let $\widehat{\boldsymbol{\Sigma}}_U^{(h+1)} = \widehat{\boldsymbol{\Sigma}}_U^{(h+1)} - \lambda_{\min}(\widehat{\boldsymbol{\Sigma}}_U^{(h+1)})\boldsymbol{I}$.

Iterate until $\max\{\frac{\|\widehat{\boldsymbol{\Sigma}}_1^{(h+1)} - \widehat{\boldsymbol{\Sigma}}_1^{(h)}\|_F}{\|\widehat{\boldsymbol{\Sigma}}_1^{(h)}\|_F}, \frac{\|\widehat{\boldsymbol{\Sigma}}_2^{(h+1)} - \widehat{\boldsymbol{\Sigma}}_2^{(h)}\|_F}{\|\widehat{\boldsymbol{\Sigma}}_2^{(h)}\|_F}, \frac{\|\widehat{\boldsymbol{\Sigma}}_U^{(h+1)} - \widehat{\boldsymbol{\Sigma}}_U^{(h)}\|_F}{\|\widehat{\boldsymbol{\Sigma}}_U^{(h)}\|_F}\} \leq \zeta$, where ζ is a user-defined stopping threshold.

3 | SIMULATION EXPERIMENTS

To the best of our knowledge, there is no existing method that can directly solve our decomposition problem. Thus, we created three ad hoc competitors to compare with our method: (i) separate decomposition of the sample Pearson

correlation matrix of observed variables, which solves

$$\left(\check{\boldsymbol{\Sigma}}_g^{\dagger}, \check{\boldsymbol{\Sigma}}_{U_g}^{\dagger}\right) = \underset{\boldsymbol{\Sigma}_g \geq 0, \boldsymbol{\Sigma}_{U_g} \geq 0}{\operatorname{argmin}} \|\check{\boldsymbol{R}}_g - \boldsymbol{\Sigma}_g - \boldsymbol{\Sigma}_{U_g}\|_F^2 + \nu_g \|\boldsymbol{\Sigma}_g\|_* + \nu_3 \|\boldsymbol{\Sigma}_{U_g}\|_1, \text{ for } g \in \{1, 2\}, \tag{14}$$

where $\check{\boldsymbol{R}}_g$ is the sample Pearson correlation matrix of observed variables for group g; (ii) joint decomposition of sample Pearson correlation matrix of observed variables, which solves

$$\left(\check{\boldsymbol{\Sigma}}_1, \check{\boldsymbol{\Sigma}}_2, \check{\boldsymbol{\Sigma}}_U\right) = \underset{\boldsymbol{\Sigma}_g \geq 0, \boldsymbol{\Sigma}_U \geq 0}{\operatorname{argmin}} \frac{1}{2} \|\check{\boldsymbol{R}}_1 + \check{\boldsymbol{R}}_2 - \boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2 - 2\boldsymbol{\Sigma}_U\|_F^2 + \nu_1 \|\boldsymbol{\Sigma}_1\|_* + \nu_2 \|\boldsymbol{\Sigma}_2\|_* + \nu_3 \|\boldsymbol{\Sigma}_U\|_1; \tag{15}$$

(iii) separate decomposition of rank-based correlation matrix of latent variables, which solves

$$\left(\widehat{\boldsymbol{\Sigma}}_g^{\dagger}, \widehat{\boldsymbol{\Sigma}}_{U_g}^{\dagger}\right) = \underset{\boldsymbol{\Sigma}_g \geq 0, \boldsymbol{\Sigma}_{U_g} \geq 0}{\operatorname{argmin}} \|\widehat{\boldsymbol{R}}_g - \boldsymbol{\Sigma}_g - \boldsymbol{\Sigma}_{U_g}\|_F^2 + \nu_g \|\boldsymbol{\Sigma}_g\|_* + \nu_3 \|\boldsymbol{\Sigma}_{U_g}\|_1, \text{ for } g \in \{1, 2\}. \tag{16}$$

We carry out simulation studies under both low- and high-dimensional settings and consider three scenarios of $\boldsymbol{\Sigma}_g$ and $\boldsymbol{\Sigma}_U$ for each setting. Scenarios 1–3 are low-dimensional and Scenarios 4–6 are high-dimensional. More details of how we generate $\boldsymbol{\Sigma}_g$ and $\boldsymbol{\Sigma}_U$ are given in Web Appendix F of the Supporting Information. In particular, we choose $\boldsymbol{\Sigma}_U$ to be a banded matrix in Scenarios 1, 2, 4, and 5, and a blockwise sparse matrix in Scenarios 3 and 6. We let the ratio of $\|\boldsymbol{\Sigma}_U\|_F / \|\boldsymbol{R}_g\|_F$ be larger in Scenarios 2 and 5 than in other scenarios to inspect how the proportion of common variation affects the decomposition performance. Table 1 tabulates the structures of the six different scenarios.

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

- **Model 1:** For $g \in \{1, 2\}$, $\boldsymbol{Y}_g = \boldsymbol{Z}_g$, $\boldsymbol{X}_g = \boldsymbol{h}_g(\boldsymbol{Y}_g)$, where $\boldsymbol{h}_g$ is defined in (1) with $C_{1,j} = 0.3$, $C_{2,j} = 0.1$ for $j \in \mathbb{B}$, 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 $j \in \mathbb{O}$.
- **Model 2:** $\boldsymbol{Y}_1 = \exp(\boldsymbol{Z}_1)$, $\boldsymbol{Y}_2 = \boldsymbol{Z}_2$, $\boldsymbol{X}_g = \boldsymbol{h}_g(\boldsymbol{Y}_g)$, where $\boldsymbol{h}_g$ is defined in (1) with $C_{1,j} = 1.5$, $C_{2,j} = 0.1$ for $j \in \mathbb{B}$, 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 $j \in \mathbb{O}$.
- **Model 3:** $\boldsymbol{Y}_1 = \exp(\boldsymbol{Z}_1)$, $\boldsymbol{Y}_2 = \boldsymbol{Z}_2^3$; $\boldsymbol{X}_g = \boldsymbol{h}_g(\boldsymbol{Y}_g)$, where $\boldsymbol{h}_g$ is defined in (1) with $C_{1,j} = 1.5$, $C_{2,j} = 0.1$ for $j \in \mathbb{B}$, 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 $j \in \mathbb{O}$.

