Submitted 1/24; Revised 4/25; Published 8/25

Sparse Semiparametric Discriminant Analysis for High-dimensional Zero-inflated Data

Hee Cheol Chung HCHUNG13@CHARLOTTE.EDU
Department of Mathematics and Statistics
University of North Carolina at Charlotte
Charlotte, NC 28223, USA

Yang Ni YANG.NI@AUSTIN.UTEXAS.EDU
Department of Statistics and Data Sciences
The University of Texas at Austin
Austin, TX 78705, USA

Irina Gaynanova IRINAGN@UMICH.EDU
Department of Biostatistics
University of Michigan
Ann Arbor, MI 48109, USA

Editor: Ji Zhu

Abstract

Sequencing-based technologies provide an abundance of high-dimensional biological data sets with highly skewed and zero-inflated measurements. Despite the computational efficiency and high interpretability offered by linear classification methods, the violation of underlying distribution assumptions, driven by high skewness and zero inflation, results in invalid classification rules and interpretations. Furthermore, existing data transformation methods addressing these violations introduce ambiguity, rendering the final model and classification performance contingent on the specific transformation employed. To tackle these challenges, we propose a novel semiparametric framework for discriminant analysis based on the truncated latent Gaussian copula model. This model accommodates skewness and zero inflation, and its estimation procedure ensures robustness against data transformations. To facilitate model interpretability, we incorporate ℓ_1 sparsity regularization and establish the consistency of the classification directions in high-dimensional settings. We validate our approach using human gut microbiome, breast cancer microRNA, and single-cell RNA sequencing data, highlighting its superior classification accuracy and robustness to data transformations.

Keywords: Latent Gaussian copula, probit regression, robust classification, sequencing data, skewed data; variable selection

1. Introduction

Linear methods are popular in classification analysis due to their computational efficiency and high interpretability. However, the complexity of modern high-dimensional data raises many challenges in its application. For example, microbiome, microRNA, and single-cell RNA sequencing data not only have a large number of variables relative to the sample size, but also are highly skewed and zero-inflated (Silverman et al., 2020). A large number of variables makes classical discriminant analysis less interpretable due to the lack of variable selection, and less accurate due to the singularity of the covariance matrix (Bickel and Levina, 2004). Moreover, deviations from normality such as skewness and zero-inflation challenge the underlying distribution assumptions, consequently affecting the reliability of linear discriminant analysis, particularly in its sensitivity to outliers, as noted by Hastie et al. (2009). To mitigate these violations, various data transformation methods like logarithmic, power, and log-ratio (Aitchison, 1983) transformations have been proposed. However, their application introduces increased ambiguity, as the signal variable identification and classification performance become contingent on the specific transformation employed (Ahlmann-Eltze and Huber, 2023; Weiss et al., 2017).

A rich body of work extends the classical linear discriminant analysis to high-dimensional settings. A common approach is to add regularization to the classification direction vector, e.g., the ℓ_2-regularization (Guo et al., 2007) or the ℓ_1-regularization (Witten and Tibshirani, 2011; Mai et al., 2012; Gaynanova et al., 2016). An alternative approach is to consider the data piling phenomenon in high dimensions, that is, observations from each class can be projected to a single point. Ahn and Marron (2010) estimate the classification direction by maximizing the distance between data piling sites, and Lee et al. (2013) regularize the degree of data piling. While these approaches improve accuracy in high-dimensional settings, they still lead to poor performance in the presence of skewness and zero inflation. Several works consider relaxations of the Gaussian assumption. Ahn et al. (2021) utilize a trace ratio formulation of Fisher's discriminant analysis, which is more robust to violations of Gaussianity than the standard formulation. Clemmensen et al. (2011) model each class using a Gaussian mixture with subclass-specific means and common covariance matrices. Hernández and Velilla (2005) consider a fully nonparametric kernel linear discriminant analysis. Lapanowski and Gaynanova (2019) consider the optimal scoring formulation of kernel linear discriminant analysis with sparsity regularization. These methods, however, do not account for zero inflation. In sequencing data, zeros typically represent the values below sequencing detection limit (Lubbe et al., 2021) and hence treating these zeros as absolute can lead to inaccurate classification and inference. While Witten (2011) and Dong et al. (2016) explore extensions of discriminant analysis tailored for sequencing data using zero-inflated Poisson and negative binomial models, respectively, our findings reveal that the skewness prevalent in real data often exceeds the tolerance of these models. Consequently, this leads to suboptimal classification accuracy, compelling the need for data transformations. However, it is important to note that while applied transformations effectively address distributional violations, they may inadvertently distort the original information (McKnight et al., 2019; Vandeputte et al., 2017; Lloréns-Rico et al., 2021).

In this work, we simultaneously address the issues of high dimensionality, skewness, and zero inflation by proposing a new transformation-robust semiparametric binary classification framework via the truncated latent Gaussian copula model (Yoon et al., 2020). The model accounts for zeros that are not necessarily absolute while simultaneously accommodating extreme skewness. Latent Gaussian copula models provide an elegant framework for the analysis of non-Gaussian data of (possibly) mixed types, such as skewed continuous (Liu et al., 2009), binary (Fan et al., 2017), ordinal (Quan et al., 2018; Feng and Ning, 2019), and zero-inflated (Yoon et al., 2020). These models capture dependencies among variables via the latent correlation matrix, which can be consistently estimated by inverting a bridge function (Fan et al., 2017; Quan et al., 2018; Yoon et al., 2020) that connects latent correlations to the rank-based association measure, Kendall's τ. As such, the estimation is invariant to monotone transformations of the data. Subsequently, these models have been used for graphical model estimation (Fan et al., 2017; Feng and Ning, 2019; Yoon et al., 2019; Chung et al., 2022) and canonical correlation analysis (Yoon et al., 2020) with non-Gaussian data. However, the use of latent Gaussian copulas in the classification context has been limited. The existing approaches (Lin and Jeon, 2003; Han et al., 2013; Mai and Zou, 2015) are restricted to continuous data type, treating zeros as absolute. Furthermore, since the linear discriminant analysis model assumes class-specific means and common covariance matrix, but the means and variances are not identifiable under the Gaussian copula, Han et al. (2013) and Lin and Jeon (2003) impose additional constraints. In particular, Han et al. (2013) assume that the marginal transformations are mean- and variance-preserving. This assumption requires the methods to rely on observation-level moment estimates, which are sensitive to zero inflation and extreme values at the right tail of the distribution. We confirm this empirically in Section 5, where we show that the performance of Han et al. (2013) is highly dependent on data transformations, and is negatively affected by extreme skewness.

There are several major difficulties in adopting the truncated Gaussian copula model for discriminant analysis. The first difficulty is the identifiability of mean and variance parameters as described above, since the copula model is invariant under shifting and scaling of the bijective marginal transformations. Second, the aforementioned unsupervised problems, graphical model estimation and canonical correlation analysis, only require a consistent estimator of the latent correlation matrix. In contrast, discriminant analysis encompasses both the estimation of the classification direction (at the latent Gaussian level) and the prediction of class labels on new data (at the observed non-Gaussian level), thus requiring mapping of observed data to the latent level. For continuous-type data, accomplishing this task involves estimating the mapping from the observation level to the latent Gaussian level (Han et al., 2013; Lin and Jeon, 2003). However, for zero-inflated data, this becomes particularly challenging as the mapping from observed zeros to underlying latent Gaussian variables is not one-to-one. Thus, a new methodology is required to establish a valid classification rule under the model, and a substantially different theoretical analysis is required to establish consistency.

To address these limitations, we propose to consider a joint binary-truncated mixed copula model, where the class label is treated as a dichotomized latent Gaussian variable, and zero-inflated covariates follow the truncated Gaussian copula. The latent continuous representation of class labels is realistic in many contexts where responses are continuous in nature, even though the observed outcomes are discrete. For example, a person's latent degree of inflammation is continuous, but the observed disease status is binary, depending on whether the degree passes a threshold of the immune system. Similarly, there is significant heterogeneity within cancer subtypes, with classifications typically based on the degree of similarity to a "prototypical subtype" rather than a strictly categorical assignment. For instance, in our second illustrative application in Section 5, the "Basal-like" breast cancer subtype is classified based on the latent similarity of a genetic profile characterized by the continuous expression of specific biomarkers (Rakha et al., 2007). At the same time, the proposed joint framework encapsulates all relationships between the class label and covariates via the joint latent correlation matrix. As a result, unlike Lin and Jeon (2003) and Han et al. (2013), our approach does not require additional conditions on the marginal transformations, and consequently, we do not rely on observation-level moment estimates. We obtain analytic approximations of posterior class probabilities under the proposed joint model, and further demonstrate that the model has an equivalent conditional linear representation on the latent Gaussian level, with the vector of coefficients being a function of full joint correlation matrix, analogous to linear regression with random Gaussian design. Since the optimal classification direction depends only on the joint latent correlation matrix, our approach is fully invariant to monotone transformations of the data. Furthermore, by adapting ℓ_1-regularization, we prove that this classification direction can be consistently estimated in high-dimensional settings with the same rate as in sparse linear regression (Bickel et al., 2009). This is a highly non-trivial result as the direct application of the element-wise consistency of rank-based estimator of joint latent correlation matrix (Yoon et al., 2020) leads to a suboptimal rate. To our knowledge, the closest result is obtained by Barber and Kolar (2018), but the proof is restricted to continuous Gaussian copula as it relies on the closed form of the inverse bridge function in that setting. In contrast, for the truncated model (Yoon et al., 2020), the bridge function is significantly more complex, with no closed-form expression for its inverse. This poses new challenges for theoretical analyses, leading us to develop a different proof technique based on newly established bounds on the first and second derivatives of the inverse bridge function. Further, we propose an estimation procedure for posterior class probabilities conditional on observed measurements (zeros and non-zeros) based on Monte Carlo draws from a multivariate truncated normal distribution, which allows us to overcome the difficulty of predicting classes based on observed non-Gaussian data. Finally, we derive a Taylor approximation of the posterior probabilities, leading to a simple classification rule where the latent measurements corresponding to zeros are substituted with their conditional expectations.

In summary, from a methodological perspective, the primary novel aspects of our work are: (a) a novel framework for discriminant analysis based on truncated latent Gaussian copula models for skewed and zero-inflated data with analytic expressions of posterior probabilities under Bayes rule; (b) a principled approach for estimation of posterior probabilities using Monte Carlo methods to approximate the analytically intractable functionals of multivariate truncated normal distribution; (c) theoretical guarantees for consistency of estimated classification direction in high-dimensional setting with novel proof techniques that bound analytically intractable 2nd derivatives of inverse bridge functions, opening the path for establishing consistency in high-dimensional settings with general semiparametric Gaussian copula regression models (Dey and Zipunnikov, 2022).

From the application perspective, the numerical results on simulated and real data consistently convey that the proposed SEmiparametric Discriminant Analysis (SEDA) method is: 1) always the best-performing method on highly-skewed and zero-inflated data, with a significant margin of error improvement compared to existing approaches; 2) the performance of other methods can be significantly improved with data transformations that mitigate skewness, but the resulting misclassification errors and selected variables are dependent on the transformation choice; 3) the proposed SEDA maintains competitive or better accuracy even when accounting for transformations while being consistent in selected variables, facilitating the robustness and reproducibility of analyses.

2. Methodology

2.1 Notation

For a vector $a \in \mathbb{R}^p$, we denote the ℓ_q-norm, $q \in [0, \infty)$, by $\|a\|_q = (\sum_{j=1}^{p} |a_j|^q)^{1/q}$ and the ℓ_∞-norm by $\|a\|_\infty = \max_{1 \leq j \leq p} |a_j|$. For two vectors of the same size, $a, b \in \mathbb{R}^p$, we write $a < b$ to denote element-wise inequalities such that $a_j < b_j$ $(j = 1, \ldots, p)$. The vectors $1_p, 0_p \in \mathbb{R}^p$ denote the one and zero vectors and matrices $I_p, 1_{pp} \in \mathbb{R}^{p \times p}$ denote the identity and matrix with ones. For a matrix $A \in \mathbb{R}^{n \times p}$, $\|A\|_\infty = \max_{jk} |a_{jk}|$ denotes its ℓ_∞-norm, and for a square matrix $T \in \mathbb{R}^{p \times p}$, $|T|$ denotes its determinant and $\lambda_{\max}(T)$ and $\lambda_{\min}(T)$ denote the largest and smallest eigenvalues of T. For two functions f and g, we denote their composite function by $f \circ g = f(g(x))$. We let $1(\cdot)$ denote the indicator function taking the value 1 when its argument is true and 0 otherwise. For a sequence of random variables, $X_1, \ldots, X_n, \ldots$, we write $X_n = O_p(a_n)$ if, for any $\varepsilon > 0$, there exist $M, N > 0$ such that $\Pr(|X_n/a_n| > M) < \varepsilon$ for all $n > N$. We let $\Phi_d(a_1, \ldots, a_d; \Sigma)$ and $\Phi(\cdot)$ denote the d-dimensional Gaussian distribution function with zero mean and correlation matrix Σ evaluated at $(a_1, \ldots, a_d)^\top \in \mathbb{R}^d$ and the univariate standard Gaussian distribution function, respectively. We use C and C_i, $i = 1, 2, \ldots$, to denote generic constants that do not depend on the sample size n, dimension p, and model parameters. The cardinality of a set $\mathcal{S}$ is denoted by $\mathrm{card}(\mathcal{S})$.

2.2 Model

Let $Y \in \{0, 1\}$ be a random variable corresponding to class label and $X = (X_1, \ldots, X_p)^\top \in \mathbb{R}^p$ be a random vector of covariates. To accommodate possibly skewed and zero-inflated X, we propose to model X using truncated latent Gaussian copula of Yoon et al. (2020). We first review the standard Gaussian copula model, also known as the nonparanormal model (Liu et al., 2009).

Definition 1 (Gaussian copula model) *A random vector $X \in \mathbb{R}^p$ satisfies the Gaussian copula model if there exist strictly increasing transformations $f = \{f_j\}_{j=1}^p$ such that $(Z_1, \ldots, Z_p)^\top = \{f_1(X_1), \ldots, f_p(X_p)\}^\top \sim \mathrm{N}_p(0, \Sigma)$, where Σ is a correlation matrix. We write $X \sim \mathrm{NPN}_p(0, \Sigma, f)$.*

While the Gaussian copula model can accommodate skewness through transformation functions $\{f_j\}_{j=1}^p$, it does not allow zero-inflated variables. The model of Yoon et al. (2020) allows for both zero inflation and skewness through the following extra truncation step.

Definition 2 (Truncated latent Gaussian copula model) *A random vector $X \in \mathbb{R}^p$ satisfies the truncated latent Gaussian copula model if there exist a random vector $X^* \sim \mathrm{NPN}_p(0, \Sigma, f)$ and constants $D_j > 0$, $j = 1, \ldots, p$, such that $X_j = 1(X_j^* > D_j)X_j^*$.*

Combining the truncated Gaussian copula model for X with the latent Gaussian copula model for binary variable (Fan et al., 2017) leads to the joint model for the class label and covariates.

Definition 3 (Binary-truncated mixed latent Gaussian copula model) *A random vector $(Y, X^\top)^\top \in \{0, 1\} \times \mathbb{R}^p$ satisfies the binary-truncated mixed latent Gaussian copula* *model if there exists a random vector* $(X_y^*, X^{*\top})^\top \sim \mathrm{NPN}_{1+p}(0, \Sigma, f)$ *and constants* D_y, $D_j > 0$, $j = 1, \dots, p$, *such that*

$$Y = 1(X_y^* > D_y), \quad X_j = 1(X_j^* > D_j)X_j^*, \quad j = 1, \dots, p. \tag{1}$$

In Model (1), the binary response Y is a dichotomized version of a latent continuous variable X_y^*. This latent variable serves as a technical device to derive an analytic expression for the conditional probability $\Pr(Y = 1 \mid X = x)$ in the presence of zero-inflated covariates as we demonstrate later in (2). Latent continuous representations have been widely used to aid understanding of discrete response regression models and improve computational efficiency (Gelman et al., 2013, Ch. 16.2). In some applications, the latent representation has an intuitive interpretation. For example, a person's latent degree of inflammation is continuous, but the observed disease status is binary, depending on whether the degree passes a threshold of the immune system or not. Similarly, the "Basal-like" breast cancer subtype is classified based on the latent similarity of a genetic profile characterized by the continuous expression of specific biomarkers (Rakha et al., 2007). While an intuitive interpretation is not always available, we use the latent continuous representation to facilitate probability modeling (Cox, 2018, Ch. 1.3), and it is not our goal to draw inference about the latent variables themselves. Additional examples of the equivalence between conditional models for discrete $Y|X$ and corresponding latent continuous representations can be found for the logistic regression (Holmes and Knorr-Held, 2003; Kinney and Dunson, 2007; Ma et al., 2022; O'brien and Dunson, 2004), probit regression (Albert and Chib, 1993; Chib and Greenberg, 1998; Fasano and Durante, 2022), and linear discriminant analysis (Hastie et al., 2009, Ch. 4.4).

The Bayes classification rule assigns a new observation X to class 1 if $\Pr(Y = 1 \mid X) > \Pr(Y = 0 \mid X)$, and to class 0, otherwise. When $D_j = -\infty$ (no truncation) and f_j are identities (no X transformation), the conditional model $Y|X$ of (1) reduces to the standard probit regression model as we demonstrate below. In this case, a standard linear model representation of Gaussian X_y^* as a function of Gaussian X holds, with the vector of coefficients being a function of the full correlation matrix Σ. Further we derive the explicit form of the Bayes classification rule in the general truncated case, connecting it to the probit model based on latent Gaussian Z_y^* (for response) and Z (for features).

2.3 Classification Rule

We first consider the special case of model (1) with $D_j = -\infty$, $j = 1, \dots, p$; that is, X follows standard Gaussian copula (without truncation). By definition, there exists a latent Gaussian vector $(Z_y, Z_1, \dots, Z_p)^\top \sim \mathrm{N}_{1+p}(0, \Sigma)$ such that $Y = 1(X_y^* > D_y) = 1(Z_y > \Delta_y)$, $\Delta_y = f_y(D_y)$, and $X_j = f_j^{-1}(Z_j)$. Let $Z = (Z_1, \dots, Z_p)^\top$. Since f_j's are strictly increasing, conditional on X, we have the following probit regression model:

$$\begin{aligned}\Pr(Y = 1 \mid X) = \Pr(Z_y > \Delta_y \mid Z) &= \Pr\left(\frac{Z_y - \beta^{*\top} Z}{v} > \frac{\Delta_y - \beta^{*\top} Z}{v} \;\middle|\; Z\right) \\ &= \Phi\left(\frac{\beta^{*\top} Z - \Delta_y}{v}\right),\end{aligned}$$

where $\beta^* = \Sigma_{22}^{-1}\Sigma_{21}$, $v^2 = 1 - \Sigma_{21}^\top\Sigma_{22}^{-1}\Sigma_{21}$, $\Sigma_{21}^\top = \mathrm{cov}(Z_y, Z)$, and $\Sigma_{22} = \mathrm{cov}(Z)$. Observe that the above representation is based on the alternative equivalent formulation of joint distribution $(Z_y, Z_1, \ldots, Z_p)^\top \sim \mathrm{N}_{1+p}(0, \Sigma)$ through the hierarchical conditional representation $Z_y|Z \sim \mathrm{N}(\beta^{*\top}Z, v^2)$, where $Z \sim \mathrm{N}_p(0, \Sigma_{22})$. The latter is equivalent to a standard linear regression model with a random Gaussian design. Thus, under model (1), the effect of Z on Z_y is linear and determined by the correlation structure between Z_y and Z , the same as in the Gaussian copula regression with continuous response (Cai and Zhang, 2018). Accordingly, we have the linear Bayes classifier $\delta(X) = 1\{\beta^{*\top}f(X) - \Delta_y > 0\}$, where $f(X) = \{f_1(X_1), \ldots, f_p(X_p)\}^\top$.

We now consider the general truncated case, where $X_j = 1(Z_j > \Delta_j)f_j^{-1}(Z_j)$ with $\Delta_j = f_j(D_j)$, $j = 1, \ldots, p$. While the same linear relationship holds between latent Z_y and Z as demonstrated above, the posterior probability expression is more complex due to a lack of one-to-one mapping between observed zeros and latent Z. We next derive the full expression of posterior probability under this more challenging setting, which subsequently allows us to derive tractable analytic approximations of $\Pr(Y = 1 \mid X)$ for estimation.

For a given vector of covariates $X \in \mathbb{R}^p$, let $X_t \in \mathbb{R}^{p_t}$ and $X_o \in \mathbb{R}^{p_o}$ be the subvectors with truncated and observed realizations, respectively, where $p_t + p_o = p$. Likewise, let Z_t and Z_o be the corresponding latent Gaussian vectors, and Δ_t and Δ_o be the corresponding threshold vectors. Then it follows that $\Pr(Y = 1 \mid X) = \Pr(Z_y > \Delta_y \mid Z_o, Z_t < \Delta_t)$. Since $Z_y \mid Z \sim \mathrm{N}(\beta^{*\top}Z,\ v^2)$ as before, the posterior probability can be expressed by integrating out Z_y from the model:

$$\begin{aligned}\Pr(Y = 1 \mid X = x) &= \Pr(Y = 1 \mid X_o = x_o, X_t = 0_{p_t}) = \Pr(Z_y > \Delta_y \mid Z_o = z_o, Z_t < \Delta_t)\\ &= \{\Pr(Z_t < \Delta_t \mid Z_o = z_o)\}^{-1}\int_{\Delta_y}^{\infty}\int_{z_t<\Delta_t} p(z_y, z_t \mid z_o)\mathrm{d}z_t\mathrm{d}z_y\\ &= \{\Pr(Z_t < \Delta_t \mid Z_o = z_o)\}^{-1}\int_{z_t<\Delta_t}\Big\{\int_{\Delta_y}^{\infty} p(z_y \mid z_t, z_o)\mathrm{d}z_y\Big\}p(z_t \mid z_o)\mathrm{d}z_t\\ &= \{\Pr(Z_t < \Delta_t \mid Z_o = z_o)\}^{-1}\int_{z_t<\Delta_t}\Phi\Big(\frac{\beta_t^{*\top}z_t + \beta_o^{*\top}z_o - \Delta_y}{v}\Big)p(z_t \mid z_o)\mathrm{d}z_t\\ &= \mathrm{E}\left\{\Phi\Big(\frac{\beta_t^{*\top}Z_t + \beta_o^{*\top}z_o - \Delta_y}{v}\Big) \mid Z_o = z_o, Z_t < \Delta_t\right\},\end{aligned} \tag{2}$$

where β_t^* and β_o^* are the subvectors of β^* corresponding to the components of Z_t and Z_o, respectively, and the expectation (2) is over the multivariate Gaussian distribution of Z_t given $Z_o = z_o$ truncated to the region $\{a \in \mathbb{R}^{p_t} \mid a < \Delta_t\}$. The Bayes classifier under model (1) relies solely on the conditional model (2), assigning a new observation X to class 1 if the expectation in (2) exceeds 0.5, and to class 0 otherwise.

The conditional expectation in (2) does not admit a closed form, and we consider two approaches for its approximation. First, we can use Monte Carlo sampling. Let $\{z_t^{(s)}\}_{s=1}^S$ be an independent sample of size S from the p_t-variate truncated Gaussian, then

$$\mathrm{E}\left\{\Phi\Big(\frac{\beta_t^{*\top}Z_t + \beta_o^{*\top}z_o - \Delta_y}{v}\Big) \mid Z_o = z_o, Z_t < \Delta_t\right\} \approx \frac{1}{S}\sum_{s=1}^{S}\Phi\Big(\frac{\beta_t^{*\top}z_t^{(s)} + \beta_o^{*\top}z_o - \Delta_y}{v}\Big). \tag{3}$$

This approach, however, is computationally demanding and makes the classification rule dependent on the scalar v (recall that the classification rules of probit and standard Gaussian copula model do not depend on v). Secondly, we consider the first-order Taylor approximation of (2) around the mean $\mu_t = \mathrm{E}(Z_t \mid Z_o = z_o, Z_t < \Delta_t)$, which leads to

$$\mathrm{E}\left\{\Phi\left(\frac{\beta_t^{*\top} Z_t + \beta_o^{*\top} z_o - \Delta_y}{v}\right) \mid Z_o = z_o, Z_t < \Delta_t\right\} \approx \Phi\left(\frac{\beta_t^{*\top}\mu_t + \beta_o^{*\top} z_o - \Delta_y}{v}\right), \qquad (4)$$

where the derivation is given in Appendix A. The mean of multivariate truncated Gaussian distribution μ_t still needs to be estimated with the Monte Carlo sample. However, the classification rule based on (4) is linear, $\delta(X) = 1\{\beta_t^{*\top}\mu_t + \beta_o^{*\top} f_o(X_o) - \Delta_y > 0\}$, so its main advantage over (3) is that it does not require multiple evaluations of the standard normal distribution function and the estimation of the scaling factor v.

2.4 Estimation of the Classification Direction

The Bayes classification rule under model (1) depends crucially on $\beta^* = \Sigma_{22}^{-1}\Sigma_{21} \in \mathbb{R}^p$, which we refer to as classification direction. The best linear unbiased predictor for Z_y, the latent Gaussian variable of the class label, is $\mathrm{E}(Z_y \mid Z) = \beta^{*\top} Z$, where β^* is the minimizer of the mean squared error criterion:

$$\beta^* = \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}}\ \mathrm{E}\left\{(Z_y - \beta^\top Z)^2\right\} = \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}}\ \left(\beta^\top\Sigma_{22}\beta - 2\beta^\top\Sigma_{12}\right). \qquad (5)$$

In practice, Σ_{12} and Σ_{22} need to be estimated from the data. However, as Z_y and Z are unobservable latent variables, Σ_{21} and Σ_{22} cannot be directly estimated using the sample. Instead, we propose to utilize rank-based estimators for Σ_{12} and Σ_{22} that take advantage of the bridge function connecting latent correlations to Kendall's τ values (Fan et al., 2017; Yoon et al., 2020). The advantage of this connection is that it enables consistent estimation of latent correlations based on ranks without requiring estimation of underlying monotone transformations f_j's.

Specifically, a strictly increasing bridge function G is defined such that $G(\Sigma_{jk}) = \mathrm{E}(\hat{\tau}_{jk}) = \tau_{jk}$, where Σ_{jk} is an element of the full correlation matrix Σ corresponding to Z_j and Z_k, τ_{jk} is the corresponding population Kendall's τ, and $\hat{\tau}_{jk}$ is the sample Kendall's τ. The sample Kendall's τ is defined as

$$\hat{\tau}_{jk} = \frac{2}{n(n-1)} \sum_{1\le i\le i'\le n} \mathrm{sign}(X_{ij} - X_{i'j})\mathrm{sign}(X_{ik} - X_{i'k}), \qquad (6)$$

where X_{ij} is the ith independent sample of X_j, n is the sample size. The specific form of the bridge function G depends on the type of observed variables. We are interested in binary-truncated (BT) pairs (correlations between the binary class label and zero-inflated variables) and truncated-truncated (TT) pairs (correlations between zero-inflated variables). The corresponding bridge functions G_{BT} and G_{TT} have the closed form expressions (Yoon et al., 2020):

$$\begin{aligned} G_{BT}(\Sigma_{jk}; \Delta_j, \Delta_k) =& 2\{1 - \Phi(\Delta_j)\}\Phi(\Delta_k) - 2\Phi_3(-\Delta_j, -\Delta_k, 0; \Sigma_{3a}) \\ & - 2\Phi_3(-\Delta_j, -\Delta_k, 0; \Sigma_{3b}), \\ G_{TT}(\Sigma_{jk}; \Delta_j, \Delta_k) =& -2\Phi_4(-\Delta_j, -\Delta_k, 0, 0; \Sigma_{4a}) + 2\Phi_4(-\Delta_j, -\Delta_k, 0, 0; \Sigma_{4b}), \end{aligned} \qquad (7)$$

with

$$\Sigma_{3a} = \begin{pmatrix} 1 & -r & 1/\sqrt{2} \\ -r & 1 & -r/\sqrt{2} \\ 1/\sqrt{2} & -r/\sqrt{2} & 1 \end{pmatrix}, \quad \Sigma_{3b} = \begin{pmatrix} 1 & 0 & -1/\sqrt{2} \\ 0 & 1 & -r/\sqrt{2} \\ -1/\sqrt{2} & -r/\sqrt{2} & 1 \end{pmatrix},$$

and

$$\Sigma_{4a} = \begin{pmatrix} 1 & 0 & 1/\sqrt{2} & -r/\sqrt{2} \\ 0 & 1 & -r/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -r/\sqrt{2} & 1 & -r \\ -r/\sqrt{2} & 1/\sqrt{2} & -r & 1 \end{pmatrix}, \quad \Sigma_{4b} = \begin{pmatrix} 1 & r & 1/\sqrt{2} & r/\sqrt{2} \\ r & 1 & r/\sqrt{2} & 1/\sqrt{2} \\ r/\sqrt{2} & 1/\sqrt{2} & 1 & r \\ 1/\sqrt{2} & r/\sqrt{2} & r & 1 \end{pmatrix},$$

where $r = \Sigma_{jk}$. The moment equation $G(\Sigma_{jk}; \Delta_j, \Delta_k) = \mathrm{E}(\hat{\tau}_{jk})$ and the strict monotonicity of G enable estimation of the latent correlation matrix Σ using the method of moments. We use the moment estimator $\hat{\Delta}_j = \Phi^{-1}(\hat{\pi}_j)$, where $\hat{\pi}_j = \sum_{i=1}^{n} 1(x_{ij} = 0)/n$ is the proportion of zeros in X_{ij}, $i = 1, \ldots, n$, leading to $\hat{\Sigma}_{jk} = G^{-1}(\hat{\tau}_{jk}; \hat{\Delta}_j, \hat{\Delta}_k)$. At the sample level, however, $\hat{\Sigma} = [\hat{\Sigma}_{jk}]_{1\leq j,k\leq p}$ is not guaranteed to be positive-definite (Fan et al., 2017; Yoon and Gaynanova, 2021). Yoon et al. (2020) propose to use $\tilde{\Sigma} = (1-\nu)\hat{\Sigma}_* + \nu I_p$ with a small positive constant ν to ensure the positive definiteness, where $\hat{\Sigma}_*$ is the projection of $\hat{\Sigma}$ on the cone of positive-semidefinite matrices. When $\nu = o\{(\log p/n)^{1/2}\}$, $\tilde{\Sigma}$ has the same consistency rates as $\hat{\Sigma}$. We refer to Corollary 2 in Fan et al. (2017) and Theorem 7 in Yoon et al. (2020). In our implementation, we set $\hat{\Sigma} = \tilde{\Sigma}$ using the R package mixedCCA (Yoon and Gaynanova, 2021), which uses the default value of $\nu = 0.01$.

Remark 4 *The sample Kendall's τ in* (6) *is also known as τ^a, and it ignores ties since the sign function is defined as* sign(0) = 0*. The closed-form derivations of bridge functions and estimation consistency results mentioned above apply to τ^a. To account for ties, one can consider an alternative Kendall's τ^b(Quan et al., 2018), but the updated bridge functions require significantly more complex derivations and do not admit closed-form expression, thus we restrict our focus to τ^a in this work.*

In summary, to estimate β^*, we propose to replace Σ_{21} and Σ_{22} in (5) with the corresponding rank-based estimators $\hat{\Sigma}_{21}$ and $\hat{\Sigma}_{22}$, respectively. In addition, we consider a ℓ_1–regularization to account for high dimensionality and to enhance the interpretability of the resulting classification rule. Specifically, we consider the following minimization problem:

$$\hat{\beta} = \mathrm{argmin}_{\beta\in\mathbb{R}^p} \left(\frac{1}{2}\beta^\top\hat{\Sigma}_{22}\beta - \beta^\top\hat{\Sigma}_{21} + \lambda\|\beta\|_1\right), \tag{8}$$

where $\lambda > 0$ is the tuning parameter that controls the sparsity level of $\hat{\beta}$. This convex optimization problem can be efficiently solved via the coordinate descent algorithm. In our numerical studies, we select the tuning parameter λ that yields the lowest misclassification rate through 5-fold cross-validation.

2.5 Estimation of the Classification Rule

Here we describe how to obtain the sample Bayes classification rule based on the optimal rule in Section 2.3 and estimated classification direction $\hat{\beta}$. The critical difficulty is that both approximations of the posterior probability (3) and (4) rely on the latent Z, which is unobservable. We first illustrate how to estimate Z_o, the subvector of Z corresponding to non-zero observed values in X. We then use the estimated Z_o to generate posterior samples of Z_t, the subvector of Z corresponding to zero values in X, for use in classification rule (3), and in computing the conditional mean for the classification rule (4).

Let $(Y_i, X_i^\top)^\top = (Y_i, X_{i1}, \dots, X_{ip})^\top \in \{0,1\} \times \mathbb{R}^p$, $i = 1, \dots, n$, be the ith sample from the latent Gaussian copula model for binary-truncated mixed data as in Definition 3. For each i, we write $X_{i,t}$ and $X_{i,o}$ to denote the truncated and observed subvectors of $X_i \in \mathbb{R}^p$, respectively. Similarly, we denote the truncated and observed subvectors of a new observation X^{new} by $X_t^{\text{new}} \in \mathbb{R}^{p_t}$ and $X_o^{\text{new}} \in \mathbb{R}^{p_o}$. We write $\hat{\Delta}_y$ and $\hat{\Delta} = (\hat{\Delta}_1, \dots, \hat{\Delta}_p)^\top$ to denote the estimated thresholds, where $\hat{\Delta}_y = \Phi^{-1}(\hat{\pi}_y)$ and $\hat{\Delta}_j = \Phi^{-1}(\hat{\pi}_j)$ with $\hat{\pi}_j = \sum_{i=1}^n 1(x_{ij} = 0)/n$, and $\hat{\pi}_y = \sum_{i=1}^n 1(y_i = 0)/n$.

To estimate z_o^{new} corresponding to observed x_o^{new}, recall from Definition 3 that $z_{j,o}^{\text{new}}$ is given by $z_{j,o}^{\text{new}} = f_j(x_{j,o}^{\text{new}}) = \Phi^{-1} \circ F_j(x_{j,o}^{\text{new}})$, where F_j is the marginal distribution function of the jth latent variable X_j^* (Liu et al., 2009). We propose to estimate F_j using the empirical cumulative distribution function, where we apply the winsorization similar to Han et al. (2013) to avoid $f_j(x_{j,o}^{\text{new}})$ being infinite. Based on observations $x_{1j}, \dots, x_{nj}$, we consider

$$\hat{F}_j(t; \delta_n, x_{1j}, \dots, x_{nj}) = W_j^{\delta_n}\Big\{\frac{1}{n}\sum_{i=1}^n 1(x_{ij} \le t)\Big\},$$

where

$$W_j^{\delta_n}(x) = \hat{\pi}_j 1(x \le \hat{\pi}_j) + x 1(\hat{\pi}_j < x \le 1 - \delta_n) + (1 - \delta_n) 1(1 - \delta_n < x).$$

In our numerical studies, we use $\delta_n = 1/(2n)$ as recommended by Han et al. (2013). Based on $\hat{F}_j$'s, we set $\hat{f}_j = \Phi^{-1} \circ \hat{F}_j$ and estimate z_o^{new} with $\hat{z}_{j,o}^{\text{new}} = \hat{f}_j(x_{j,o}^{\text{new}})$.