First, we investigate how the rank-based correlation matrix estimator compares with the Pearson correlation estimator. We denote $\check{\boldsymbol{R}}_g$ as the sample Pearson correlation coefficient for group g, where its (j, k)th element is defined as

$$\check{R}_{g;(j,k)} = \sum_{i=1}^{n_g} (X_{g;i,j} - \bar{X}_{g;j})(X_{g;i,k} - \bar{X}_{g;k}) \Big/ \times \left[\left\{ \sum_{i=1}^{n_g} (X_{g;i,j} - \bar{X}_{g;j})^2 \right\} \left\{ \sum_{i=1}^{n_g} (X_{g;i,k} - \bar{X}_{g;k})^2 \right\} \right]^{1/2}, \tag{17}$$

$X_{g;i,j}$ is the (i, j)th element of $\boldsymbol{X}_g$ and $\bar{X}_{g;j} = (1/n_g) \sum_{i=1}^{n_g} X_{g;i,j}$ for $j = 1, \ldots, p$. Panel (a) of Figure 1 gives the boxplots of $\|\widehat{\boldsymbol{R}}_g - \boldsymbol{R}_g\|_F$ and $\|\check{\boldsymbol{R}}_g - \boldsymbol{R}_g\|_F$. Figure 1 appears in color in the electronic version of this article, and any mention of color refers to that version. It is seen that $\widehat{\boldsymbol{R}}_g$ (red) outperforms $\check{\boldsymbol{R}}_g$ (blue) in all scenarios.

Next, we compare the estimation errors 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 Panel (b) of Figure 1 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 $\boldsymbol{\Sigma}_U$. We define sensitivity as the proportion of nonzero entries in $\boldsymbol{\Sigma}_U$ being estimated as nonzeros and specificity as the proportion of zero entries in $\boldsymbol{\Sigma}_U$ being estimated as zeros. Figure 2 demonstrates the sensitivity and specificity of four competitors over 100 simulations. This figure appears in color in the electronic version of this article, and any mention of color refers to that version. 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, the $\boldsymbol{\Sigma}_U$'s have 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.

TABLE 1 Simulation scenarios

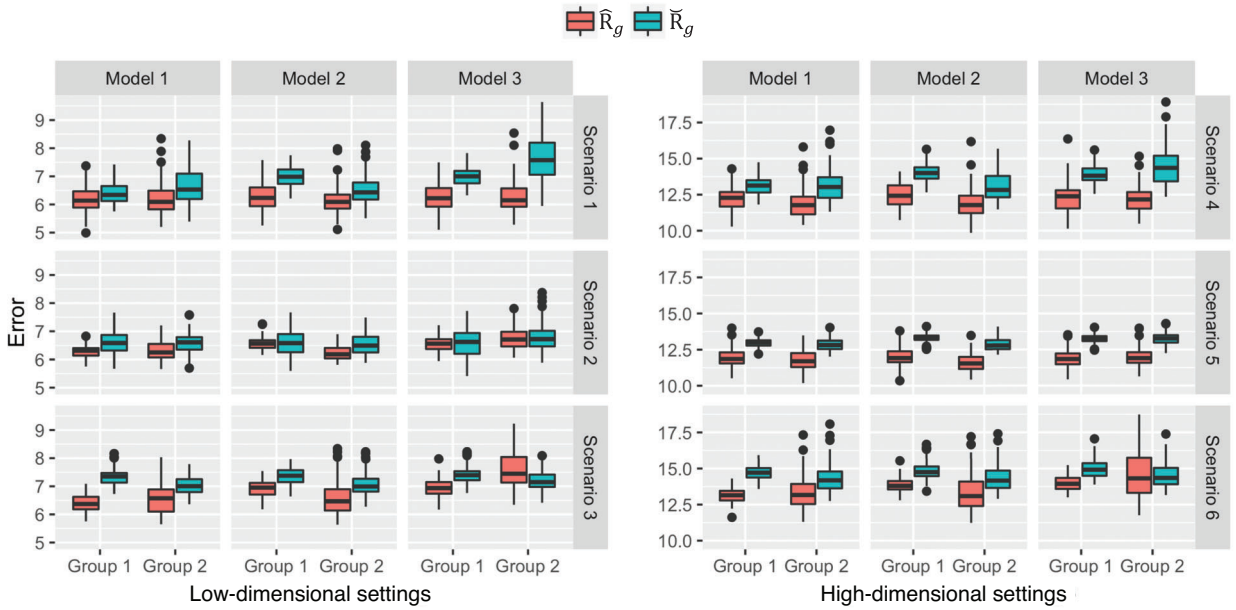
Low dimension $(n_1, n_2, p) = (100, 100, 60)$			
	Scenario 1	**Scenario 2**	**Scenario 3**
Σ_1	$\Sigma_1 = Q_1 D_1 Q_1^T$, D_1 is a diagonal matrix, Q_1 is an orthonormal matrix	$\Sigma_1 = Q_1 D_1 Q_1^T$, D_1 is a diagonal matrix, Same Q_1 as in Scenario 1	Same as in Scenario 1
Σ_2	$\Sigma_2 = \mathbf{w}_2\mathbf{w}_2^T$, $\mathbf{w}_2 = (\mathbf{w}_{21}, \mathbf{w}_{22})$	$\Sigma_2 = \mathbf{w}_2\mathbf{w}_2^T$, $\mathbf{w}_2 = (\mathbf{w}_{21}, \mathbf{w}_{22})$	$\Sigma_2 = \mathbf{w}_2\mathbf{w}_2^T$, $\mathbf{w}_2 = (\mathbf{w}_{21}, \mathbf{w}_{22})$
	$w_{22;j} = \sqrt{1 - \sigma_{u,j} - w_{21;j}^2}$	$w_{22;j} = \sqrt{1 - \sigma_{u,j} - w_{21;j}^2}$	$w_{22;j} = \sqrt{1 - \sigma_{u,j} - w_{21;j}^2}$
Σ_U	$\text{diag}(\Sigma_U) = 1 - \text{diag}(\Sigma_1)$	$\text{diag}(\Sigma_U) = 1 - \text{diag}(\Sigma_1)$	$\text{diag}(\Sigma_U) = 1 - \text{diag}(\Sigma_1)$
	$\sigma_{u,ij} = \begin{cases} \sigma_{u,i}\sigma_{u,j}\rho^{\lvert i-j \rvert} & \text{if } \lvert i-j \rvert = 1 \\ \sigma_{u,ij} = 0 & \text{otherwise} \end{cases}$	$\sigma_{u,ij} = \begin{cases} \sigma_{u,i}\sigma_{u,j}\rho^{\lvert i-j \rvert} & \text{if } \lvert i-j \rvert \leq 2 \\ \sigma_{u,ij} = 0 & \text{otherwise} \end{cases}$	blockwise sparse
$\lVert\Sigma_U\rVert_F / \lVert R_1 \rVert_F$	44.89%	68.41%	42.09%
$\lVert\Sigma_U\rVert_F / \lVert R_2 \rVert_F$	36.82%	62.72%	33.31%
High dimension $(n_1, n_2, p) = (50, 50, 90)$			
	Scenario 4	**Scenario 5**	**Scenario 6**
Σ_1	$\Sigma_1 = Q_1 D_1 Q_1^T$, D_1 is a diagonal matrix, Q_1 is an orthonormal matrix	$\Sigma_1 = Q_1 D_1 Q_1^T$, D_1 is a diagonal matrix, Same Q_1 as in Scenario 4	Same as in Scenario 4
Σ_2	$\Sigma_2 = \mathbf{w}_2\mathbf{w}_2^T$, $\mathbf{w}_2 = (\mathbf{w}_{21}, \mathbf{w}_{22})$	$\Sigma_2 = \mathbf{w}_2\mathbf{w}_2^T$, $\mathbf{w}_2 = (\mathbf{w}_{21}, \mathbf{w}_{22})$	$\Sigma_2 = \mathbf{w}_2\mathbf{w}_2^T$, $\mathbf{w}_2 = (\mathbf{w}_{21}, \mathbf{w}_{22})$
	$w_{22;j} = \sqrt{1 - \sigma_{u,j} - w_{21;j}^2}$	$w_{22;j} = \sqrt{1 - \sigma_{u,j} - w_{21;j}^2}$	$w_{22;j} = \sqrt{1 - \sigma_{u,j} - w_{21;j}^2}$
Σ_U	$\text{diag}(\Sigma_U) = 1 - \text{diag}(\Sigma_1)$	$\text{diag}(\Sigma_U) = 1 - \text{diag}(\Sigma_1)$	$\text{diag}(\Sigma_U) = 1 - \text{diag}(\Sigma_1)$
	$\sigma_{u,ij} = \begin{cases} \sigma_{u,i}\sigma_{u,j}\rho^{\lvert i-j \rvert} & \text{if } \lvert i-j \rvert = 1 \\ \sigma_{u,ij} = 0 & \text{otherwise} \end{cases}$	$\sigma_{u,ij} = \begin{cases} \sigma_{u,i}\sigma_{u,j}\rho^{\lvert i-j \rvert} & \text{if } \lvert i-j \rvert \leq 2 \\ \sigma_{u,ij} = 0 & \text{otherwise} \end{cases}$	blockwise sparse
$\lVert\Sigma_U\rVert_F / \lVert R_1 \rVert_F$	36.37%	69.39%	33.2%
$\lVert\Sigma_U\rVert_F / \lVert R_2 \rVert_F$	28.83%	62.82%	25.47%

4 | AN ANALYSIS OF A *C. TRACHOMATIS* GENITAL TRACT INFECTION STUDY

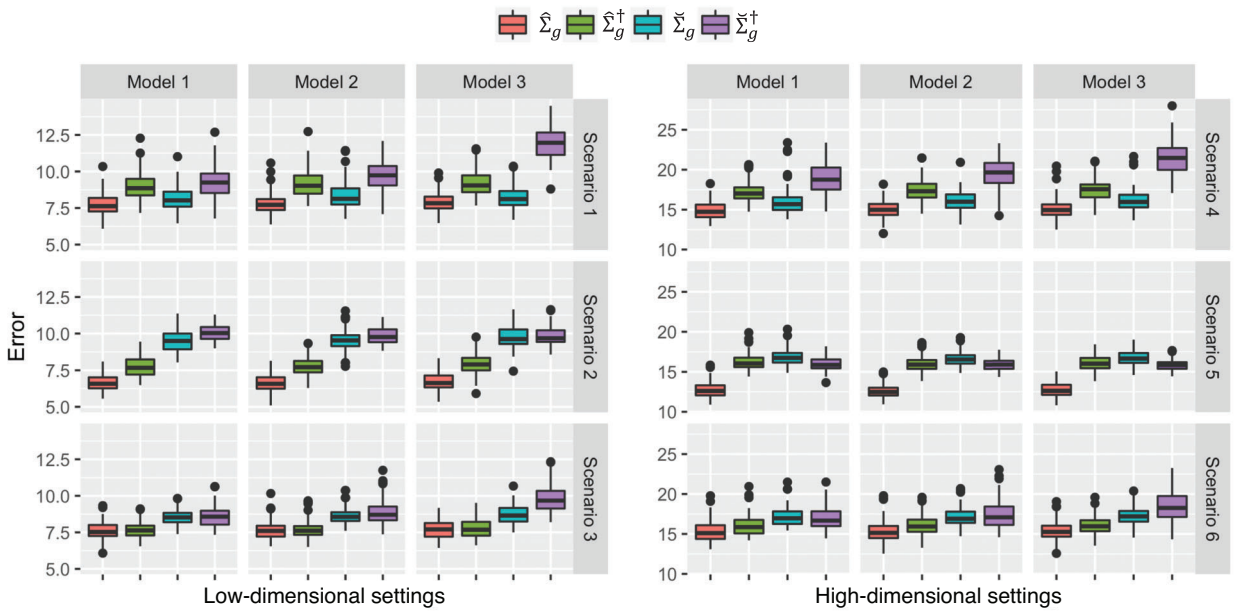
We applied our method to the multimodal data from the T cell response against Chlamydia (TRAC) cohort (Russell *et al.*, 2016), which is designed for studying chlamydial genital tract infection. *C. trachomatis* can ascend from the cervix to the uterus and fallopian tubes in some women, and potentially 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 modeling 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 cervical cytokine expression data and genotype data from 128 women in TRAC, who either had both cervical and endometrial infection (Endo+ group, $n = 60$) or had only limited cervical infection (Endo- group, $n = 68$). Descriptions of the TRAC cohort, processing and quality control of cervical cytokines data and genotype data have been published in detail previously (Poston *et al.*, 2019). There are 48 cytokines in the cervical cytokines data. Cytokine levels were determined using Milliplex Magnetic Bead Assay. The cytokine values were log2 transformed, and treated as normally distributed continuous variables. Directly genotyped SNPs were used in this study, while imputed genotypes were excluded. We treated the genotypes as ordinal variables with three levels.