For prediction, the posterior probability (2) is estimated with sample versions of (3) and (4), respectively. Let $\hat{\Delta}_t$ is the subvector of $\hat{\Delta}$ corresponding to Z_t^{new}. We generate an independent sample of Z_t^{new}, $\{z_t^{\text{new}(s)}\}_{s=1}^S$, conditional on $Z_o^{\text{new}} = \hat{z}_o^{\text{new}}$ and $Z_t^{\text{new}} < \hat{\Delta}_t$ from the following multivariate truncated Gaussian distribution. Let $\text{var}(Z_t^{\text{new}}) = \Sigma_t$, $\text{var}(Z_o^{\text{new}}) = \Sigma_o$, and $\text{cov}(Z_o^{\text{new}}, Z_t^{\text{new}}) = \Sigma_{ot}$. By properties of the multivariate Gaussian distribution, $\text{E}(Z_t \mid Z_o) = \Sigma_{ot}^\top \Sigma_o^{-1} Z_o$ and $\text{var}(Z_t \mid Z_o) = \Sigma_t - \Sigma_{ot}^\top \Sigma_o^{-1} \Sigma_{ot}$. By plugging estimators $\hat{Z}_o^{\text{new}}$, $\hat{\Sigma}_t$, $\hat{\Sigma}_o$, $\hat{\Sigma}_{ot}$, and $\hat{\Delta}_t$, we obtain the multivariate truncated Gaussian distribution with the probability density

$$p(z_t \mid \hat{z}_o^{\text{new}}, z_t < \hat{\Delta}_t) = \frac{\phi_{p_t}(z_t \ ; \ \hat{\gamma}, \hat{\Gamma})}{\Pr(Z_t < \hat{\Delta}_t \mid Z_o = \hat{z}_o^{\text{new}})} 1(z_t < \hat{\Delta}_t), \tag{9}$$

where $\phi_p(z \ ; \ \gamma, \Gamma)$ denotes the p-dimensional Gaussian density with mean γ and covariance matrix Γ, $\hat{\gamma} = \hat{\Sigma}_{ot}^\top \hat{\Sigma}_o^{-1} \hat{z}_o^{\text{new}}$, and $\hat{\Gamma} = \hat{\Sigma}_t - \hat{\Sigma}_{ot}^\top \hat{\Sigma}_o^{-1} \hat{\Sigma}_{ot}$. Combining $\{z_t^{\text{new}(s)}\}_{s=1}^S$ with the estimates $\hat{\beta}$ from Section 2.4, $\hat{v} = (1 - \hat{\Sigma}_{21}^\top \hat{\Sigma}_{22}^{-1} \hat{\Sigma}_{21})^{1/2}$, and $\hat{\Delta}_y$, a sample plug-in version of the posterior probability (3) is given by

$$\widehat{\Pr}(Y = 1 \mid X^{\text{new}}) = S^{-1} \sum_{s=1}^{S} \Phi \left(\frac{\hat{\beta}_t^\top z_t^{\text{new}(s)} + \hat{\beta}_o^\top \hat{z}_o^{\text{new}} - \hat{\Delta}_y}{\hat{v}} \right). \tag{10}$$

Similarly, using the Monte Carlo estimate $\tilde{\mu}_t = S^{-1} \sum_{s=1}^{S} z_t^{\text{new}(s)}$ for $\mathrm{E}(Z_t \mid Z_o = \hat{z}_o^{\text{new}}, Z_t < \hat{\Delta}_t)$, we have a sample version of the posterior probability (4) as

$$\widehat{\Pr}(Y = 1 \mid X^{\text{new}}) = \Phi \left(\frac{\hat{\beta}_t^\top \tilde{\mu}_t + \hat{\beta}_o^\top \hat{z}_o^{\text{new}} - \hat{\Delta}_y}{\hat{v}} \right). \tag{11}$$

The corresponding sample Bayes rule assigns a new observation X^{new} to class 1 if the sample posterior probability is greater than 0.5 and to class 0, otherwise. While the expression (11) for the posterior probability depends on $\hat{v}$, the corresponding classification rule does not—similar to the standard linear discriminant analysis rule.

3. Theoretical Results

In this section, we demonstrate that the estimated classification direction $\hat{\beta}$ from (8) is a consistent estimator for β^* under the following assumptions.

Assumption 1 (Latent correlations) *All the off-diagonal elements of* Σ *satisfy* $|\Sigma_{jk}| \leq 1 - \varepsilon_r$ *for some constant* $\varepsilon_r \in (0, 1)$.

Assumption 2 (Thresholds) *All the thresholds* Δ_j *satisfy* $|\Delta_j| \leq M$ *for some constant* M.

Assumption 3 (Condition number) $\mathsf{C}(\Sigma) = \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)} \leq C_{\text{cov}}$ *for some constant* C_{cov}.

Assumption 4 (Sparsity) β^* *is sparse with the support* $\mathcal{S} = \{j : \beta_j^* \neq 0\}$ *with* $s = \mathrm{card}(\mathcal{S})$.

Assumption 5 (Sample size) $s \log p = o(n)$.

Assumptions 1-2 are needed to guarantee consistency of estimated latent correlations in $\hat{\Sigma}_{22}$ and $\hat{\Sigma}_{21}$. Under these assumptions, the level of zero inflation, represented by the sizes of thresholds Δ_j and Δ_k, affects the convergence rate at most by introducing a constant factor (Fan et al., 2017; Yoon et al., 2020). Assumptions 3–5 are used to account for the high-dimensional setting, where p is large, potentially much greater than n. We also take advantage of the restricted eigenvalue condition.

Definition 5 (Restricted eigenvalue condition) *A* $p \times p$ *matrix* Σ *satisfies the restricted eigenvalue condition* $RE(s, 3)$ *with parameter* $\gamma = \gamma(\Sigma)$ *if for all sets* $\mathcal{S} \subset \{1, \ldots, p\}$ *with* $\mathrm{card}(\mathcal{S}) \leq s$, *and for all* $a \in \mathcal{C}(\mathcal{S}, 3) = \{a \in \mathbb{R}^p : \|a_{\mathcal{S}^c}\|_1 \leq 3\|a_{\mathcal{S}}\|_1\}$, *it holds that*

$$a^\top \Sigma a \geq \gamma^{-1} \|a_{\mathcal{S}}\|_2^2.$$

First, we provide the deterministic bound on estimation error, a standard bound for high-dimensional sparse regression (Bickel et al., 2009; Hastie et al., 2015; Negahban et al., 2012). For completeness, the proof is presented in Appendix C.

Theorem 6 *Under Assumption 4, if $\lambda \geq 2\|\hat{\Sigma}_{21} - \hat{\Sigma}_{22}\beta^*\|_\infty$ and $\hat{\Sigma}_{22}$ satisfy RE(s,3) with parameter γ, then*

$$\|\hat{\beta} - \beta^*\|_2 \leq \frac{15}{2}\gamma\sqrt{s}\lambda.$$

To derive the probabilistic bound, we need to control the size of the tuning parameter, that is $\|\hat{\Sigma}_{21} - \hat{\Sigma}_{22}\beta^*\|_\infty$, and also ensure that the restricted eigenvalue condition on Σ implies the condition on $\hat{\Sigma}$. The existing results on the consistency of $\hat{\Sigma}$ (Yoon et al., 2020) provide the following high probability bounds: $\|\hat{\Sigma}_{21} - \Sigma_{21}\|_\infty \leq C\sqrt{\log p/n}$ and $\|\hat{\Sigma}_{22} - \Sigma_{22}\|_\infty \leq C\sqrt{\log p/n}$. A direct application of these results to control $\|\hat{\Sigma}_{21} - \hat{\Sigma}_{22}\beta^*\|_\infty$ gives

$$\begin{aligned}\|\hat{\Sigma}_{21} - \hat{\Sigma}_{22}\beta^*\|_\infty &\leq \|\hat{\Sigma}_{21} - \Sigma_{21}\|_\infty + \|(\Sigma_{22} - \hat{\Sigma}_{22})\beta^*\|_\infty \\ &\leq C\sqrt{\log p/n} + \|\hat{\Sigma}_{22} - \Sigma_{22}\|_\infty\|\beta^*\|_1 \leq C_1\sqrt{\log p/n}\|\beta^*\|_1.\end{aligned}$$

The above bound is sub-optimal as $\|\beta^*\|_1$ scales approximately $s^{1/2}$, implying that the knowledge of true sparsity level is required to choose the tuning parameter λ. In contrast, the results from sparse high-dimensional regression (Bickel et al., 2009; Hastie et al., 2015; Negahban et al., 2012) suggest that the optimal rate should be of the order $\sqrt{\log p/n}$ without the extra dependence on s. Our main theoretical contribution is obtaining the optimal bound for $\|\hat{\Sigma}_{21} - \hat{\Sigma}_{22}\beta^*\|_\infty$ under the model (1).

Theorem 7 *Under Assumptions 1–5, for any $\eta \in (0, 1)$, there exists some constant $C > 0$ such that*

$$\|\hat{\Sigma}_{21} - \hat{\Sigma}_{22}\beta^*\|_\infty \leq C\sqrt{\frac{\log(p\eta^{-1})}{n}}$$

with probability at least $1 - \eta$.

The full proof is presented in Appendix C, and here we summarize the argument at a high level. To our knowledge, the only similar result is obtained by Barber and Kolar (2018) in the case of continuous Gaussian copula of Definition 1. Their proof, however, takes advantage of the closed form of the inverse bridge function G^{-1} in the continuous case, which is a scaled cosine function. Due to the significantly higher complexity of the bridge function G_{TT} in (7) for the truncated Gaussian copula case, its inverse lacks a closed-form expression, which makes the proof more challenging. An additional complication arises from substituting true thresholds Δ_j with their estimators $\hat{\Delta}_j$, these thresholds being unique to the truncated case. To overcome these challenges, we consider the 2nd-order Taylor expansion of $\hat{\sigma}_{jk} = G_{TT}^{-1}(\hat{\tau}_{jk}, \hat{\Delta}_j, \hat{\Delta}_k)$ with respect to $\sigma_{jk} = G_{TT}^{-1}(\tau_{jk}, \Delta_j, \Delta_k)$. To control the first-order terms, we combine the bound on first derivatives of inverse bridge functions (Yoon et al., 2020) with the concentration bound for deviations of quadratic forms involving the Kendall's τ correlation matrix (Barber and Kolar, 2018)[Lemma E.2] and sign sub-Gaussian property of the Gaussian vectors (Barber and Kolar, 2018)[Lemma 4.5]. To control the second-order terms, we show that the second derivatives of inverse bridge functions are bounded and use these bounds in conjunction with element-wise convergence of $\hat{\Sigma}$ and $\hat{\Delta}$. Due to the inverse bridge function not being available in a closed form, establishing bounds on the second derivative is highly non-trivial and is a major technical part of the proof. A similar technique proves that the restricted eigenvalue condition on Σ implies the condition holds for $\hat{\Sigma}$ (Lemma A.6 in Appendix C.3), leading to our final estimation bound.

Theorem 8 *Under Assumptions 1–5, if* $\lambda = C\sqrt{\log p/n}$ *for some constant* $C > 0$ *and* Σ_{22} *satisfies RE(s,3) with parameter* γ*, then*

$$\|\hat{\beta} - \beta^*\|_2^2 = O_p\left(\gamma^2 \frac{s \log p}{n}\right).$$

The obtained rate in estimation error coincides with the optimal rate in sparse linear regression (Bickel et al., 2009; Hastie et al., 2015; Negahban et al., 2012). The developed technique can be applied to establish estimation consistency within the general semiparametric Gaussian copula regression framework (Dey and Zipunnikov, 2022), encompassing all continuous, binary, ordinal, and truncated variable types.

4. Simulation

We empirically evaluate the performance of the proposed method, which we name SEmiparametric Discriminant Analysis (SEDA), with two approaches for estimating class-conditional probabilities as described in Section 2.5: formula (10) based on the Monte Carlo approximation and formula (11) based on the Taylor approximation, where we set the size of Monte Carlo samples $S = 100$ from (9). Given a fixed tuning parameter λ, we solve (8) with a solver implemented in `C` in the `R` package `MGSDA` (Gaynanova, 2021). For comparison, we consider high-dimensional copula discriminant analysis (CODA) of Han et al. (2013), sparse semiparametric discriminant analysis (SSDA) of Mai and Zou (2015), negative binomial linear discriminant analysis (NBLDA) of Dong et al. (2016), classification and clustering of sequencing data using a Poisson model (PoiClaClu) of Witten (2011), random forest (RF) of Breiman (2001), sparse logistic regression (S-Logistic) of Friedman et al. (2010), and sparse support vector machine (S-SVM) of Yi and Huang (2017). Comprehensive implementation details are provided in Appendix B.1.

To generate synthetic data, we fix the number of covariates $p = 300$ and consider three correlation structures for the latent Gaussian vector associated with the covariates:

(1) Autoregressive (AR), $\Sigma_{22} = [0.7^{|j-j'|}]_{1\leq j,j'\leq p}$.

(2) Compound symmetry (CS), $\Sigma_{22} = 0.7I_p + 0.31_{pp}$; where I_p and 1_{pp} are the $p \times p$ identity matrix and matrix of ones, respectively.

(3) Geometric decaying eigenvalues (GD): $\Sigma_{22} = \Gamma N \Gamma^\top$ where Γ is generated from the uniform distribution on p-dimensional orthogonal group (Chikuse, 2003, Theorem

2.2.1) and N is a diagonal matrix with geometrically decaying eigenvalues $\nu_1 > \nu_2 > \cdots > \nu_p$, where

$$\nu_j = \frac{p(0.9^{j-1} - 0.9^j)}{(1 - 0.9^p)}, \quad j = 1, \ldots, p.$$

In the AR setting, each variable is strongly positively correlated with only a few variables, whereas in the CS setting, all variables are moderately positively correlated. In the GD setting, the correlation structure mimics the high-dimensional real data (Lee et al., 2013), with a wide range of correlations from -0.7 to 0.7.

We consider two model settings: joint model (as in Definition 3) and mixture model (proposed in Han et al. (2013), thus favoring CODA). For both models, we use the quantitative microbiome profiling data of Vandeputte et al. (2017) as a reference to mimic the real-world data. The true classification direction vector β^* is set so that only the first $s = 0.05p = 15$ variables are non-zero. For the joint model, we consider three levels of zero-inflation (truncation): no truncation (0% zeros), low truncation (10%–50% zeros in each variable) and high truncation (50%–80% zeros in each variable). We use the empirical cumulative distribution functions of the corresponding reference variables to generate $p = 300$ covariates (since the number of reference variables is fewer than 300, we recycle the empirical cumulative distribution functions to generate multiple synthetic variables).

For the mixture model, we exclusively consider the non-truncation case for a fair comparison with CODA. As a benchmark for classification accuracy, we define an Oracle rule for each model, using the Bayes plug-in rule that incorporates the true β^* while estimating the underlying transformations as in (A1) and (A3). Comprehensive details regarding the data generation mechanism employed for each model and the oracle rules are contained in Appendix B."

For each model and correlation structure, we consider equally (50:50) and unequally (20:80) proportioned class sizes and fix the sample sizes of training and test data at $n = 150$ and $n_{\text{test}} = 300$, respectively. We consider 100 replications for each population setting, correlation structure, and truncation level (for the joint setting).

To assess the prediction performance, we evaluate average misclassification rates on test data, which are reported in Table 1. For SEDA, we found that both approximations (10) and (11) of the conditional class probability led to practically the same misclassification rates, and thus, we only report the results for (11) (the results for (10) are in Appendix Table A1). When compared to parametric models in joint settings (NBLDA, PoiClaClu, S-Logistic, S-SVM), the proposed SEDA performs significantly better, especially with the increase in levels of zero inflation, confirming that existing models struggle to simultaneously address skewness and zero-inflation. Compared with semiparametric CODA, the proposed SEDA still performs significantly better, even under the population model settings favorable to CODA (i.e., the mixture setting without zero inflation). We suspect this is due to extreme skewness in simulated data, which affects the quality of observation level moment estimates that CODA relies on. SSDA, another semiparametric approach, has better performance than CODA and is comparable to SEDA in the mixture setting with no zero inflation, but is worse than SEDA in the joint setting as the proportion of zeros increases. When compared to the fully nonparametric random forest, SEDA has competitive performance in the equal class size case and is significantly better in the unequal class size case.

Joint	Oracle	CODA	NBLDA	PoiClaClu	RF	S-Logistic	SSDA	S-SVM	SEDA_L
(50:50)				No truncation					
AR	7.9 (0.2)	17.5 (0.3)	19.5 (0.3)	16.0 (0.3)	14.0 (0.2)	15.3 (0.2)	13.4 (0.3)	21.9 (0.8)	13.0 (0.3)
CS	7.6 (0.2)	17.2 (0.3)	22.7 (0.3)	20.4 (0.3)	13.7 (0.2)	16.2 (0.2)	15.7 (0.3)	15.0 (0.2)	14.8 (0.3)
GD	8.7 (0.2)	16.7 (0.3)	22.0 (0.4)	22.3 (0.4)	20.3 (0.3)	19.6 (0.4)	17.2 (0.3)	21.7 (0.4)	16.4 (0.3)
				Low truncation					
AR	8.8 (0.2)	23.9 (0.3)	31.1 (0.4)	19.8 (0.3)	14.9 (0.2)	23.4 (0.3)	17.0 (0.3)	33.8 (0.6)	15.0 (0.3)
CS	8.3 (0.2)	19.1 (0.3)	27.2 (0.3)	22.5 (0.5)	14.0 (0.2)	19.3 (0.2)	17.2 (0.3)	16.5 (0.3)	14.9 (0.3)
GD	8.3 (0.2)	22.9 (0.4)	24.7 (0.4)	22.7 (0.4)	20.5 (0.3)	25.9 (0.5)	19.9 (0.3)	30.5 (0.6)	17.0 (0.3)
				High truncation					
AR	11.0 (0.2)	36.1 (0.6)	33.5 (0.3)	17.5 (0.3)	16.7 (0.2)	28.8 (0.3)	20.1 (0.4)	37.5 (0.6)	18.5 (0.3)
CS	9.6 (0.2)	32.9 (0.6)	32.0 (0.4)	29.8 (0.5)	14.3 (0.2)	20.9 (0.2)	18.8 (0.3)	17.1 (0.3)	15.8 (0.2)
GD	8.6 (0.2)	29.4 (0.6)	25.1 (0.4)	23.0 (0.4)	21.0 (0.4)	30.3 (0.7)	22.6 (0.4)	32.7 (0.6)	18.2 (0.3)
(20:80)				No truncation					
AR	5.5 (0.1)	14.6 (0.2)	15.8 (0.3)	17.2 (0.3)	17.4 (0.2)	16.3 (0.3)	10.9 (0.2)	20.0 (0.2)	9.8 (0.2)
CS	5.3 (0.1)	11.9 (0.2)	16.1 (0.2)	16.5 (0.4)	11.3 (0.1)	13.0 (0.2)	12.8 (0.2)	10.2 (0.2)	11.0 (0.2)
GD	6.1 (0.1)	13.3 (0.2)	18.6 (0.3)	26.4 (0.5)	16.4 (0.2)	16.7 (0.3)	13.7 (0.2)	16.1 (0.2)	12.6 (0.2)
				Low truncation					
AR	6.8 (0.1)	22.0 (1.1)	19.1 (0.2)	19.3 (0.5)	18.8 (0.2)	19.9 (0.2)	14.3 (0.2)	20.2 (0.2)	12.3 (0.2)
CS	6.4 (0.1)	14.6 (0.2)	16.8 (0.2)	13.6 (0.4)	10.8 (0.1)	15.1 (0.3)	12.9 (0.2)	11.1 (0.2)	11.0 (0.2)
GD	6.0 (0.1)	16.7 (0.3)	17.3 (0.3)	27.5 (0.6)	16.8 (0.2)	19.2 (0.2)	14.9 (0.2)	20.1 (0.4)	12.9 (0.2)
				High truncation					
AR	9.5 (0.2)	21.2 (0.6)	19.9 (0.3)	23.2 (0.4)	19.7 (0.2)	19.8 (0.2)	20.3 (0.2)	20.0 (0.2)	16.7 (0.2)
CS	7.6 (0.1)	21.5 (0.9)	17.1 (0.2)	27.3 (1.2)	11.2 (0.2)	16.7 (0.3)	16.5 (0.2)	12.7 (0.3)	11.7 (0.2)
GD	6.5 (0.1)	21.4 (1.1)	17.0 (0.3)	26.7 (0.4)	17.3 (0.2)	19.6 (0.2)	17.7 (0.3)	20.1 (0.3)	13.6 (0.2)

Mixture	Oracle	CODA	NBLDA	PoiClaClu	RF	S-Logistic	SSDA	S-SVM	SEDA_L
(50:50)				No truncation					
AR	10.0 (0.2)	15.3 (0.3)	41.4 (1.0)	41.5 (1.0)	11.7 (0.2)	11.6 (0.2)	12.5 (0.3)	12.8 (0.4)	12.2 (0.2)
CS	10.2 (0.2)	14.9 (0.3)	15.9 (0.2)	16.7 (0.3)	13.1 (0.2)	12.0 (0.2)	13.1 (0.3)	14.2 (0.2)	12.9 (0.2)
GD	9.8 (0.2)	13.9 (0.2)	21.6 (0.4)	22.0 (0.4)	13.5 (0.2)	12.4 (0.2)	13.1 (0.3)	14.3 (0.3)	13.2 (0.3)
(20:80)				No truncation					
AR	7.3 (0.1)	12.6 (0.3)	41.5 (0.4)	44.9 (0.5)	11.6 (0.2)	10.7 (0.2)	9.5 (0.2)	11.3 (0.2)	9.1 (0.2)
CS	7.3 (0.1)	11.7 (0.2)	15.3 (0.3)	15.9 (0.3)	9.6 (0.2)	11.2 (0.2)	10.7 (0.3)	9.9 (0.2)	9.8 (0.2)
GD	7.2 (0.2)	11.0 (0.2)	21.1 (0.4)	22.4 (0.5)	11.5 (0.2)	11.5 (0.3)	10.3 (0.2)	10.8 (0.2)	10.0 (0.2)

Table 1: Average misclassification rates (%) for simulated data based on 100 replications, with standard errors in parentheses

To assess the variable selection performance, we calculate the Matthews correlation coefficient (Matthews, 1975). A larger value represents a better variable selection performance, with the two boundary values, 1 and -1, indicating the completely correct and incorrect variable selections, respectively. Average MCC values across 100 replications for each method except random forest (which does not provide variable selection) are provided in Table A2. The proposed SEDA performs the best in almost all settings and is worse by the magnitude of at most 0.06 when it's second best. Overall, the proposed SEDA obtains the best or second-best results in both classification accuracy and model selection performance and is the best overall performing method when considering both metrics.

5. Application to Sequencing Data

We assess the classification performance of SEDA and competing methods on three sequencing data sets: the Quantitative Microbiome Profiling (QMP) data of Vandeputte et al. (2017), microRNA data from breast cancer patients available through The Cancer Genome Atlas Project (Cancer Genome Atlas Network, 2012), and single-cell RNA (scRNA) sequencing data from the 10x Genomics website (`https://www.10xgenomics.com`).

The QMP data are not compositional but quantitative (Vandeputte et al., 2017), and we use the data set processed as in Yoon et al. (2019). The QMP data contain $p = 101$ genera from $n_0 = 29$ patients with Crohn's disease ($Y = 0$) and $n_1 = 106$ healthy controls ($Y = 1$). The proportions of zeros across all 101 genera range from 1% to 80%, with 14 genera having no zeros. As the data are heavily skewed, we also consider two popularly used transformations: log transformation after adding the pseudo count one (log-transformed) and modified centered log-ratio transformation (Yoon et al., 2019) (mclr-transformed). Unlike log transformation, mclr transformation is not monotone; thus, we expect the results of all methods, including the proposed SEDA, to change. The mclr transformation is designed for compositional rather than quantitative microbiome data (Yoon et al., 2019) and can be viewed as evaluating the robustness of the results to the total per-sample count. Our goal is to construct a classification rule that can separate patients with Crohn's disease from healthy controls and to identify key genera that influence the rule.

The microRNA breast cancer data contain abundances of $p = 423$ microRNAs from $n = 348$ breast cancer patients, where $n_0 = 66$ patients have Basal-like tumor type ($Y = 0$) and $n_1 = 282$ patients have other sub-types ($Y = 1$). The proportions of zeros across all 423 variables range from 0.3% to 49%, with 206 variables having no zeros. As the data are skewed, we consider square-root and log transformations, where log transformation is applied after adding a pseudo-count of one. We aim to construct a classification rule that separates the Basal-like tumor type from other types.

The scRNA sequencing data contain $p = 329$ genes of $n_0 = 77$ CXCR4-negative centrocyte B-cells and $n_1 = 188$ immature B-Cells from the lymphoblastoid cell line (LCL) GM12878. The proportions of zeros across 329 genes range from 2% to 90%, with 6 genes having no zeros. As well as the original data, we consider two transformations: log transformation after adding the pseudo count one (log-transformed) and square-root transformation.

For each data set, we apply the same methods as in Section 4, using 100 random splits into training (4/5) and testing (1/5). Table 2 displays the average misclassification rates and model sizes. Moreover, we evaluate the stability of variable selection across data sets with distinct transformations utilizing the Jaccard index (Jaccard, 1912). The Jaccard index compares the size of shared elements, $\text{card}(\mathcal{S}_1 \cap \mathcal{S}_2)$, to the total unique elements, $\text{card}(\mathcal{S}_1 \cup \mathcal{S}_2)$, between two estimated signal sets, $\mathcal{S}_1$ and $\mathcal{S}_2$, from distinct transformed data sets. The index's scale ranges from zero (indicating an empty intersection) to one (representing complete agreement). For more than two sets, we employ the average pairwise Jaccard indices (Lang, 2022). For each fold, we compute the Jaccard index across transformed data sets and present the average values in Table 2.

Under the untransformed original data sets, the proposed method significantly outperforms all other methods having the lowest misclassification rates. Under the transformed data sets, the error rates of most other methods are notably improved. In contrast, the

	CODA	NBLDA	PoiClaClu	RF	S-Logistic	SSDA	S-SVM	SEDA$_L$
QMP (original)								
Error	10.25 (0.56)	12.96 (0.64)	4.32 (0.36)	5.79 (0.43)	14.00 (0.50)	7.57 (0.47)	10.46 (0.50)	2.93 (0.31)
Size	37.65 (1.22)	101 (0)	37.39 (2.18)	–	6.51 (0.31)	35.05 (1.77)	50.69 (1.27)	24.09 (1.51)
QMP (log-transformed)								
Error	2.18 (0.25)	3.79 (0.39)	3.64 (0.34)	5.75 (0.45)	3.93 (0.31)	7.57 (0.47)	6.79 (0.41)	2.93 (0.31)
Size	8.09 (0.43)	84.93 (2.22)	38.60 (1.54)	–	12.94 (0.37)	35.05 (1.77)	19.45 (2.12)	24.09 (1.51)
QMP (mclr-transformed)								
Error	2.75 (0.31)	3.54 (0.30)	3.57 (0.31)	4.32 (0.35)	2.21 (0.22)	3.61 (0.32)	4.36 (0.39)	2.79 (0.31)
Size	10.41 (1.12)	68.17 (2.70)	36.04 (2.32)	–	11.12 (0.33)	41.17 (2.40)	23.24 (1.36)	10.45 (0.95)
Stability	22.74 (0.70)	73.20 (1.51)	63.69 (1.44)	–	30.73 (0.53)	55.60 (1.10)	30.63 (1.52)	59.14 (1.47)
Breast cancer microRNA (original)								
Error	4.85 (0.23)	16.08 (0.49)	4.30 (0.24)	4.06 (0.19)	6.39 (0.26)	3.56 (0.20)	7.00 (0.25)	3.08 (0.20)
Size	52.92 (0.92)	423 (0)	385.48 (6.69)	–	20.51 (0.4)	70.03 (4.34)	44.05 (1.94)	86.34 (7.67)
Breast cancer microRNA (log-transformed)								
Error	3.17 (0.18)	5.97 (0.28)	5.97 (0.29)	3.93 (0.20)	3.58 (0.17)	3.56 (0.20)	3.23 (0.18)	3.08 (0.20)
Size	22.42 (3.28)	216.68 (20.78)	42.33 (7.37)	–	9.12 (0.34)	70.03 (4.34)	17.79 (1.37)	86.34 (7.67)
Breast cancer microRNA (square-root-transformed)								
Error	3.07 (0.19)	4.99 (0.24)	4.24 (0.24)	3.99 (0.20)	4.87 (0.21)	3.56 (0.20)	4.39 (0.21)	3.08 (0.20)
Size	50.74 (2.08)	419.06 (3.40)	387.27 (6.46)	–	13.78 (0.39)	70.03 (4.34)	31.29 (2.06)	86.34 (7.67)
Stability	20.41 (0.37)	67.22 (3.31)	39.59 (1.21)	–	40.55 (0.62)	100 (0)	31.57 (0.83)	100 (0)
B-cell scRNA (original)								
Error	7.29 (0.36)	10.71 (0.48)	6.77 (0.32)	4.21 (0.24)	5.73 (0.31)	6.13 (0.33)	12.02 (0.60)	3.50 (0.23)
Size	23.33 (0.74)	328.17 (0.32)	278.18 (4.29)	–	26.65 (0.33)	10.14 (0.23)	43.63 (1.89)	59.53 (4.21)
B-cell scRNA (log-transformed)								
Error	9.08 (0.34)	4.08 (0.22)	3.71 (0.22)	4.10 (0.24)	4.19 (0.25)	6.13 (0.33)	5.83 (0.29)	3.50 (0.23)
Size	10.41 (1.28)	149.42 (9.70)	164.89 (7.47)	–	26.72 (0.45)	10.14 (0.23)	40.23 (1.60)	59.53 (4.21)
B-cell scRNA (square-root-transformed)								
Error	7.65 (0.34)	5.88 (0.32)	5.94 (0.33)	4.08 (0.24)	4.63 (0.29)	6.13 (0.33)	6.12 (0.31)	3.50 (0.23)
Size	16.30 (0.55)	326.41 (1.82)	258.74 (5.36)	–	27.34 (0.41)	10.14 (0.23)	46.24 (1.97)	59.53 (4.21)
Stability	31.15 (4.32)	68.11 (8.27)	60.42 (4.58)	98.42 (0.24)	63.13 (2.80)	100 (0)	44.51 (4.00)	100 (0)

Table 2: Average misclassification rates (%), model sizes, and stability measured by the Jaccard indices (%) for real data sets based on 100 random splits, with standard errors in parentheses.

results of PoiClaClu, RF, SSDA, and the proposed method are not much affected by the transformations. In particular, the results of the proposed method and SSDA are not affected by the log and square root transformations at all, as the transformation preserves the rank of observed values. Specifically, the rank-based correlation estimator and the marginal transformations based on the empirical cumulative distribution function do not depend on the variables' scales but on the observations' rank. On the contrary, the mclr transformation changes the rank of the observations by applying observation-wise transformation. Hence, the error rate of the proposed method and SSDA on the mclr-transformed QMP data are slightly different from the original QMP data. However, SSDA shows overall inferior misclassification rates, particularly on QMP and scRNA data sets, likely because of the relatively higher zero-inflation compared to the microRNA data set. Although CODA employs the rank correlation and marginal transformation estimators as ours, its error rates differ substantially between original and transformed data sets. This sensitivity is possibly due to the moment matching constraints as the applied transformations dramatically change the first moments of the two classes, resulting in different CODA classifiers. Furthermore, CODA exhibits the lowest variable selection stability by selecting significantly different signal sets across various transformations.

On QMP data sets, PoiClaClu demonstrates relatively robust performance in class prediction and variable selection across diverse transformations. This robustness might stem from the inherent power transformation of data within the `R` package `PoiClaClu`, aimed at addressing overdispersion under the Poisson model (Witten, 2011). However, noticeable differences in variable selection between the original and log-transformed microRNA and scRNA data sets suggest sensitivity of model selection to potential misspecification.

NBLDA exhibits high variable selection stability, potentially attributed to its selection of a large number (or all) of variables. However, its classification error rates are larger than those of other methods and vary significantly across transformed data sets. The average misclassification rates and model sizes of the proposed method exhibit modest (mlcr-transformed) or no (log-transformed or square-root-transformed) changes, demonstrating its expected robustness.

Overall, the results from the three data sets consistently convey that: 1) the proposed SEDA is always the best-performing method on original highly-skewed and zero-inflated data, with a significant margin of error improvement; 2) the performance of other methods can be significantly improved with data transformations that mitigate skewness, but the resulting misclassification errors and selected variables are dependent on the transformation choice; 3) the proposed SEDA maintains competitive or better accuracy even when accounting for transformations while being consistent in selected variables, facilitating the robustness and reproducibility of analyses.

6. Discussion

There are several further research directions that could be pursued. First, our estimation consistency result is non-trivial, leading us to develop a new technique to facilitate underlying theoretical analyses by combining sub-Gaussian properties of sign vector with newly established bounds on second derivatives of inverse bridge function for truncated/truncated cases. The theoretical techniques introduced in this work can be extended beyond the specific application considered in the manuscript. While we have focused on the binary-truncated mixed model in Definition 3, the framework can flexibly accommodate the bridge functions of continuous, binary, ordinal, and truncated variables, ensuring estimation consistency in high-dimensional settings under a semiparametric latent Gaussian copula regression model, such as the model considered in Dey and Zipunnikov (2022). This generalization requires establishing bounds on the first and second derivatives of corresponding inverse bridge functions, which, while technical, can be accomplished similarly as in Appendix C.3. Secondly, our focus here is on the binary classification problem, and multi-class extensions are of interest. In case the classes have a natural ordering, e.g., a disease classification as "Mild", "Moderate," or "Severe," the extension is straightforward by considering the ordinal-truncated mixed model, with corresponding bridge function for ordinal-truncated case as derived in Huang et al. (2021). However, it is unclear how to incorporate unordered class labels due to ambiguity in the underlying latent Gaussian representation, making it a compelling question for future study. The `R` implementing SEDA are available at `https://github.com/heech31/SEDA`.

Acknowledgments

Chung's research was partially supported by the Texas A&M Institute of Data Science and the Texas A&M Strategic Transformative Research Program. Ni's research was supported by National Institutes of Health grant 1R01GM148974-01 and National Science Foundation grant DMS-2112943. Gaynanova's research was supported by National Science Foundation grant DMS-2044823. This research were conducted with the advanced computing resources provided by Texas A&M and UNC Charlotte.

Appendix A. First-order Taylor approximation of the posterior probability

Let $\mu_t = \mathrm{E}(Z_t \mid Z_o = z_o, Z_t \leq \Delta_t)$ and

$$g(Z_t) = \Phi\left(\frac{\beta_t^{*\top} Z_t + \beta_o^{*\top} z_o - \Delta_y}{v}\right).$$

Then, by Taylor expansion,

$$\begin{aligned}
\Pr(Y = 1 \mid X) &= \mathrm{E}\left\{g(Z_t) \mid Z_o = z_o, Z_t < \Delta_t\right\} \\
&\approx \mathrm{E}\left\{g(\mu_t) + \nabla g(\mu_t)^\top (Z_t - \mu_t) \mid Z_o = z_o, Z_t < \Delta_t\right\} \\
&= g(\mu_t) + \nabla g(\mu_t)^\top \mathrm{E}\left\{(Z_t - \mu_t) \mid Z_o = z_o, Z_t < \Delta_t\right\} \\
&= g(\mu_t).
\end{aligned}$$

Appendix B. Additional numerical results

B.1 Implementation details of the methods

We consider high-dimensional COpula Discriminant Analysis (CODA) of Han et al. (2013), Negative Binomial Linear Discriminant Analysis (NBLDA) of Dong et al. (2016), Classification and Clustering of Sequencing Data Based on a Poisson Model (PoiClaClu) of Witten (2011), Random Forest (RF) of Breiman (2001), Sparse Logistic regression (S-Logistic) of Friedman et al. (2010), Sparse Semiparametric Discriminant Analysis (SSDA) of Mai and Zou (2015), and Sparse Support Vector Machine (S-SVM) of Yi and Huang (2017) using `R` packages `NBLDA` (Goksuluk et al., 2022), `PoiClaclu` (Witten, 2019), `randomForest` (Liaw and Wiener, 2002), `glmnet` (Friedman et al., 2010), and `sparseSVM` (Yi and Zeng, 2018), respectively.

Since CODA, SSDA, and SEDA do not have available software, we use the default settings described in the original papers. For all methods, we use the default settings in the associated software for selecting tuning parameters and classification.