Expression quantitative trait loci (eQTLs) are the SNPs that influence expression levels of mRNA transcripts, which provide functional interpretation of the correlation



(a) Estimation errors of $\boldsymbol{R}_g$ given by the Kendall's τ estimator ($\widehat{\boldsymbol{R}}_g$) and the Pearson correlation estimator ($\breve{\boldsymbol{R}}_g$)



(b) Estimation errors of $\boldsymbol{\Sigma}_g$ given by the four methods.

$\hat{\Sigma}_g$ is the estimator given by our method.

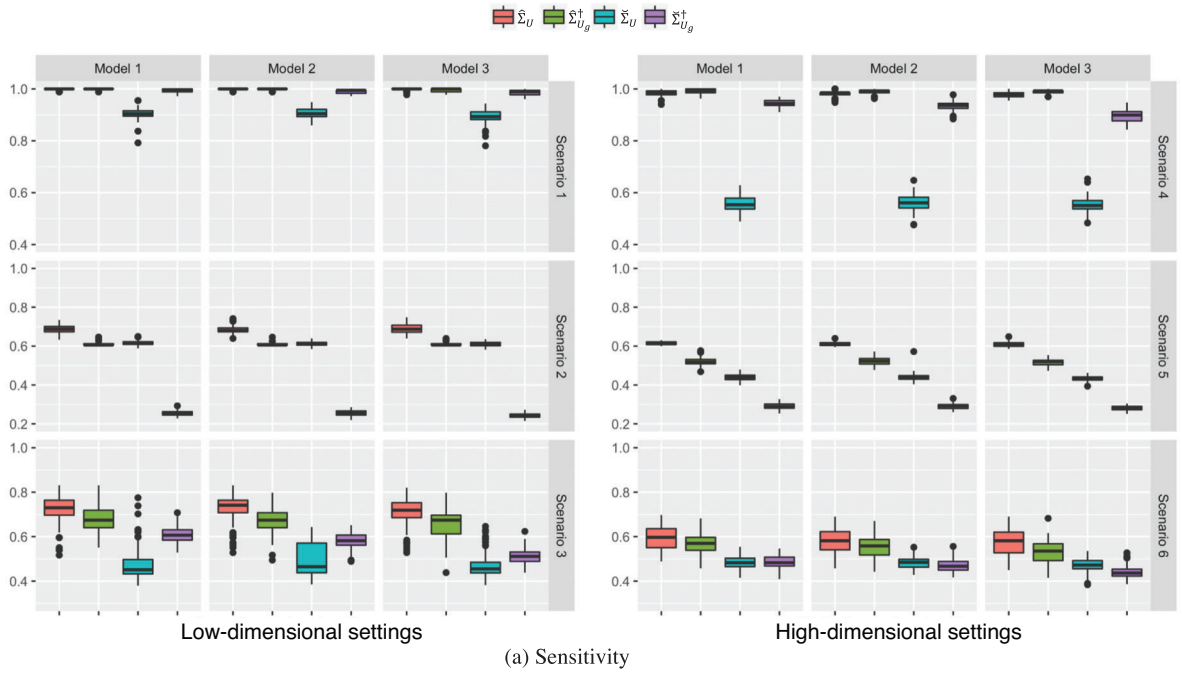
$\hat{\Sigma}_g^\dagger$ is the estimator given by separately decomposing the Kendall's τ correlation.

$\breve{\Sigma}_g$ is the estimator given by jointly decomposing the Pearson correlation.