For both CODA and SSDA, the sparsity tuning parameter is chosen via 5-fold cross-validation to minimize the misclassification error rate. In both methods, the intercept in the classification rule is set with the optimal intercept of Mai et al. (2012). Specifically, given $\hat{\beta}$, whether from CODA or SSDA, the corresponding optimal intercept $\hat{\beta}_0^{\text{opt}}$ is

$$\hat{\beta}_0^{\text{opt}} = -\hat{\mu}_a^\top \hat{\beta} + \frac{\hat{\beta}^\top \hat{S} \hat{\beta}}{\hat{\mu}_d^\top \hat{\beta}} \log\left(\frac{n_1}{n_0}\right),$$

where $\hat{S}$, $\hat{\mu}_a$, and $\hat{\mu}_d$ are the estimated common covariance matrix, global mean, and mean difference, respectively. Using the optimal intercept, the sample classification rule assigns a new observation X^{new} to class 1 if

$$\{\hat{f}(X^{\text{new}}) - \hat{\mu}_a\}^\top \hat{\beta} + \hat{\beta}_0^{\text{opt}} > 0$$

and to class 0, otherwise, where $\hat{f} = (\hat{f}_1, \ldots, \hat{f}_p)^\top$ denotes the estimated copula transformation. For SEDA, both the sparsity tuning parameter λ and the intercept Δ_y are selected based on 5-fold cross-validation to minimize the misclassification error rate using a grid of 100 values for each.

B.2 Joint model

We generate data from the latent Gaussian copula model for binary/truncated mixed data as in Definition 3. Recall that given full correlation matrix Σ, the population direction is $\beta^* = \Sigma_{22}^{-1}\Sigma_{21}$. To generate β^* with a given support $\mathcal{S} = \{j : \beta_j^* \neq 0\}$ for each of the three correlation structures Σ_{22} from above, we define Σ_{21} as follows.

Let $b = (b_1, \ldots, b_p)^\top \in \{0,1\}^p$ be the indicator vector for the signal variables such that $b_j = 1$ if $j \in \mathcal{S}$ and $b_j = 0$, otherwise. Let $v^2 = 1 - \Sigma_{21}^\top \Sigma_{22}^{-1} \Sigma_{21} = 0.05$ be the prespecified conditional variance of $Z_y | Z_1, \ldots, Z_p$. We set $\Sigma_{21} = \{(1-v^2)/b^\top \Sigma_{22} b\}^{1/2} \Sigma_{22} b$ to ensure positive-definiteness of the full correlation matrix Σ with the desired sparsity of $\beta^* = \{(1-v^2)/b^\top \Sigma_{22} b\}^{1/2} b$.

Given Σ, we follow the synthetic microbiome data generation mechanism proposed in Yoon et al. (2019). Specifically, we select monotone transformations and truncation levels so that the resulting synthetic X follows the empirical marginal cumulative distributions of the reference data of Vandeputte et al. (2017). To investigate the effect of truncation, we divide all 101 reference variables according to three truncation levels: no truncation (0%), low truncation (10%-50%), and high truncation (40%-80%). For each level, we use the empirical cumulative distribution functions of the corresponding reference variables to generate $p = 300$ covariates (as the number of the reference variables is less than 300, we use the same empirical cdf to generate multiple synthetic variables).

Let $\tilde{F}_j$ be the empirical cumulative distribution function chosen to represent variable X_j. For $i = 1, \ldots, n$, we generate $(Z_{i,y}, Z_i)^\top \sim \mathrm{N}_{1+p}(0, \Sigma)$ and obtain Y_i and $X_i = (X_{i1}, \ldots, X_{ip})^\top$ as $Y_i = 1(Z_{i,y} > \Delta_y)$, $X_{ij} = \tilde{F}_j^- \circ \Phi(Z_{ij})$, $j = 1, \ldots, p$, where $\tilde{F}_j^-(u) = \min_i\{X_{ij} \mid \tilde{F}_j(X_{ij}) \geq u\}$. For the balanced and unbalanced class settings, we set $\Delta_y =$

$\Phi^{-1}(0.5) = 0$ and $\Delta_y = \Phi^{-1}(0.2) = -0.842$ resulting in $\Pr(Y = 0) = 0.5$ and $\Pr(Y = 0) = 0.2$, respectively. Marginally, this data generation scheme for X_j is the uniform sampling with replacement of the observations of the jth reference variable, but the joint association structure is induced by the prespecified latent correlation matrix Σ_{22}. Under the joint population, we define the Oracle classification rule as

$$\delta_J(X^{\text{new}}) = 1\left\{\beta_t^\top \tilde{\mu}_t + \beta_o^\top \hat{f}_o(X_o^{\text{new}}) - \Delta_y > 0\right\}, \tag{A1}$$

where the marginal transformation for the observed variables $\hat{f}_o$ is estimated with the training sample, $\tilde{\mu}_t$ is estimated with $\hat{Z}_o^{\text{new}}$ and the true latent correlation matrix (11), β_t and β_o are from the true latent correlation matrix Σ, and Δ_y is the population threshold.

B.3 Mixture model

Han et al. (2013) consider the following model:

$$X|(Y = g) \sim \text{NPN}(\mu_g, \Sigma, f) \quad (g = 0, 1), \tag{A2}$$

where $\mu_g \in \mathbb{R}^p$ is the mean of class $g = 0, 1$ and $\Sigma \in \mathbb{R}^{p\times p}$ is a common covariance matrix. Thus, unlike Definition (1), CODA allows latent Gaussian vector to have a non-zero mean and covariance matrix by restricting monotone transformations, f_j $(j = 1, \ldots, p)$, to be mean and variance preserving, i.e., each f_j satisfies the following moment matching conditions:

$$\begin{aligned} \text{E}\{f_j(X_j)|Y = g\} &= \text{E}(X_j|Y = g) = \mu_{j,g} \quad (j = 1, \ldots, p), \\ \text{var}\{f_j(X_j)|Y = g\} &= \text{var}(X_j|Y = g) = \sigma_j^2 \quad (j = 1, \ldots, p). \end{aligned}$$

This model does not account for zero inflation; it assumes continuous X.

To generate realistic simulation data, we set $\Sigma = S\Sigma_{22}S$, where $S = \text{diag}(s_1, \ldots, s_p)$ contains the sample standard deviations of the reference variables and Σ_{22} is one of the three correlation structures described in Section 4. We set the class means μ_g $(g = 0, 1)$ and discriminant direction β^* as the following.

Let $\mu_a = (\mu_1 + \mu_0)/2$ and $\mu_d = \mu_1 - \mu_0$ be the global mean and mean difference, respectively. Under model (A2), the Bayes classification direction is given by $\beta^* = \Sigma^{-1}\mu_d$. When $\Pr(Y = 1) = \Pr(Y = 0)$, the Bayes error rate is $\alpha = \Phi\left(-q^{1/2}/2\right)$, where $q = \beta^{*\top}\Sigma\beta^*$. Given the support $\mathcal{S}$, let $b \in \{0, 1\}^p$ be the corresponding indicator vector such that $b_j = 1$ if $j \in \mathcal{S}$ and $b_j = 0$, otherwise. Fixing the Bayes error rate at $\alpha = 0.1$, we generate β^* as $\beta^* = -2\Phi^{-1}(\alpha)b/(b^\top\Sigma b)^{1/2}$, and obtain $\mu_d = \Sigma\beta^*$. Finally, we set $\mu_0 = C1_p$ and $\mu_1 = \mu_0 + \mu_d$, where the constant $C > 0$ is chosen sufficiently large to mimic the means of the reference data, leading to generated synthetic data with non-negative values.

Given μ_g and Σ from above, we generate $Z_i|(Y_i = g) \sim \text{N}(\mu_g, \Sigma)$ $(i = 1, \ldots, n;\ g = 0, 1)$, where class sizes are $n_0 = n_1 = n/2 = 75$ for the balanced setting, and $n_0 = 30$ and $n_1 = 120$ for the unbalanced setting. To obtain continuous $X_1^*, \ldots, X_n^*$ that follow model (A2), we use the identity transformation (Han et al., 2013; Liu et al., 2009), i.e., $Z_j = X_j^* = X_j$, $j = 1, \ldots, p$.

Under the mixture model, the Oracle classification rule is

$$\delta_M(X) = 1\left\{\left(\hat{f}(X) - \mu_a\right)^\top \beta^* + \log\frac{\Pr(Y=1)}{\Pr(Y=0)} > 0\right\}, \tag{A3}$$

where $\hat{f} = (\hat{f}_1, \ldots, \hat{f}_p)^\top$ is estimated from the training data as in Section 3.2. of Han et al. (2013), μ_a is the true population global mean, $\beta^* = \Sigma^{-1}\mu_d$, and $\Pr(Y = g)$ is the population proportion of the class g.

B.4 Additional results on simulated data

This section provides complete simulation results. Table A1 display average misclassification rates including SEDA with (10), and Table A2 displays the average Matthews correlation coefficient (Matthews, 1975), where a larger value represents a better variable selection performance. The two boundary values, 1 and -1, indicate the completely correct and incorrect variable selections, respectively. Random forest is omitted, as it does not provide variable selection. The two approximations, Taylor and Monte Carlo, of the conditional class probability results in practically the same misclassification rates and variable selections.

Table A1: Average missclassification rates (%) for simulated data based on 100 replications, with standard errors in parentheses.

Joint	CODA	NBLDA	PoiClaClu	RF	S-Logistic	SSDA	S-SVM	SEDA_{L}	SEDA_{MC}
(50:50)				No truncation					
AR	17.5 (0.3)	19.5 (0.3)	16.0 (0.3)	14.0 (0.2)	15.3 (0.2)	13.4 (0.3)	21.9 (0.8)	13.0 (0.3)	13.0 (0.3)
CS	17.2 (0.3)	22.7 (0.3)	20.4 (0.3)	13.7 (0.2)	16.2 (0.2)	15.7 (0.3)	15.0 (0.2)	14.8 (0.3)	14.8 (0.3)
GD	16.7 (0.3)	22.0 (0.4)	22.3 (0.4)	20.3 (0.3)	19.6 (0.4)	17.2 (0.3)	21.7 (0.4)	16.4 (0.3)	16.4 (0.3)
				Low truncation					
AR	23.9 (0.3)	31.1 (0.4)	19.8 (0.3)	14.9 (0.2)	23.4 (0.3)	17.0 (0.3)	33.8 (0.6)	15.0 (0.3)	15.0 (0.3)
CS	19.1 (0.3)	27.2 (0.3)	22.5 (0.5)	14.0 (0.2)	19.3 (0.2)	17.2 (0.3)	16.5 (0.3)	14.9 (0.3)	14.9 (0.3)
GD	22.9 (0.4)	24.7 (0.4)	22.7 (0.4)	20.5 (0.3)	25.9 (0.5)	19.9 (0.3)	30.5 (0.6)	17.0 (0.3)	17.0 (0.3)
				High truncation					
AR	36.1 (0.6)	33.5 (0.3)	17.5 (0.3)	16.7 (0.2)	28.8 (0.3)	20.1 (0.4)	37.5 (0.6)	18.5 (0.3)	18.6 (0.3)
CS	32.9 (0.6)	32 (0.4)	29.8 (0.5)	14.3 (0.2)	20.9 (0.2)	18.8 (0.3)	17.1 (0.3)	15.8 (0.2)	15.8 (0.2)
GD	29.4 (0.6)	25.1 (0.4)	23 (0.4)	21.0 (0.4)	30.3 (0.7)	22.6 (0.4)	32.7 (0.6)	18.2 (0.3)	18.2 (0.3)
(20:80)				No truncation					
AR	14.6 (0.2)	15.8 (0.3)	17.2 (0.3)	17.4 (0.2)	16.3 (0.3)	10.9 (0.2)	20.0 (0.2)	9.8 (0.2)	9.8 (0.2)
CS	11.9 (0.2)	16.1 (0.2)	16.5 (0.4)	11.3 (0.1)	13.0 (0.2)	12.8 (0.2)	10.2 (0.2)	11.0 (0.2)	11.0 (0.2)
GD	13.3 (0.2)	18.6 (0.3)	26.4 (0.5)	16.4 (0.2)	16.7 (0.3)	13.7 (0.2)	16.1 (0.2)	12.6 (0.2)	12.6 (0.2)
				Low truncation					
AR	22.0 (1.1)	19.1 (0.2)	19.3 (0.5)	18.8 (0.2)	19.9 (0.2)	14.3 (0.2)	20.2 (0.2)	12.3 (0.2)	12.3 (0.2)
CS	14.6 (0.2)	16.8 (0.2)	13.6 (0.4)	10.8 (0.1)	15.1 (0.3)	12.9 (0.2)	11.1 (0.2)	11.0 (0.2)	11.0 (0.2)
GD	16.7 (0.3)	17.3 (0.3)	27.5 (0.6)	16.8 (0.2)	19.2 (0.2)	14.9 (0.2)	20.1 (0.4)	12.9 (0.2)	12.9 (0.2)
				High truncation					
AR	21.2 (0.6)	19.9 (0.3)	23.2 (0.4)	19.7 (0.2)	19.8 (0.2)	20.3 (0.2)	20.0 (0.2)	16.7 (0.2)	16.7 (0.2)
CS	21.5 (0.9)	17.1 (0.2)	27.3 (1.2)	11.2 (0.2)	16.7 (0.3)	16.5 (0.2)	12.7 (0.3)	11.7 (0.2)	11.7 (0.2)
GD	21.4 (1.1)	17.0 (0.3)	26.7 (0.4)	17.3 (0.2)	19.6 (0.2)	17.7 (0.3)	20.1 (0.3)	13.6 (0.2)	13.5 (0.2)

Mixture	CODA	NB-DA	PoiClaClu	RF	S-Logistic	SSDA	S-SVM	SEDA_{L}	SEDA_{MC}
(50:50)				No truncation					
AR	15.3 (0.3)	41.4 (1.0)	41.5 (1.0)	11.7 (0.2)	11.6 (0.2)	12.5 (0.3)	12.8 (0.4)	12.2 (0.2)	12.2 (0.2)
CS	14.9 (0.3)	15.9 (0.2)	16.7 (0.3)	13.1 (0.2)	12.0 (0.2)	13.1 (0.3)	14.2 (0.2)	12.9 (0.2)	12.9 (0.2)
GD	13.9 (0.2)	21.6 (0.4)	22.0 (0.4)	13.5 (0.2)	12.4 (0.2)	13.1 (0.3)	14.3 (0.3)	13.2 (0.3)	13.2 (0.3)
(20:80)				No truncation					
AR	12.6 (0.3)	41.5 (0.4)	44.9 (0.5)	11.6 (0.2)	10.7 (0.2)	9.5 (0.2)	11.3 (0.2)	9.1 (0.2)	9.1 (0.2)
CS	11.7 (0.2)	15.3 (0.3)	15.9 (0.3)	9.6 (0.2)	11.2 (0.2)	10.7 (0.3)	9.9 (0.2)	9.8 (0.2)	9.8 (0.2)
GD	11.0 (0.2)	21.1 (0.4)	22.4 (0.5)	11.5 (0.2)	11.5 (0.3)	10.3 (0.2)	10.8 (0.2)	10.0 (0.2)	10.0 (0.2)

Table A2: Average Matthews correlation coefficients and standard errors over 100 replications from the joint and mixture populations.

Joint	CODA	NBLDA	PoiClaClu	S-Logistic	SSDA	S-SVM	SEDA$_{\mathrm{L}}$	SEDA$_{\mathrm{MC}}$
(50:50)	No truncation							
AR	0.20 (0.01)	0.00 (0.00)	0.62 (0.02)	0.65 (0.01)	0.70 (0.01)	0.54 (0.02)	0.72 (0.01)	0.72 (0.01)
CS	0.21 (0.01)	0.00 (0.00)	0.00 (0.01)	0.22 (0.01)	0.23 (0.01)	0.03 (0.00)	0.20 (0.01)	0.20 (0.01)
GD	0.14 (0.00)	0.00 (0.00)	0.09 (0.01)	0.16 (0.01)	0.23 (0.01)	0.11 (0.01)	0.18 (0.01)	0.17 (0.01)
	Low truncation							
AR	0.11 (0.00)	0.00 (0.00)	0.55 (0.02)	0.48 (0.01)	0.67 (0.01)	0.43 (0.01)	0.69 (0.02)	0.69 (0.02)
CS	0.12 (0.00)	0.00 (0.00)	0.05 (0.01)	0.10 (0.01)	0.20 (0.01)	0.00 (0.00)	0.18 (0.01)	0.18 (0.01)
GD	0.07 (0.01)	0.00 (0.00)	0.08 (0.01)	0.08 (0.01)	0.17 (0.01)	0.08 (0.01)	0.15 (0.01)	0.15 (0.01)
	High truncation							
AR	0.15 (0.01)	0.00 (0.00)	0.63 (0.02)	0.45 (0.01)	0.62 (0.01)	0.38 (0.01)	0.65 (0.01)	0.65 (0.01)
CS	0.06 (0.01)	0.00 (0.00)	0.02 (0.01)	0.10 (0.01)	0.15 (0.01)	0.02 (0.00)	0.16 (0.01)	0.16 (0.01)
GD	0.09 (0.00)	0.00 (0.00)	0.09 (0.01)	0.09 (0.01)	0.14 (0.01)	0.09 (0.01)	0.13 (0.01)	0.13 (0.01)
(20:80)	No truncation							
AR	0.20 (0.01)	0.00 (0.00)	0.55 (0.02)	0.51 (0.01)	0.58 (0.01)	0.19 (0.02)	0.66 (0.01)	0.66 (0.01)
CS	0.20 (0.01)	0.00 (0.00)	0.02 (0.00)	0.15 (0.01)	0.16 (0.01)	0.03 (0.00)	0.15 (0.01)	0.15 (0.01)
GD	0.11 (0.01)	0.00 (0.00)	0.08 (0.01)	0.05 (0.01)	0.17 (0.01)	0.04 (0.01)	0.15 (0.01)	0.15 (0.01)
	Low truncation							
AR	0.04 (0.00)	0.00 (0.00)	0.47 (0.03)	0.19 (0.02)	0.50 (0.01)	0.06 (0.01)	0.62 (0.01)	0.62 (0.01)
CS	0.10 (0.00)	0.00 (0.00)	0.03 (0.00)	0.07 (0.01)	0.12 (0.01)	-0.02 (0.00)	0.13 (0.01)	0.13 (0.01)
GD	0.05 (0.00)	0.00 (0.00)	0.08 (0.01)	-0.01 (0.00)	0.06 (0.01)	-0.01 (0.00)	0.12 (0.01)	0.12 (0.01)
	High truncation							
AR	0.07 (0.01)	0.00 (0.00)	0.58 (0.02)	0.02 (0.01)	0.35 (0.01)	0.04 (0.01)	0.55 (0.01)	0.55 (0.01)
CS	0.04 (0.01)	0.00 (0.00)	0.00 (0.00)	0.05 (0.01)	0.06 (0.01)	0.01 (0.00)	0.08 (0.01)	0.08 (0.01)
GD	0.04 (0.01)	0.00 (0.00)	0.08 (0.01)	-0.01 (0.00)	0.03 (0.01)	-0.01 (0.00)	0.11 (0.01)	0.11 (0.01)

Mixture	CODA	NB-LDA	PoiClaClu	S-Logistic	SSDA	S-SVM	SEDA$_{\mathrm{L}}$	SEDA$_{\mathrm{MC}}$
(50:50)	No truncation							
AR	0.11 (0.00)	0.13 (0.01)	0.14 (0.01)	0.60 (0.01)	0.54 (0.02)	0.71 (0.02)	0.65 (0.01)	0.65 (0.01)
CS	0.03 (0.00)	-0.02 (0.00)	-0.03 (0.00)	0.22 (0.01)	0.20 (0.01)	0.03 (0.00)	0.22 (0.01)	0.22 (0.01)
GD	0.03 (0.00)	0.00 (0.00)	0.01 (0.00)	0.21 (0.01)	0.20 (0.01)	0.13 (0.01)	0.19 (0.01)	0.19 (0.01)
(20:80)	No truncation							
AR	0.09 (0)	0.02 (0.01)	0.09 (0.01)	0.58 (0.01)	0.49 (0.01)	0.36 (0.01)	0.61 (0.01)	0.61 (0.01)
CS	0.00 (0)	-0.01 (0.00)	0.00 (0.00)	0.20 (0.01)	0.17 (0.01)	0.08 (0.00)	0.16 (0.01)	0.16 (0.01)
GD	0.03 (0)	0.00 (0.00)	0.00 (0.00)	0.21 (0.01)	0.19 (0.01)	0.09 (0.01)	0.17 (0.01)	0.17 (0.01)

Appendix C. Proofs of theoretical results

C.1 Notation and assumptions

In this section, we prove the results stated in the main manuscript. For completeness, we start with restating the notations and assumptions from the main manuscript as follows. For a vector $a \in \mathbb{R}^p$, we denote the ℓ_q-norm, $q \in [0, \infty)$, by $\|a\|_q = (\sum_{j=1}^p |a_j|^q)^{1/q}$ and the ℓ_∞-norm by $\|a\|_\infty = \max_{1 \leq j \leq p} |a_j|$. For two vectors with the same size, $a, b, c \in \mathbb{R}^p$, we write $a < b$ to denote element-wise inequalities such that $a_j < b_j$, $j = 1, \dots, p$ and $a \in (b, c)$ whether $b < a < c$ or $c < a < b$. The vectors $1_p, 0_p \in \mathbb{R}^p$ denote the one and zero vectors and matrices $I_p, 1_{pp} \in \mathbb{R}^{p \times p}$ denote the identity and matrix with ones. For a matrix $A \in \mathbb{R}^{n \times p}$, $\|A\|_\infty = \max_{j,k} |a_{jk}|$ denotes its ℓ_∞-norm, and for a square matrix $T \in \mathbb{R}^{p \times p}$, $|T|$ denotes its determinant, and $\lambda_{\max}(T)$ and $\lambda_{\min}(T)$ denote the largest and smallest eigenvalues of T. For two matrices with the same size, $A, B \in \mathbb{R}^{n \times p}$, $A \circ B$ denotes the Hadamard product defined as $A \circ B = [a_{jk}b_{jk}] \in \mathbb{R}^{n\times p}$. For two functions f and g, we denote their composite function by $f \circ g = f(g(x))$. We let $1(\cdot)$ denote the indicator function taking the value 1 when its argument is true and 0 otherwise. For a sequence of random variables, $X_1, \dots, X_n, \dots$, we write $X_n = O_p(a_n)$ if, for any $\varepsilon \in (0,1)$, there exist $M, N > 0$ such that $\Pr(|X_n/a_n| > M) < \varepsilon$ for all $n > N$. We let $\Phi_d(a_1, \dots, a_d; \Sigma)$ and $\Phi(\cdot)$ denote the d-dimensional Gaussian distribution function with zero mean and correlation matrix Σ evaluated at $(a_1, \dots, a_d)^\top$ and the univariate standard Gaussian distribution function, respectively. We use C and $C_i, i = 1, 2, \dots$, to denote generic constants that do not depend on the sample size n, dimension p, and model parameters, where their values may change from line to line. We write $\mathrm{card}(\mathcal{S})$ to denote the cardinality of a set $\mathcal{S}$.

Throughout, we use G to denote the bridge function such that for TT case $\tau_{jk} = G(\Sigma_{jk}, \Delta_j, \Delta_k)$ with $\Sigma_{22} = \left[G^{-1}(\tau_{jk}, \Delta_j, \Delta_k)\right]_{1\le j,k\le p} = G^{-1}(T, \Delta)$ and $\hat{\Sigma}_{22} = G^{-1}(\hat{T}, \hat{\Delta})$. Here T and $\hat{T}$ are the population and sample Kendall's τ matrices, respectively, $\Delta = (\Delta_1, \dots, \Delta_p)^\top$, $\hat{\Delta} = (\hat{\Delta}_1, \dots, \hat{\Delta}_p)^\top$ with $\Delta_j = \Phi^{-1}(\pi_j)$ with $\pi_j = \Pr(Z_j \le \Delta_j) = \Pr(X_j = 0)$, and $\hat{\Delta}_j = \Phi^{-1}(\hat{\pi}_j)$ with $\hat{\pi}_j = \sum_{i=1}^n 1(X_{ij} = 0)/n$ being the sample zero proportion of the jth variable.

C.2 Proofs of the theorems from the main manuscript

Proof of Theorem 6 The proof follows the proof of Theorem 2 in Gaynanova (2020). For completeness, we provide the full proof as follows. By the optimality condition of equation (9) in the main manuscript, we have

$$\hat{\Sigma}_{22}\hat{\beta} - \hat{\Sigma}_{21} + \lambda g = 0,$$

where g is a subgradient of $\|\beta\|_1$ at $\hat{\beta}$. This gives

$$(\hat{\beta} - \beta^*)^\top(\hat{\Sigma}_{22}\hat{\beta} - \hat{\Sigma}_{21} + \lambda g) = 0,$$

and thus

$$(\hat{\beta} - \beta^*)^\top\hat{\Sigma}_{22}(\hat{\beta} - \beta^*) - (\hat{\beta} - \beta^*)^\top(\hat{\Sigma}_{21} - \hat{\Sigma}_{22}\beta^*) + \lambda(\hat{\beta} - \beta^*)^\top g = 0. \tag{A4}$$

Since g is a subgradient of $\|\beta\|_1$ at $\hat{\beta}$,

$$(\hat{\beta} - \beta^*)^\top g \ge \|\hat{\beta}\|_1 - \|\beta^*\|_1. \tag{A5}$$

By combining (A4), (A5), and Hölder's and triangle inequalities, we have

$$\begin{aligned}(\hat{\beta} - \beta^*)^\top\hat{\Sigma}_{22}(\hat{\beta} - \beta^*) &\le (\hat{\beta} - \beta^*)^\top(\hat{\Sigma}_{21} - \hat{\Sigma}_{22}\beta^*) + \lambda\|\beta^*\|_1 - \lambda\|\hat{\beta}\|_1 \\ &\le \|\hat{\beta} - \beta^*\|_1\|\hat{\Sigma}_{21} - \hat{\Sigma}_{22}\beta^*\|_\infty + \lambda\|\beta^*\|_1 - \lambda\|\hat{\beta}\|_1.\end{aligned}$$

Using the condition on λ and Assumption 4,

$$\begin{aligned}(\hat{\beta} - \beta^*)^\top\hat{\Sigma}_{22}(\hat{\beta} - \beta^*) &\le \frac{\lambda}{2}\|\hat{\beta} - \beta^*\|_1 + \lambda\|\beta^*\|_1 - \lambda\|\hat{\beta}\|_1 \\ &= \frac{\lambda}{2}\|\hat{\beta}_{\mathcal{S}} - \beta^*_{\mathcal{S}}\|_1 + \frac{\lambda}{2}\|\hat{\beta}_{\mathcal{S}^c}\|_1 + \lambda\|\beta^*\|_1 - \lambda\|\hat{\beta}\|_1 \\ &= \frac{\lambda}{2}\|\hat{\beta}_{\mathcal{S}} - \beta^*_{\mathcal{S}}\|_1 + \frac{\lambda}{2}\|\hat{\beta}_{\mathcal{S}^c}\|_1 + \lambda\|\beta^*_{\mathcal{S}}\|_1 - \lambda\|\hat{\beta}\|_1 \\ &= \frac{\lambda}{2}\|\hat{\beta}_{\mathcal{S}} - \beta^*_{\mathcal{S}}\|_1 + \frac{\lambda}{2}\|\hat{\beta}_{\mathcal{S}^c}\|_1 + \lambda\|\beta^*_{\mathcal{S}}\|_1 - \lambda\|\hat{\beta}_{\mathcal{S}}\|_1 - \lambda\|\hat{\beta}_{\mathcal{S}^c}\|_1.\end{aligned}$$

Using the triangle inequality,

$$(\hat{\beta}-\beta^*)^\top\hat{\Sigma}_{22}(\hat{\beta}-\beta^*) \leq \frac{\lambda}{2}\|\hat{\beta}_{\mathcal{S}}-\beta^*_{\mathcal{S}}\|_1 + \frac{\lambda}{2}\|\hat{\beta}_{\mathcal{S}^c}\|_1 + \lambda\|\hat{\beta}_{\mathcal{S}}-\beta^*_{\mathcal{S}}\|_1 - \lambda\|\hat{\beta}_{\mathcal{S}^c}\|_1 \tag{A6}$$

$$= \frac{3\lambda}{2}\|\hat{\beta}_{\mathcal{S}}-\beta^*_{\mathcal{S}}\|_1 - \frac{\lambda}{2}\|\hat{\beta}_{\mathcal{S}^c}\|_1 \tag{A7}$$

$$\leq \frac{3\lambda}{2}\|\hat{\beta}_{\mathcal{S}}-\beta^*_{\mathcal{S}}\|_1. \tag{A8}$$

As $(\hat{\beta}-\beta^*)^\top\hat{\Sigma}_{22}(\hat{\beta}-\beta^*)$ is non-negative and $\beta^*_{\mathcal{S}^c}=0$, (A7) implies that $\|\hat{\beta}_{\mathcal{S}^c}\|_1 \leq 3\|\hat{\beta}_{\mathcal{S}}-\beta^*_{\mathcal{S}}\|_1$, and thus, $\hat{\beta}-\beta^*$ is in the cone $\mathcal{C}(\mathcal{S},3)$. Since $\hat{\Sigma}_{22}$ satisfies RE$(s,3)$ with paramter γ, we have

$$\|\hat{\beta}_{\mathcal{S}}-\beta^*_{\mathcal{S}}\|_2 \leq \{\gamma(\hat{\beta}-\beta^*)^\top\hat{\Sigma}_{22}(\hat{\beta}-\beta^*)\}^{1/2}. \tag{A9}$$

Since $\|\hat{\beta}_{\mathcal{S}}-\beta^*_{\mathcal{S}}\|_1 \leq s^{1/2}\|\hat{\beta}_{\mathcal{S}}-\beta^*_{\mathcal{S}}\|_2$, (A8) and (A9) imply that

$$(\hat{\beta}-\beta^*)^\top\hat{\Sigma}_{22}(\hat{\beta}-\beta^*) \leq \frac{9}{4}\gamma s\lambda^2 \tag{A10}$$

The bound for $\|\hat{\beta}_{\mathcal{S}}-\beta^*_{\mathcal{S}}\|_2$ can be obtained as follows. Since $\hat{\beta}-\beta^* \in \mathcal{C}(\mathcal{S},3)$,

$$\|\hat{\beta}-\beta^*\|_1 = \|\hat{\beta}_{\mathcal{S}}-\beta^*_{\mathcal{S}}\|_1 + \|\hat{\beta}_{\mathcal{S}^c}-\beta^*_{\mathcal{S}^c}\|_1 \tag{A11}$$

$$\leq 4\|\hat{\beta}_{\mathcal{S}}-\beta^*_{\mathcal{S}}\|_1 \tag{A12}$$

$$\leq 4s^{1/2}\|\hat{\beta}_{\mathcal{S}}-\beta^*_{\mathcal{S}}\|_2 \tag{A13}$$

$$\leq 4\{s\gamma(\hat{\beta}-\beta^*)^\top\hat{\Sigma}_{22}(\hat{\beta}-\beta^*)\}^{1/2} \tag{A14}$$

$$\leq 4(s\gamma)^{1/2}\frac{3}{2}(\gamma s\lambda^2)^{1/2} = 6s\gamma\lambda. \tag{A15}$$

Let $a = \hat{\beta}-\beta^*$ for notational simplicity. For $j=0,1,\ldots,J$, let T_j be the index set of $(j+1)$th s largest (in absolute) elements of a. Then, $a \in \mathcal{C}(T_0,3)$ as

$$\begin{aligned}\|a_{T_0^c}\|_1 = \|a\|_1 - \|a_{T_0}\|_1 &\leq \|a\|_1 - \|a_{\mathcal{S}}\|_1 \\ &= \|a_{\mathcal{S}^c}\|_1 \\ &\leq 3\|a_{\mathcal{S}}\|_1 \quad \text{since } a \in \mathcal{C}(\mathcal{S},3) \\ &\leq 3\|a_{T_0}\|_1.\end{aligned}$$

Furthermore, it follows that $\|a_{T_j}\|_0 = s$ for $j=0,\ldots,J-1$ with last $\|a_{T_J}\|_0 \leq s$. Also, for $j \geq 1$, $\|a_{T_j}\|_2 \leq s^{1/2}\|a_{T_j}\|_\infty \leq s^{1/2}s^{-1}\|a_{T_{j-1}}\|_1$. Thus, by the triangle inequality,

$$\|a\|_2 \leq \|a_{T_0}\|_2 + \sum_{j=1}^{J}\|a_{T_j}\|_2 \leq \|a_{T_0}\|_2 + \sum_{j=1}^{J}s^{1/2}\|a_{T_j}\|_\infty \tag{A16}$$

$$\leq \|a_{T_0}\|_2 + \sum_{j=0}^{J-1}s^{1/2}s^{-1}\|a_{T_j}\|_1 \leq \|a_{T_0}\|_2 + s^{-1/2}\|a\|_1. \tag{A17}$$

Using that $\hat{\Sigma}_{22}$ satisfies RE$(s,3)$ and $a \in \mathcal{C}(T_0, 3)$,

$$\begin{aligned}\|\hat{\beta} - \beta^*\|_2 = \|a\|_2 &\le \|a_{T_0}\|_2 + s^{-1/2}\|a\|_1 \\ &\le \{\gamma(\hat{\beta} - \beta^*)^\top \hat{\Sigma}_{22}(\hat{\beta} - \beta^*)\}^{1/2} + 6s^{1/2}\gamma\lambda \quad \text{as (A9) and (A15)} \\ &\le \frac{3}{2}\gamma s^{1/2}\lambda + 6s^{1/2}\gamma\lambda \quad \text{by (A10)} \\ &= \frac{15}{2}\gamma s^{1/2}\lambda.\end{aligned}$$

■

Proof of Theorem 7 Using $\beta^* = \Sigma_{22}^{-1}\Sigma_{21}$ and triangle inequality, we have

$$\begin{aligned}\|\hat{\Sigma}_{21} - \hat{\Sigma}_{22}\beta^*\|_\infty &= \|\hat{\Sigma}_{21} - \hat{\Sigma}_{22}\beta^* + \Sigma_{21} - \Sigma_{21}\|_\infty \\ &= \|\hat{\Sigma}_{21} - \Sigma_{21} + \Sigma_{21} - \hat{\Sigma}_{22}\beta^*\|_\infty \\ &\le \|\hat{\Sigma}_{21} - \Sigma_{21}\|_\infty + \|(\Sigma_{22} - \hat{\Sigma}_{22})\beta^*\|_\infty.\end{aligned}$$

For $\|\hat{\Sigma}_{21} - \Sigma_{21}\|_\infty$, it follows from Theorem 7 of Yoon et al. (2020) that

$$\|\hat{\Sigma}_{21} - \Sigma_{21}\|_\infty \le C_1\sqrt{\frac{\log(p\eta^{-1})}{n}} \tag{A18}$$

with probability at least $1 - \eta$.

Consider $\|(\Sigma_{22} - \hat{\Sigma}_{22})\beta^*\|_\infty$. Recall that $\Sigma_{22} = G^{-1}(T, \Delta) = \left[G^{-1}(\tau_{jk}, \Delta_j, \Delta_k)\right]_{1\le j,k\le p}$, and $\hat{\Sigma}_{22} = G^{-1}(\hat{T}, \hat{\Delta})$. Let $G_\tau^{-1} = \partial G^{-1}(\tau, \Delta_j, \Delta_k)/\partial\tau$ be the partial derivative of the inverse bridge function with respect to τ. By adding and subtracting $G^{-1}(\hat{T}, \Delta)$ from $G^{-1}(\hat{T}, \hat{\Delta})$ and applying the mean value theorem to $G^{-1}(\hat{T}, \Delta)$ with respect to $\hat{T}$,

$$\begin{aligned}\hat{\Sigma}_{22} = G^{-1}(\hat{T}, \hat{\Delta}) &= G^{-1}(\hat{T}, \Delta) + \{G^{-1}(\hat{T}, \hat{\Delta}) - G^{-1}(\hat{T}, \Delta)\} \\ &= G^{-1}(T, \Delta) + G_\tau^{-1}(\tilde{T}, \Delta) \circ (\hat{T} - T) + \{G^{-1}(\hat{T}, \hat{\Delta}) - G^{-1}(\hat{T}, \Delta)\} \\ &= \Sigma_{22} + G_\tau^{-1}(\tilde{T}, \Delta) \circ (\hat{T} - T) + \{G^{-1}(\hat{T}, \hat{\Delta}) - G^{-1}(\hat{T}, \Delta)\},\end{aligned}$$

where $\tilde{T} = [\tilde{\tau}_{jk}]_{1\le j,k\le p}$ and $\tilde{\tau}_{jk} \in (\hat{\tau}_{jk}, \tau_{jk})$.