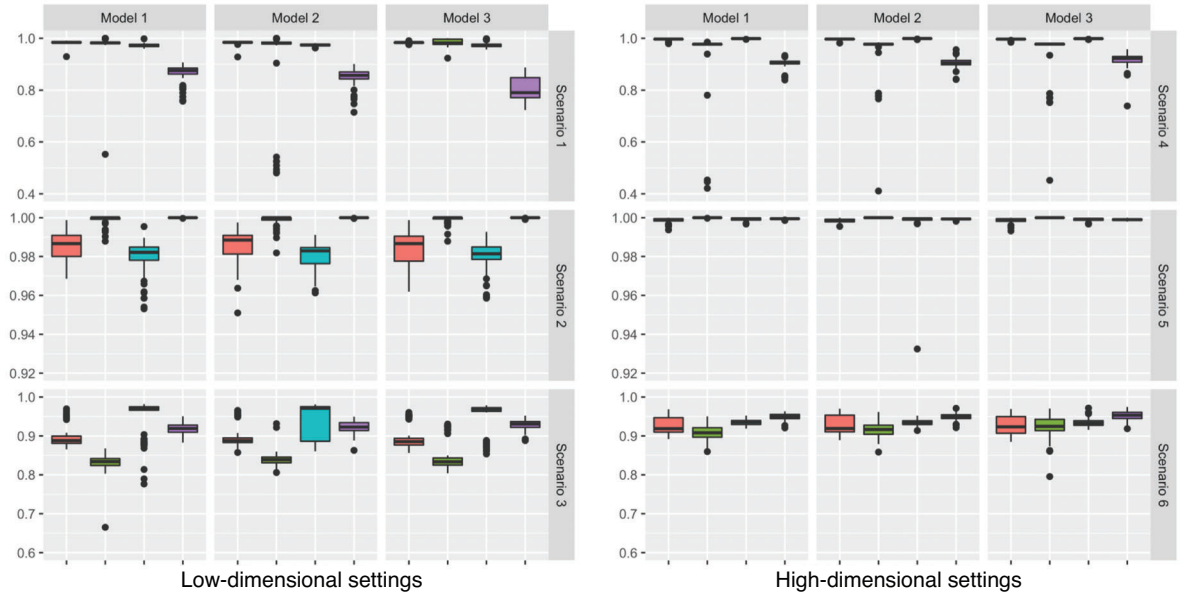
$\breve{\Sigma}_g^\dagger$ is the estimator given by separately decomposing the Pearson correlation.

FIGURE 1 Estimation errors. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

between SNPs and cytokines. We thus primarily focused on SNPs that were cis-eQTLs of the cytokines, defined as SNPs within 1 MB 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, and preferentially kept the



(a) Sensitivity



(b) Specificity

$\hat{\Sigma}_U$ is the estimator given by our method.
$\hat{\Sigma}^{\dagger}_{U_g}$ is the estimator given by separately decomposing the Kendall's τ correlation.
$\check{\Sigma}_U$ is the estimator given by jointly decomposing the Pearson correlation.
$\check{\Sigma}^{\dagger}_{U_g}$ is the estimator given by separately decomposing the Pearson correlation.

FIGURE 2 Variable selection accuracy of $\mathbf{\Sigma}_U$ given by the four methods. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

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 correlation with another SNP is greater than the upper 0.1% quantile of the absolute values in the latent correlation matrix, while keeping the more significant SNPs in the cis-eQTL detection. Final data set for each group had a total of 227 variables, including 42 cytokines and 185 SNPs.

We applied our proposed method to this data set to obtain $\widehat{\boldsymbol{R}}_g$, $\widehat{\mathbf{\Sigma}}_g$ and $\widehat{\mathbf{\Sigma}}_U$. Figures 3 and 4 represent their

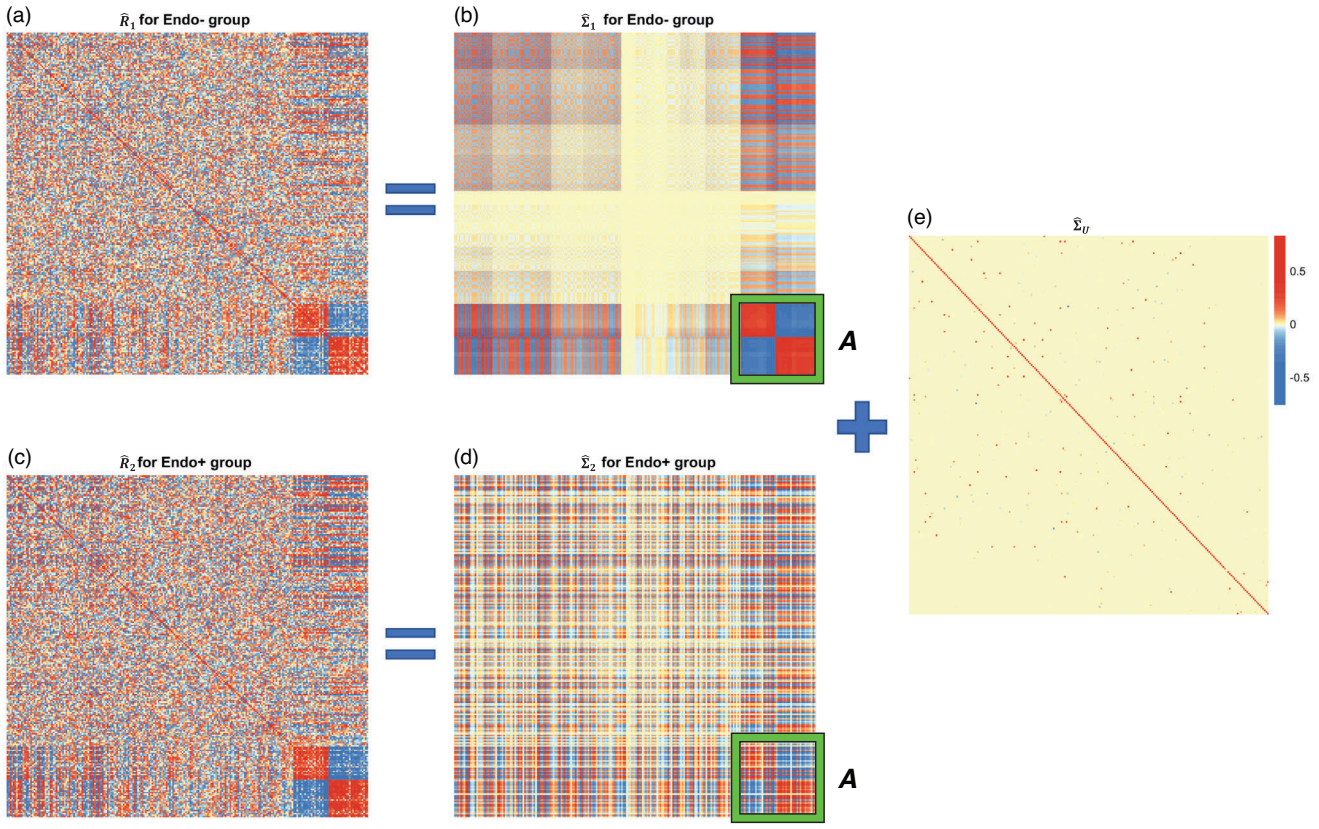


FIGURE 3 Heatmaps of $\widehat{\boldsymbol{R}}_g$, $\widehat{\boldsymbol{\Sigma}}_g$, and $\widehat{\boldsymbol{\Sigma}}_U$ for Endo- ($g = 1$) and Endo+ ($g = 2$) groups. Rows and columns of all heatmaps were ordered by applying clustering to the absolute value of $\widehat{\boldsymbol{\Sigma}}_1$. 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. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