Therefore

$$(\Sigma_{22} - \hat{\Sigma}_{22})\beta^* = -G_\tau^{-1}(\tilde{T}, \Delta) \circ (\hat{T} - T)\beta^* - \{G^{-1}(\hat{T}, \hat{\Delta}) - G^{-1}(\hat{T}, \Delta)\}\beta^*.$$

By letting

$$G_\tau^{-1}(\tilde{T}, \Delta) = G_\tau^{-1}(T, \Delta) + \{G_\tau^{-1}(\tilde{T}, \Delta) - G_\tau^{-1}(T, \Delta)\},$$

we further have, by the triangle inequality, that

$$\begin{aligned}\|(\Sigma_{22} - \hat{\Sigma}_{22})\beta^*\|_\infty \le\ & \underbrace{\|G_\tau^{-1}(T, \Delta) \circ (\hat{T} - T)\beta^*\|_\infty}_{:=I_1} + \underbrace{\|\{G_\tau^{-1}(\tilde{T}, \Delta) - G_\tau^{-1}(T, \Delta)\} \circ (\hat{T} - T)\beta^*\|_\infty}_{:=I_2} \\ & + \underbrace{\|\{G^{-1}(\hat{T}, \hat{\Delta}) - G^{-1}(\hat{T}, \Delta)\}\beta^*\|_\infty}_{:=I_3}.\end{aligned}$$

We separately bound I_1, I_2, and I_3 in Lemmas A.1, A.2, and A.3, respectively. Combining these bounds with (A18) completes the proof. ■

Proof of Theorem 8 From Theorem 6, if $\hat{\Sigma}_{22}$ satisfies $RE(s,3)$ and $\lambda \geq 2\|\hat{\Sigma}_{21} - \hat{\Sigma}_{22}\beta^*\|_\infty$, then

$$\|\hat{\beta} - \beta^*\|_2^2 \leq C_1\gamma^2 s\lambda^2.$$

From Theorem 7, if $\lambda = C\sqrt{\log p/n}$, then $\lambda \geq 2\|\hat{\Sigma}_{21} - \hat{\Sigma}_{22}\beta\|_\infty$ holds with high probability. From Lemma A.6, if Σ_{22} satisfies $RE(s,3)$ with parameter $\gamma(\Sigma_{22})$, then with high probability so does $\hat{\Sigma}_{22}$ with $\gamma(\hat{\Sigma}_{22}) = C_2\gamma(\Sigma_{22})$. Combining these results gives that, with high probability,

$$\|\hat{\beta} - \beta^*\|_2^2 \leq C_3\gamma^2 s\frac{\log p}{n},$$

leading to the desired bound. ■

C.3 Main supporting lemmas

Lemma A.1 *Under Assumptions 1—5, for any fixed $\eta \in (0,1)$, there exists some constant $C > 0$ such that*

$$\|G_\tau^{-1}(T,\Delta) \circ (\hat{T} - T)\beta^*\|_\infty \leq C\sqrt{\frac{\log(p\eta^{-1})}{n}}$$

with probability at least $1 - \eta$.

Proof Let $e_j \in \mathbb{R}^p$ be the vector with 1 in the jth component and 0 otherwise. Then

$$\begin{aligned}\|G_\tau^{-1}(T,\Delta) \circ (\hat{T} - T)\beta^*\|_\infty &= \max_{1\leq j\leq p} |e_j^\top G_\tau^{-1}(T,\Delta) \circ (\hat{T} - T)\beta^*| \\ &= \max_{1\leq j\leq p} \|m_j\|_2 |u^\top(\hat{T} - T)e_j|.\end{aligned}$$

where $m_j = (G_\tau^{-1}(\tau_{1j},\Delta_1,\Delta_j)\beta_1^*, \ldots, G_\tau^{-1}(\tau_{pj},\Delta_p,\Delta_j)\beta_p^*)^\top$ and $u = m_j/\|m_j\|_2$ is a deterministic unit vector. Since $|G_\tau^{-1}| \leq C_0$ for some constant $C_0 > 0$ by Theorem 6 of Yoon et al. (2020), we have $\|m_j\|_2 \leq C_0\|\beta^*\|_2 \leq C_0 C_{\text{cov}}$ by Lemma A.4. Therefore,

$$\|G_\tau^{-1}(T,\Delta) \circ (\hat{T} - T)\beta^*\|_\infty \leq C_0 C_{\text{cov}} \max_{1\leq j\leq p} |u^\top(\hat{T} - T)e_j|.$$

Consider $u^\top(\hat{T} - T)e_j$. By Lemma A.8, for any $\epsilon > 0$ and $0 < t \leq n/C_{\text{cov}}$,

$$\begin{aligned}\Pr\left(\max_{1\leq j\leq p} |u^\top(\hat{T} - T)e_j| \geq \epsilon\right) &\leq p \cdot \Pr\left(|u^\top(\hat{T} - T)e_1| \geq \epsilon\right) \\ &\leq 2p \cdot \Pr\left\{\exp\left(tu^\top(\hat{T} - T)e_1\right) \geq \exp(t\epsilon)\right\} \\ &\leq 2p \cdot \mathrm{E}\left[\exp\left(tu^\top(\hat{T} - T)e_1\right)\right]\exp(-t\epsilon) \\ &\leq 2p \cdot \exp\left(\frac{t^2 C_{\text{cov}}^2}{n} - t\epsilon\right) \\ &= 2\exp\left(\log p + \frac{t^2 C_{\text{cov}}^2}{n} - t\epsilon\right).\end{aligned}$$

Letting $\epsilon = 2C_{\text{cov}}\sqrt{\log(2p\eta^{-1})/n}$ and $t = C_{\text{cov}}^{-1}\sqrt{n\log(2p\eta^{-1})}$, we have

$$\Pr\left\{\max_{1\le j\le p}|u^\top(\hat{T}-T)e_j| \ge 2C_{\text{cov}}\sqrt{\frac{\log(2p\eta^{-1})}{n}}\right\} \le 1-\eta.$$

Thus, for some constant $C>0$, $\|G_\tau^{-1}(T,\Delta)\circ(\hat{T}-T)\beta^*\|_\infty \le C\sqrt{\log(p\eta^{-1})/n}$ with probability at least $1-\eta$. ■

Lemma A.2 *Under Assumptions 1—5, for any fixed $\eta\in(0,1)$, there exists some constant $C>0$ such that*

$$\left\|\left\{G_\tau^{-1}(\tilde{T},\Delta)-G_\tau^{-1}(T,\Delta)\right\}\circ(\hat{T}-T)\beta^*\right\|_\infty \le C\sqrt{\frac{\log(p\eta^{-1})}{n}}$$

with probability at least $1-\eta$.

Proof By the mean value theorem, for some $\bar{T}=[\bar{\tau}_{jk}]_{1\le j,k\le p}$, where $\bar{\tau}_{jk}\in(\tilde{\tau}_{jk},\tau_{jk})$, we have

$$\|\{G_\tau^{-1}(\tilde{T},\Delta)-G_\tau^{-1}(T,\Delta)\}\circ(\hat{T}-T)\beta^*\|_\infty = \|G_{\tau\tau}^{-1}(\bar{T},\Delta)\circ(\tilde{T}-T)\circ(\hat{T}-T)\beta^*\|_\infty,$$

where $G_{\tau\tau}^{-1}$ is the 2nd partial derivative of inverse bridge function with respect to τ. By Lemma A.11, $|G_{\tau\tau}^{-1}|\le C_0$ for some constant $C_0>0$. Since $|\bar{\tau}_{jk}-\tau_{jk}|\le|\tilde{\tau}_{jk}-\tau_{jk}|\le|\hat{\tau}_{jk}-\tau_{jk}|$, by Hölder's inequality,

$$\begin{aligned}\|\{G_\tau^{-1}(\tilde{T},\Delta)-G_\tau^{-1}(T,\Delta)\}\circ(\hat{T}-T)\beta^*\|_\infty &\le \|G_{\tau\tau}^{-1}(\bar{T},\Delta)\circ(\tilde{T}-T)\circ(\hat{T}-T)\|_\infty\|\beta^*\|_1\\ &\le C_0\|\tilde{T}-T\|_\infty\|\hat{T}-T\|_\infty\|\beta^*\|_1\\ &\le C_0\|\hat{T}-T\|_\infty^2\|\beta^*\|_1.\end{aligned}$$

By Lemma A.9, $\|\hat{T}-T\|_\infty^2\le C_1\log(p\eta^{-1})/n$ with probability at least $1-\eta$, and by Lemma A.4, $\|\beta^*\|_1\le\sqrt{s}C_{\text{cov}}$. Thus, under Assumption 5, for sufficiently large n and some constant $C>0$,

$$\|I_2\|_\infty \le C\frac{\sqrt{s}\log(p\eta^{-1})}{n} \le C\frac{\log(p\eta^{-1})}{n}$$

with probability at least $1-\eta$. ■

Lemma A.3 *Under Assumptions 1—5, for any fixed $\eta\in(0,1)$, there exists some constant $C>0$ such that*

$$\left\|\left\{G^{-1}(\hat{T},\hat{\Delta})-G^{-1}(\hat{T},\Delta)\right\}\beta^*\right\|_\infty \le C\sqrt{\frac{\log(p\eta^{-1})}{n}}$$

with probability at least $1-\eta$.

Proof We reparameterize $\Delta_j = \Phi^{-1}(\pi_j)$ and write $G^{-1}(T, \Delta) = G^{-1}(T, \pi)$, where $\pi = (\pi_1, \dots, \pi_p)^\top$. We also write $G^{-1}_{\pi_1} = \partial G^{-1}(\tau, \pi_1, \pi_2)/\partial \pi_1$ and $G^{-1}_{\pi_2} = \partial G^{-1}(\tau, \pi_1, \pi_2)/\partial \pi_2$. For each element of $G^{-1}(\hat{T}, \hat{\pi}) - G^{-1}(\hat{T}, \pi)$, the multivariate mean value theorem gives

$$G^{-1}(\hat{\tau}_{jk}, \hat{\pi}_j, \hat{\pi}_k) - G^{-1}(\hat{\tau}_{jk}, \pi_j, \pi_k) = \underbrace{G^{-1}_{\pi_1}(\hat{\tau}_{jk}, \tilde{\pi}_j, \tilde{\pi}_k)(\hat{\pi}_j - \pi_j)}_{:=I_{3,1}} + \underbrace{G^{-1}_{\pi_2}(\hat{\tau}_{jk}, \tilde{\pi}_j, \tilde{\pi}_k)(\hat{\pi}_k - \pi_k)}_{:=I_{3,2}},$$

for some $\tilde{\pi}_j \in (\hat{\pi}_j, \pi_j)$ and $\tilde{\pi}_k \in (\hat{\pi}_k, \pi_k)$, respectively.

Consider $I_{3,1}$. By adding and subtracting $G^{-1}_{\pi_1}(\tau_{jk}, \pi_j, \pi_k)(\hat{\pi}_j - \pi_j)$ to $I_{3,1}$, we have that

$$\begin{aligned}
&G^{-1}_{\pi_1}(\hat{\tau}_{jk}, \tilde{\pi}_j, \tilde{\pi}_k)(\hat{\pi}_j - \pi_j) \\
&\qquad = \left\{G^{-1}_{\pi_1}(\hat{\tau}_{jk}, \tilde{\pi}_j, \tilde{\pi}_k) - G^{-1}_{\pi_1}(\tau_{jk}, \pi_j, \pi_k) + G^{-1}_{\pi_1}(\tau_{jk}, \pi_j, \pi_k)\right\}(\hat{\pi}_j - \pi_j) \\
&\qquad = \left\{G^{-1}_{\pi_1}(\hat{\tau}_{jk}, \tilde{\pi}_j, \tilde{\pi}_k) - G^{-1}_{\pi_1}(\tau_{jk}, \pi_j, \pi_k)\right\}(\hat{\pi}_j - \pi_j) \\
&\qquad\quad + G^{-1}_{\pi_1}(\tau_{jk}, \pi_j, \pi_k)(\hat{\pi}_j - \pi_j).
\end{aligned}$$

By applying the multivariate mean value theorem to $\{G^{-1}_{\pi_1}(\hat{\tau}_{jk}, \tilde{\pi}_j, \tilde{\pi}_k) - G^{-1}_{\pi_1}(\tau_{jk}, \pi_j, \pi_k)\}$, we further have that

$$\begin{aligned}
&G^{-1}_{\pi_1}(\hat{\tau}_{jk}, \tilde{\pi}_j, \tilde{\pi}_k)(\hat{\pi}_j - \pi_j) \\
&\qquad = G^{-1}_{\pi_1\tau}(\bar{\tau}_{jk}, \bar{\pi}_j, \bar{\pi}_k)(\hat{\tau} - \tau)(\hat{\pi}_j - \pi_j) + G^{-1}_{\pi_1\pi_1}(\bar{\tau}_{jk}, \bar{\pi}_j, \bar{\pi}_k)(\tilde{\pi}_j - \pi_j)(\hat{\pi}_j - \pi_j) \\
&\qquad\quad + G^{-1}_{\pi_1\pi_2}(\bar{\tau}_{jk}, \bar{\pi}_j, \bar{\pi}_k)(\tilde{\pi}_k - \pi_k)(\hat{\pi}_j - \pi_j) + G^{-1}_{\pi_1}(\tau_{jk}, \pi_j, \pi_k)(\hat{\pi}_j - \pi_j)
\end{aligned}$$

for some $\bar{\tau}_{jk} \in (\hat{\tau}_{jk}, \tau_{jk})$, $\bar{\pi}_j \in (\tilde{\pi}_j, \pi_j)$, and $\bar{\pi}_k \in (\tilde{\pi}_k, \pi_k)$. Thus, by the triangle inequality,

$$\begin{aligned}
\|G^{-1}_{\pi_1}(\hat{T}, \tilde{\pi}) \circ (\hat{\Pi} - \Pi)\beta^*\|_\infty \le & \|G^{-1}_{\pi_1\tau}(\bar{T}, \bar{\pi}) \circ (\hat{T} - T) \circ (\hat{\Pi} - \Pi)\beta^*\|_\infty && \text{(A19)} \\
& + \|G^{-1}_{\pi_1\pi_1}(\bar{T}, \bar{\pi}) \circ (\tilde{\Pi} - \Pi) \circ (\hat{\Pi} - \Pi)\beta^*\|_\infty && \text{(A20)} \\
& + \|G^{-1}_{\pi_1\pi_2}(\bar{T}, \bar{\pi}) \circ (\hat{\Pi}^\top - \Pi^\top) \circ (\hat{\Pi} - \Pi)\beta^*\|_\infty && \text{(A21)} \\
& + \|G^{-1}_{\pi_1}(T, \pi) \circ (\hat{\Pi} - \Pi)\beta^*\|_\infty, && \text{(A22)}
\end{aligned}$$

where $\Pi = \pi 1_p^\top$ and $\hat{\Pi} = \hat{\pi} 1_p^\top$. We consider (A19)—(A21) and (A22) separately as follows.

For (A19)—(A21), we know that $|G^{-1}_{\pi_1\tau}|$, $|G^{-1}_{\pi_1\pi_1}|$, and $|G^{-1}_{\pi_1\pi_2}|$ are bounded above by some positive constants by Lemmas A.14, A.12, A.13, respectively. Also, for some constants $C_0, C_1 > 0$ and a fixed $\eta \in (0,1)$, $\|\hat{\pi} - \pi\|_\infty \le C_0\sqrt{\log(p\eta^{-1})/n}$ and $\|\hat{T} - T\|_\infty \le C_1\sqrt{\log(p\eta^{-1})/n}$ with probability at least $1 - \eta$ by Lemmas A.5 and A.9. Since $|\tilde{\pi}_j - \pi_j| \le |\hat{\pi}_j - \pi_j|$, it follows that $\|\tilde{\Pi} - \Pi\|_\infty \le \|\hat{\Pi} - \Pi\|_\infty$ and

$$\|\hat{\Pi} - \Pi\|_\infty = \|\hat{\Pi}^\top - \Pi^\top\|_\infty = \|\hat{\pi} - \pi\|_\infty.$$

Hence, we can show that, for some constant $C > 0$, (A19)—(A21) are bounded above by $C\sqrt{\log(p\eta^{-1})/n}$ with probability at least $1 - \eta$ by following the steps of Lemma A.2.

For (A22), we know that $|G^{-1}_{\pi_1}|$ is bounded above by some positive constant by Lemma A.10. Thus, by following the steps of Lemma A.1 with $(\hat{T} - T)$ being replaced by $(\hat{\Pi} - \Pi)$ and using Lemma A.5, it follows that, for some constant $C > 0$, (A22) is bounded above by

$C\sqrt{\log(p\eta^{-1})/n}$ with probability at least $1-\eta$. By combining these results, we have that, for some constant $C>0$,

$$\|G_{\pi_1}^{-1}(\hat{T},\tilde{\pi}) \circ (\hat{\Pi}-\Pi)\beta^*\|_\infty \le C\sqrt{\frac{\log(p\eta^{-1})}{n}} \tag{A23}$$

with probability at least $1-\eta$.