heatmaps. Rows and columns of the heatmaps in Figure 3 were ordered by applying hierarchical clustering to the absolute value of $\widehat{\boldsymbol{\Sigma}}_1$, and those in Figure 4 were ordered by applying clustering to the absolute value of $\widehat{\boldsymbol{\Sigma}}_2$.

We highlighted the cluster of variables most distinct from the rest variables in $\widehat{\boldsymbol{\Sigma}}_1$ for the Endo- group (Figure 3b) and the same group of variables in $\widehat{\boldsymbol{\Sigma}}_2$ for Endo+ group (Figure 3d) with green squares, namely, Block $\boldsymbol{A}$, which consists seven cytokines (CXCL13, EGF, IL17A, IL23A, CXCL10, CCL7, and CCL23) and 40 SNPs. Figure 3 appears in color in the electronic version of this article, and any mention of color refers to that version. Among the 40 SNPs, 28 (70%) are cis-eQTLs of these 7 cytokines, and 12 are trans-eQTLs of these cytokines.

These seven cytokines formed two subnetworks, one includes IL17A, IL23A, CXCL10, CXCL13, and their eQTLs. These four 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 the differentiation of naive CD4+ T cells into Th17 cells (Iwakura and Ishigame, 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). In addition, the connectivity of CXCL13 and IL-17A has been evidenced experimentally (Rangel-Moreno *et al.*, 2011).

The other subnetwork in block $\boldsymbol{A}$ includes CCL7, CCL23, and EGF and their eQTLs. These their 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 most distinct from all other clusters in the unique low-rank part $\widehat{\boldsymbol{\Sigma}}_2$ for Endo+ group (Figure 4d) and the same group of variables in the

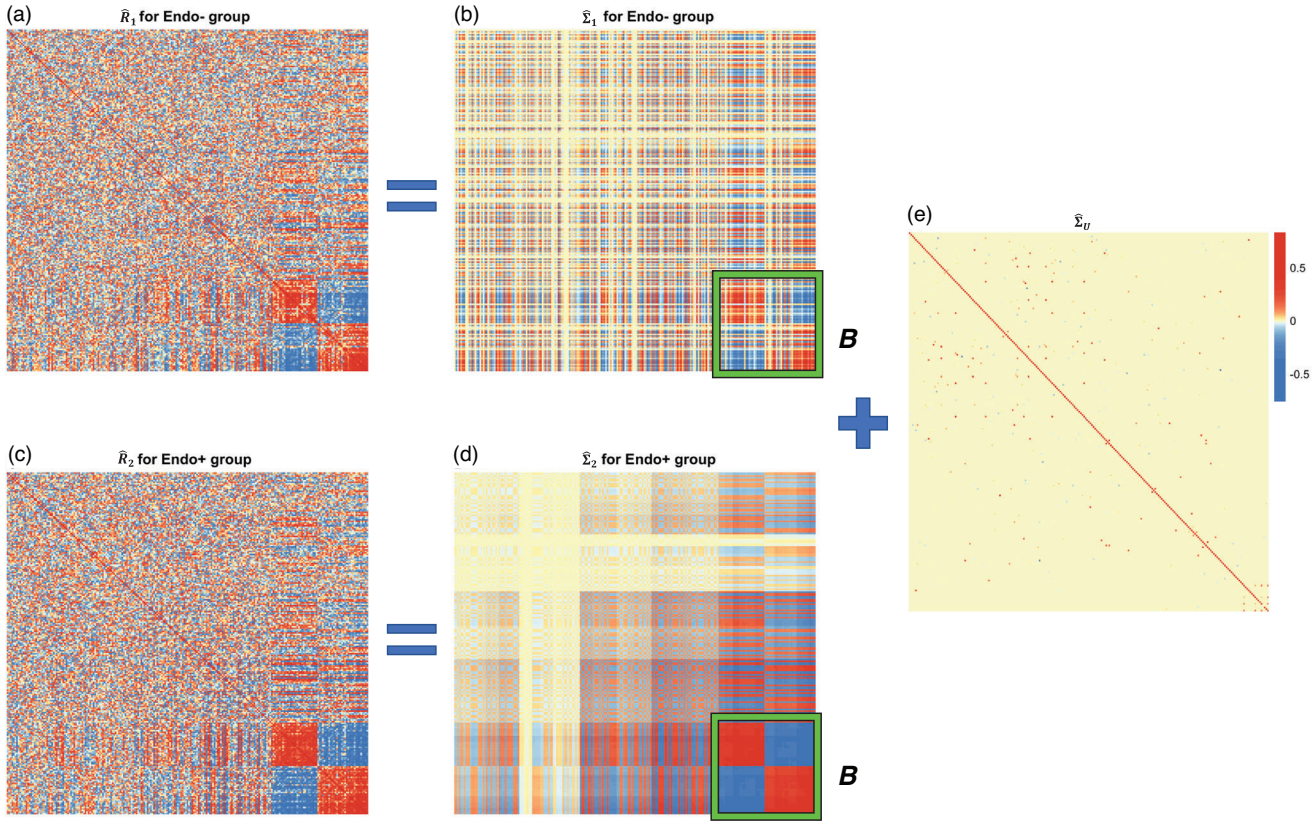


FIGURE 4 Heatmaps of $\hat{\boldsymbol{R}}_g$, $\hat{\boldsymbol{\Sigma}}_g$, and $\hat{\boldsymbol{\Sigma}}_U$ for Endo- ($g = 1$) and Endo+ ($g = 2$) groups. Rows and columns of all heatmaps were ordered by applying clustering to the absolute value of $\hat{\boldsymbol{\Sigma}}_2$. The cluster that is most distinct from all other clusters in $\hat{\boldsymbol{\Sigma}}_2$ is highlighted in the green square. The same group of variables in $\hat{\boldsymbol{\Sigma}}_1$ is also highlighted in the green square. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

unique low-rank part $\hat{\boldsymbol{\Sigma}}_1$ for Endo- (Figure 4b) with green squares, namely, Block $\boldsymbol{B}$. Figure 4 appears in color in the electronic version of this article, and any mention of color refers to that version. Block $\boldsymbol{B}$ consists of five cytokines (CSF3, FLT3LG, TNFSF10, CCL5, and CCL23) and 56 SNPs.

All these five 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 cell populations. TNFSF10 is critical in promoting infection-induced inflammation, 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 S1 and S2. The cytokine networks among CXCL14, IL15, IL-16, PDGF-A, and 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.

5 | DISCUSSION

We propose a novel method to decompose the correlations of mixed variables from multigroup subjects into a shared component and a group-specific component. Our main contributions are two folds. First, we derive the bridge functions for measuring correlations between three-level ordinal and other variables and prove their monotonicity. These results together with the existing works (Liu *et al.*, 2009; Fan *et al.*, 2017; Feng and Ning, 2019; Yoon *et al.*, 2020) on measuring correlations among continuous, binary, and truncated variables, provide a unified framework to quantify correlations among the four types of variables. Second, we provide a decomposition method to dissect the group-specific variation from the common

variation in the background. Our method applies to mixed variables from multiple groups. Our numerical studies demonstrate its advantage over the group-by-group analysis and its usefulness in gene network analysis. Some other technical details, such as rank selection, alternative loss function, and numerical convergence of our algorithm are discussed in Web Appendices C, D, and E.

Our method can be extended to handle more than two groups. In general, if there are G groups of subjects, we can first obtain the rank-based correlation estimator $\widehat{\boldsymbol{R}}_g$ for each group. Then, we can solve a problem of

$$(\widehat{\boldsymbol{\Sigma}}_g, \widehat{\boldsymbol{\Sigma}}_U) = \underset{\boldsymbol{\Sigma}_g \geq 0, \boldsymbol{\Sigma}_U \geq 0}{\operatorname{argmin}} \left\{ G^{-1} || \sum_{g=1}^{G} (\widehat{\boldsymbol{R}}_g - \boldsymbol{\Sigma}_g - \boldsymbol{\Sigma}_U) ||_F^2 + \sum_{g=1}^{G} \nu_g \|\boldsymbol{\Sigma}_g\|_* + \nu_{G+1} \|\boldsymbol{\Sigma}_U\|_1 \right\}, \tag{18}$$

which is an extension of (13). Such a problem can also be solved by a similar proximal gradient descent algorithm as described in Algorithm 1. But, our R code only deals with two groups.

We point out that our rank-based correlation estimator only applies to three-level ordinal variables. Quan *et al.* (2018) derived the bridge function for measuring the correlation between continuous and ordinary variables with arbitrarily many levels. However, it is hard to derive the counterpart for the correlation between ordinary variables with arbitrarily many levels, as the breakdowns of these variables are complicated (Quan *et al.*, 2018). This can be a future research topic. Another extension is that if we have prior information on the structure of $\boldsymbol{\Sigma}_g$, we can add more regularization terms in (18) to ensure the resulting estimators reflect such a structure.

ACKNOWLEDGMENTS

The authors thank the associate editor and two reviewers for their valuable comments, which have led to the great improvement of the manuscript. This work was partially supported by the National Institutes of Health grants U19AI084024, U19AI144181 to TD; U19AI144181 to XZ; and R01AG073259 to QL.

DATA AVAILABILITY STATEMENT

The data that support the findings in this paper are available from the corresponding authors upon reasonable request.

ORCID

Quefeng Li https://orcid.org/0000-0003-0707-2763

REFERENCES

Ahn, S.C. and Horenstein, A.R. (2013) Eigenvalue ratio test for the number of factors. *Econometrica*, 81, 1203–1227.

Alter, O., Brown, P.O. and Botstein, D. (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, 3351–3356.

Amar, D., Safer, H. and Shamir, R. (2013) Dissection of regulatory networks that are altered in disease via differential co-expression. *PLoS Computational Biology*, 9, e1002955.

Andrew, D.W., Cochrane, M., Schripsema, J.H., Ramsey, K.H., Dando, S.J., O'Meara, C.P. et al. (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, e76664.

Candès, E.J. and Recht, B. (2009) Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9, 717–772.

Candès, E.J. and Tao, T. (2010) The power of convex relaxation: near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56, 2053–2080.

Choi, Y. and Kendziorski, C. (2009) Statistical methods for gene set co-expression analysis. *Bioinformatics*, 25, 2780–2786.

Darville, T., Albritton, H.L., Zhong, W., Dong, L., O'Connell, C.M., Poston, T.B. 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, e13103.

De Vito, R., Bellio, R., Trippa, L. and Parmigiani, G. (2019) Multistudy factor analysis. *Biometrics*, 75, 337–346.

Fan, J., Liu, H., Ning, Y. and Zou, H. (2017) High dimensional semiparametric latent graphical model for mixed data. *Journal of the Royal Statistical Society: Series B*, 79, 405–421.

Feng, Q., Jiang, M., Hannig, J. and Marron, J. (2018) Angle-based joint and individual variation explained. *Journal of Multivariate Analysis*, 166, 241–265.

Feng, H. and Ning, Y. (2019) High-dimensional mixed graphical model with ordinal data: parameter estimation and statistical inference. In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, Vol. 89, pp. 654–663.

Ha, M.J., Baladandayuthapani, V. and Do, K.-A. (2015) DINGO: differential network analysis in genomics. *Bioinformatics*, 31, 3413–3420.

Huang, M., Müller, C.L. and Gaynanova, I. (2021) latentcor: An R package for estimating latent correlations from mixed data types. *Journal of Open Source Software*, 6, 3634–3638.

Iwakura, Y. and Ishigame, H. (2006) The IL-23/IL-17 axis in inflammation. *Journal of Clinical Investigation*, 116, 1218–1222.

Keates, S., Han, X., Kelly, C.P. and Keates, A.C. (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. *Journal of Immunology*, 178, 8013–8021.

Kiviat, N., Wolner-Hanssen, P., Eschenbach, D., Wasserheit, J., Paavonen, J., Bell, T. et al. (1990) Endometrial histopathology in patients with culture-proved upper genital tract infection and laparoscopically diagnosed acute salpingitis. *American Journal of Surgical Pathology*, 14, 167–175.

Lam, C. and Yao, Q. (2012) Factor modeling for high-dimensional time series: inference for the number of factors. *Annals of Statistics*, 40, 694–726.

Li, G. and Gaynanova, I. (2018) A general framework for association analysis of heterogeneous data. *Annals of Applied Statistics*, 12, 1700–1726.

Lijek, R.S., Helble, J.D., Olive, A.J., Seiger, K.W. and Starnbach, M.N. (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, 2216–2221.

Liu, H., Lafferty, J. and Wasserman, L. (2009) The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10, 2295–2328.

Lock, E.F., Hoadley, K.A., Marron, J.S. and Nobel, A.B. (2013) Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Annals of Applied Statistics*, 7, 523–542.

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

Marino, J., Furmento, V.A., Zotta, E. and Roguin, L.P. (2009) Peritumoral administration of granulocyte colony-stimulating factor induces an apoptotic response on a murine mammary adenocarcinoma. *Cancer Biology & Therapy*, 8, 1737–1743.

Mazumder, R., Hastie, T. and Tibshirani, R. (2010) Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11, 2287–2322.

Nastase, M.V., Zeng-Brouwers, J., Beckmann, J., Tredup, C., Christen, U., Radeke, H.H. et al. (2018) Biglycan, a novel trigger of Th1 and Th17 cell recruitment into the kidney. *Matrix Biology*, 68, 293–317.

Ponnapalli, S.P., Saunders, M.A., Van Loan, C.F. and Alter, O. (2011) A higher-order generalized singular value decomposition for comparison of global mRNA expression from multiple organisms. *PloS One*, 6, e28072.

Poston, T.B., Lee, D.E., Darville, T., Zhong, W., Dong, L., O'Connell, C.M. et al. (2019) Cervical cytokines associated with Chlamydia trachomatis susceptibility and protection. *Journal of Infectious Diseases*, 220, 330–339.

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

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

Rangel-Moreno, J., Carragher, D.M., de la Luz Garcia-Hernandez, M., Hwang, J.Y., Kusser, K., Hartson, L. et al. (2011) The development of inducible bronchus-associated lymphoid tissue depends on IL-17. *Nature Immunology*, 12, 639–646.

Russell, A.N., Zheng, X., O'Connell, C.M., Taylor, B.D., Wiesenfeld, H.C., Hillier, S.L. et al. (2016) Analysis of factors driving incident and ascending infection and the role of serum antibody in Chlamydia trachomatis genital tract infection. *Journal of Infectious Diseases*, 213, 523–531.

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

Shin, S.Y., Lee, D.H., Lee, J., Choi, C., Kim, J.-Y., Nam, J.-S. et al. (2017) C-C motif chemokine receptor 1 (CCR1) is a target of the EGF-AKT-mTOR-STAT3 signaling axis in breast cancer cells. *Oncotarget*, 8, 94591–94605.

Shu, H., Wang, X. and Zhu, H. (2020) D-CCA: a decomposition-based canonical correlation analysis for high-dimensional datasets. *Journal of the American Statistical Association*, 115, 292–306.

Tesson, B.M., Breitling, R. and Jansen, R.C. (2010) DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinformatics*, 11, 497–505.

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

van Dam, S., Vosa, U., van der Graaf, A., Franke, L. and de Magalhaes, J.P. (2018) Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in Bioinformatics*, 19, 575–592.

Watson, M. (2006) CoXpress: differential co-expression in gene expression data. *BMC Bioinformatics*, 7, 509–520.

Yoon, G., Carroll, R.J. and Gaynanova, I. (2020) Sparse semiparametric canonical correlation analysis for data of mixed types. *Biometrika*, 107, 609–625.

Yoon, G. and Gaynanova, I. (2021) *mixedCCA: Sparse Canonical Correlation Analysis for High-Dimensional Mixed Data*. R Package Version 1.4.6.

Zhao, T., Roeder, K. and Liu, H. (2014) Positive semidefinite rank-based correlation matrix estimation with application to semiparametric graph estimation. *Journal of Computational and Graphical Statistics*, 23, 895–922.

Zhong, W., Dong, L., Poston, T.B., Darville, T., Spracklen, C.N., Wu, D. et al. (2020) Inferring regulatory networks from mixed observational data using directed acyclic graphs. *Frontiers in Genetics*, 11, 8.

Zhou, G., Cichocki, A., Zhang, Y. and Mandic, D.P. (2015) Group component analysis for multiblock data: common and individual feature extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 27, 2426–2439.

Zhu, H., Li, G. and Lock, E.F. (2020) Generalized integrative principal component analysis for multi-type data with block-wise missing structure. *Biostatistics*, 21, 302–318.

SUPPORTING INFORMATION

Web Appendices, Tables, Figures, and Proofs referenced in Sections 2, 3, 4 are available with this paper at the Biometrics website on Wiley Online Library. The R code for implementing the proposed method is available at the Biometrics website on Wiley Online Library.

How to cite this article: Liu, Y., Darville, T., Zheng, X., Li, Q. Decomposition of variation of mixed variables by a latent mixed Gaussian copula model. *Biometrics*. 2022;1–14.
https://doi.org/10.1111/biom.13660