By symmetrically applying above steps to $I_{3,2}$, we also have that, for some constant $C>0$,

$$\|G_{\pi_2}^{-1}(\hat{T},\tilde{\pi}) \circ (\hat{\Pi}-\Pi)^\top\beta^*\|_\infty \le C\sqrt{\frac{\log(p\eta^{-1})}{n}} \tag{A24}$$

with probability at least $1-\eta$. Combining (A23) and (A24) completes the proof. ■

Lemma A.4 *Let $\beta^* = \Sigma_{22}^{-1}\Sigma_{21}$. Under Assumptions 3—4*

$$\|\beta^*\|_2 < C_{\text{cov}} \quad \textit{and} \quad \|\beta^*\|_1 < \sqrt{s}C_{\text{cov}}.$$

Proof Under Assumption 3

$$\|\beta^*\|_2 = \|\Sigma_{22}^{-1}\Sigma_{21}\|_2 \le \|\Sigma_{22}^{-1}\|_{\text{op}}\|\Sigma_{21}\|_2.$$

At the same time, using $e_1 = (1,0,\ldots,0)^\top$,

$$\lambda_{\max}(\Sigma) \ge \|\Sigma e_1\|_2 = \sqrt{1+\|\Sigma_{21}\|_2^2} > \|\Sigma_{21}\|_2.$$

Since $\|\Sigma_{22}^{-1}\|_{\text{op}} = \{\lambda_{\min}(\Sigma)\}^{-1}$, we have that

$$\|\beta^*\|_2 \le \|\Sigma_{22}^{-1}\|_{\text{op}}\|\Sigma_{21}\|_2 < \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)} \le C_{\text{cov}}.$$

Under Assumption 4, it follows

$$\|\beta^*\|_1 \le \sqrt{s}\|\beta^*\|_2 < \sqrt{s}C_{\text{cov}}.$$

■

Lemma A.5 *For $j=1,\ldots,p$, let $\pi_j = \Phi(\Delta_j)$ and $\hat{\pi}_j = n^{-1}\sum_{i=1}^n 1(X_{ij}=0)$, where $\mathrm{E}(\hat{\pi}_j) = \pi_j$. Also, let $\pi = (\pi_1,\ldots,\pi_p)^\top$ and $\hat{\pi} = (\hat{\pi}_1,\ldots,\hat{\pi}_p)^\top$. Then for any deterministic $\|u\|_2 = 1$ and $t \ge 0$, and some constant $C>0$*

$$\mathrm{E}\left[\exp\{tu^\top(\hat{\pi}-\pi)\}\right] \le \exp\left(\frac{t^2C}{n}\right).$$

Proof For $i = 1, \dots, n$, let $b_i = (1(X_{i1} = 0), \dots, 1(X_{ip} = 0))^\top$ such that $n^{-1}\sum_{i=1}^n b_i = \hat{\pi}$. By definition of the truncated latent Gaussian copula model, we have

$$\begin{aligned} b_{ij} &= 1(X_{ij} = 0) = 1(X_{ij}^* \leq D_j) = 1(Z_{ij} \leq \Delta_j) \\ &= 1(Z_{ij} - \Delta_j \leq 0) = \frac{\mathrm{sign}(Z_{ij} - \Delta_j) + 1}{2}. \end{aligned}$$

Since $\tilde{Z}_i = Z_i - \Delta \overset{iid}{\sim} \mathrm{N}_p(-\Delta, \Sigma_{22})$, where $Z_i = (Z_{i1}, \dots, Z_{ip})^\top$ and $\Delta = (\Delta_1, \dots, \Delta_p)^\top$, $\mathrm{sign}(\tilde{Z}_i) - \mathrm{E}\{\mathrm{sign}(\tilde{Z}_i)\}$ is $\mathsf{C}(\Sigma_{22})$-subgaussian by Lemma A.7, and thus $\hat{\pi} - \pi = n^{-1}\sum_{i=1}^n \{b_i - \mathrm{E}(b_i)\}$ is sum of n iid $\mathsf{C}(\Sigma_{22})$-subgaussians. Thus,

$$\begin{aligned} \mathrm{E}\left\{\exp(tu^\top(\hat{\pi} - \pi))\right\} &= \mathrm{E}\left[\exp\left\{\frac{t}{n}\sum_{i=1}^n u^\top(b_i - \mathrm{E}(b_i))\right\}\right] \\ &\leq \prod_{i=1}^n \exp\left(\frac{t^2 C}{n^2}\right) = \exp\left(\frac{t^2 C}{n}\right). \end{aligned}$$

■

Lemma A.6 *Let Σ_{22} satisfy RE$(s, 3)$ with parameter $\gamma = \gamma(\Sigma_{22})$. Let Assumptions 1, 2 and 5 hold. Then with probability $1 - O(p^{-1})$, $\hat{\Sigma}_{22}$ satisfies RE$(s, 3)$ with*

$$\hat{\gamma} = \gamma(\hat{\Sigma}_{22}) \leq C\gamma$$

for some constant $C > 1$.

Proof Let $a \in \mathcal{C}(S, 3) = \{a \in \mathbb{R}^p : \|a_{S^c}\|_1 \leq 3\|a_S\|_1\}$. Let T_0 be the index set of the s largest (in absolute) elements of a. Then it holds that $a \in \mathcal{C}(T_0, 3)$, and

$$\|a\|_1 = \|a_S\|_1 + \|a_S^c\|_1 \leq 4\|a_S\|_1 \leq 4s^{1/2}\|a_S\|_2 \leq 4s^{1/2}\|a_{T_0}\|_2. \tag{A25}$$

Furthermore, following (A17), it holds that

$$\|a\|_2 \leq \|a_{T_0}\|_2 + s^{-1/2}\|a\|_1 \leq 5\|a_{T_0}\|_2. \tag{A26}$$

Consider

$$a^\top \hat{\Sigma}_{22} a = a^\top \Sigma_{22} a + a^\top(\hat{\Sigma}_{22} - \Sigma_{22})a \geq a^\top \Sigma_{22} a - |a^\top(\hat{\Sigma}_{22} - \Sigma_{22})a|. \tag{A27}$$

Following the proof of Theorem 7 and reparameterization of Δ in terms of $\Pi = \pi 1_p^\top$, we have the following decomposition

$$\begin{aligned}\hat{\Sigma}_{22} - \Sigma_{22} &= G_\tau^{-1}(T,\Delta)\circ(\hat{T}-T) + G_{\tau\tau}^{-1}(\bar{T},\Delta)\circ(\tilde{T}-T)\circ(\hat{T}-T)\\ &\quad + \{G^{-1}(\hat{T},\hat{\Delta}) - G^{-1}(\hat{T},\Delta)\}\\ &= G_\tau^{-1}(T,\Delta)\circ(\hat{T}-T) + G_{\tau\tau}^{-1}(\bar{T},\Delta)\circ(\tilde{T}-T)\circ(\hat{T}-T)\\ &\quad + \{G^{-1}(\hat{T},\hat{\Pi}) - G^{-1}(\hat{T},\Pi)\}\\ &= G_\tau^{-1}(T,\Delta)\circ(\hat{T}-T) + G_{\tau\tau}^{-1}(\bar{T},\Delta)\circ(\tilde{T}-T)\circ(\hat{T}-T)\\ &\quad + G_{\pi_1}^{-1}(\hat{T},\tilde{\pi})\circ(\hat{\Pi}-\Pi) + G_{\pi_2}^{-1}(\hat{T},\tilde{\pi})\circ(\hat{\Pi}-\Pi)^\top\\ &= G_\tau^{-1}(T,\Delta)\circ(\hat{T}-T) + G_{\pi_1}^{-1}(T,\pi)\circ(\hat{\Pi}-\Pi) + G_{\pi_2}^{-1}(T,\pi)\circ(\hat{\Pi}-\Pi)^\top\\ &\quad + G_{\tau\tau}^{-1}(\bar{T},\Delta)\circ(\tilde{T}-T)\circ(\hat{T}-T)\\ &\quad + G_{\pi_1\tau}^{-1}(\bar{T},\bar{\pi})\circ(\hat{T}-T)\circ(\hat{\Pi}-\Pi) + G_{\pi_2\tau}^{-1}(\bar{T},\bar{\pi})\circ(\hat{T}-T)\circ(\hat{\Pi}-\Pi)^\top\\ &\quad + G_{\pi_1\pi_1}^{-1}(\bar{T},\bar{\pi})\circ(\tilde{\Pi}-\Pi)\circ(\hat{\Pi}-\Pi) + G_{\pi_2\pi_2}^{-1}(\bar{T},\bar{\pi})\circ(\tilde{\Pi}-\Pi)^\top\circ(\hat{\Pi}-\Pi)^\top\\ &\quad + G_{\pi_1\pi_2}^{-1}(\bar{T},\bar{\pi})\circ(\hat{\Pi}-\Pi)^\top\circ(\hat{\Pi}-\Pi) + G_{\pi_2\pi_1}^{-1}(\bar{T},\bar{\pi})\circ(\hat{\Pi}-\Pi)\circ(\hat{\Pi}-\Pi)^\top,\end{aligned} \tag{A28}$$

where $\bar{\tau}_{jk} \in (\tilde{\tau}_{jk}, \tau_{jk})$, $\tilde{\tau}_{jk} \in (\hat{\tau}_{jk}, \tau_{jk})$, $\bar{\pi}_j \in (\tilde{\pi}_j, \pi_j)$, and $\tilde{\pi}_j \in (\hat{\pi}_j, \pi_j)$. We will use one technique to bound all first order terms in (A28), and another technique to bound all second-order terms.

Consider second-order terms in (A28). Each term is bounded in the same way using Hölder's inequality and bounds on second derivatives. Concretely, consider the term corresponding to $G_{\tau\tau}^{-1}$, that is

$$|a^\top G_{\tau\tau}^{-1}(\bar{T},\Delta)\circ(\tilde{T}-T)\circ(\hat{T}-T)a| \leq \|a\|_1^2 \|G_{\tau\tau}^{-1}(\bar{T},\Delta)\circ(\tilde{T}-T)\circ(\hat{T}-T)\|_\infty.$$

By Lemma A.11, the 2nd derivative is bounded $|G_{\tau\tau}^{-1}| \leq C$, thus, since $\tilde{\tau}_{jk}$ is between $\hat{\tau}_{jk}$ and τ_{jk},

$$|a^\top G_{\tau\tau}^{-1}(\bar{T},\Delta)\circ(\tilde{T}-T)\circ(\hat{T}-T)a| \leq C\|a\|_1^2 \|\tilde{T}-T\|_\infty \|\hat{T}-T\|_\infty \leq C\|a\|_1^2\|\hat{T}-T\|_\infty^2.$$

Using the bound (A25) on $\|a\|_1$, the condition that Σ_{22} satisfies $\mathrm{RE}(s,3)$, and by Lemma A.9, it follows that, for any constant $\eta \in (0,1)$,

$$|a^\top G_{\tau\tau}^{-1}(\bar{T},\Delta)\circ(\tilde{T}-T)\circ(\hat{T}-T)a| \leq C_1\|a_S\|_2^2 \frac{s\log(p\eta^{-1})}{n} \leq a^\top\Sigma_{22}a\ C_1\gamma\frac{s\log(p\eta^{-1})}{n}$$

with probability at least $1-\eta$. All the remaining 2nd order terms have the same bound as all the second derivatives are bounded, that is $|G_{\pi_j\pi_j}^{-1}| \leq C$ by Lemma A.12, $|G_{\pi_j\tau}^{-1}| \leq C$ by Lemma A.14, and $|G_{\pi_j\pi_k}^{-1}| \leq C$ by Lemma A.13. Also $\|\hat{\pi}-\pi\|_\infty \leq C_1\sqrt{\log(p\eta^{-1})/n}$ with probability at least $1-\eta$ by Hoeffding's inequality combined with union bound, and $\|\hat{T}-T\|_\infty \leq C_2\sqrt{\log(p\eta^{-1})/n}$ with probability at least $1-\eta$ by Lemma A.9.

Consider first-order terms in (A28). Each term is bounded in the same way using sub-gaussian properties in Lemma A.5 (for $\hat{\pi}$) and Lemma A.8 (for $\hat{T}$) combined with the fact that the first derivatives are both bounded and fixed. Concretely, consider the term corresponding to G_τ^{-1}, that is

$$\begin{aligned}\left|a^\top G_\tau^{-1}(T,\Delta)\circ(\hat{T}-T)a\right| &= \left|\left(\sum_{j=1}^p a_j e_j\right)^\top G_\tau^{-1}(T,\Delta)\circ(\hat{T}-T)a\right| \\ &= \left|\sum_{j=1}^p a_j\left\{e_j^\top G_\tau^{-1}(T,\Delta)\circ(\hat{T}-T)a\right\}\right| \\ &\leq \|a\|_1 \max_{1\leq j\leq p}\left|e_j^\top G_\tau^{-1}(T,\Delta)\circ(\hat{T}-T)a\right| \\ &\leq 4\sqrt{s}\|a_{T_0}\|_2 \max_{1\leq j\leq p}|e_j^\top(\hat{T}-T)b_j|,\end{aligned}$$

where $e_j \in \mathbb{R}^p$ be the vector with 1 in the jth component and 0 otherwise, $b_j = a \circ G_\tau^{-1}(T_j, \Delta)$, and the last inequality follows from (A25). From Theorem 6 of Yoon et al. (2020), $|G_\tau^{-1}| \leq C$ for some constant $C > 0$, hence using (A26)

$$\|b_j\|_2 \leq C\|a\|_2 \leq C_1\|a_{T_0}\|_2.$$

Combining this bound with Lemma A.9, and following the proof of Lemma A.1 gives, with probability at least $1-\eta$,

$$\max_j |e_j^\top(\hat{T}-T)b_j| \leq C_2\|a_{T_0}\|_2\sqrt{\frac{\log(p\eta^{-1})}{n}},$$

and, using that Σ_{22} satisfies $RE(s,3)$ gives

$$|a^\top G_\tau^{-1}(T,\Delta)\circ(\hat{T}-T)a| \leq C\|a_{T_0}\|_2^2\sqrt{\frac{s\log(p\eta^{-1})}{n}} \leq a^\top\Sigma_{22}a\ C\gamma\sqrt{\frac{\log(p\eta^{-1})}{n}}.$$

All the remaining first order terms have the same bound as the first derivatives $G_{\pi_j}^{-1}$ are fixed and bounded by Lemma A.10, and $\hat{\pi}$ satisfy Lemma A.5.

Combining the bounds on the first and second-order terms coupled with Assumption 5 gives

$$\left|a^\top(\hat{\Sigma}_{22}-\Sigma_{22})a\right| \leq a^\top\Sigma_{22}aC_3\gamma\sqrt{\frac{s\log(p\eta^{-1})}{n}}$$

with probability at least $1-\eta$. Combining this bound with (A27) gives

$$a^\top\hat{\Sigma}_{22}a \geq a^\top\Sigma_{22}a\left\{1 - C_3\gamma\sqrt{\frac{s\log(p\eta^{-1})}{n}}\right\}.$$

with probability at least $1-\eta$. Under the scaling of Assumption 5, $C_3\gamma\sqrt{s\log p/n} = o(1)$, thus it follows that with probability at least $1-\eta$, $\gamma(\hat{\Sigma}) \leq C\gamma$ for some constant $C > 1$. ■

C.4 Supporting lemmas based on existing results

Lemma A.7 (Barber and Kolar (2018) Lemma 4.5) *Let $Z \sim \mathrm{N}_p(\mu, \Sigma)$. Then* $\operatorname{sign}(Z) - \mathrm{E}\{\operatorname{sign}(Z)\}$ *is* $\mathsf{C}(\Sigma)$*-subgaussian.*

Lemma A.8 (Barber and Kolar (2018) Lemma E.2) *For fixed u and v with $\|u\|_2, \|v\|_2 \leq 1$, for any $|t| \leq n/C$,*

$$\mathrm{E}\left[\exp\left\{tu^\top\left(\hat{T} - T\right)v\right\}\right] \leq \exp\left(\frac{t^2C^2}{n}\right).$$

Lemma A.9 (De la Pena and Giné (2012) Theorem 4.1.8) *For any $\eta \in (0, 1)$,*

$$\|\hat{T} - T\|_\infty \leq \sqrt{\frac{4\log\left(2\binom{p}{2}/\eta\right)}{n}}$$

with probability at least $1 - \eta$.

C.5 Bounds on partial derivatives of the inverse bridge function

Lemma A.10 *Let $G^{-1}(\tau)$ be the inverse bridge function for TT case, where $\tau = G_{TT}(r; \Delta_j, \Delta_k)$. Under Assumptions 1–2, $|\partial G^{-1}(\tau)/\partial\pi_j| \leq C$ and $|\partial G^{-1}(\tau)/\partial\pi_k| \leq C$ for some constant $C > 0$.*

Proof By the multivariate chain rule, we have

$$\begin{aligned}\frac{\partial G^{-1}(\tau)}{\partial\pi_j} = \frac{\partial G^{-1}(\tau)}{\partial\tau}\frac{\partial\tau}{\partial\Delta_j}\frac{\partial\Delta_j}{\partial\pi_j} &= \frac{\partial G^{-1}(\tau)}{\partial\tau}\frac{\partial G(r;\Delta_j,\Delta_k)}{\partial\Delta_j}\frac{\partial\Delta_j}{\partial\pi_j} \\ &:= A_1A_2A_3.\end{aligned} \tag{A29}$$

By Theorem 6 in Yoon et al. (2020), $|A_1| \leq C$. By Lemma A.15, A_2 is bounded. By Lemma A.21, A_3 is bounded. The proof for π_k is analogous. ■

Lemma A.11 *Let $G^{-1}(\tau)$ be the inverse bridge function for TT case, where $\tau = G_{TT}(r; \Delta_j, \Delta_k)$. Under Assumptions 1– 2, $|G^{-1}_{\tau\tau} = \partial^2 G^{-1}(\tau)/\partial\tau^2| \leq C$ for some constant $C > 0$ independent of r, Δ_j, Δ_k.*

Proof Let $h(r) = \partial G(r; \Delta_j, \Delta_k)/\partial r$ and consider

$$\begin{aligned}\frac{\partial^2 G^{-1}(\tau)}{\partial\tau^2} &= \frac{\partial}{\partial\tau}\left\{\frac{\partial G(r;\Delta_j,\Delta_k)}{\partial r}\right\}^{-1} = \frac{\partial}{\partial r}\left\{\frac{\partial G(r;\Delta_j,\Delta_k)}{\partial r}\right\}^{-1}\frac{\partial r}{\partial\tau} \\ &= \frac{\partial}{\partial r}\left\{\frac{1}{h(r)}\right\}\left(\frac{\partial\tau}{\partial r}\right)^{-1} = \frac{\partial}{\partial r}\left\{\frac{1}{h(r)}\right\}\left\{\frac{\partial G(r;\Delta_j,\Delta_k)}{\partial r}\right\}^{-1} \\ &= -\frac{1}{h(r)^2}\frac{\partial h(r)}{\partial r}\frac{1}{h(r)} = -\frac{1}{h(r)^3}\frac{\partial h(r)}{\partial r}.\end{aligned}$$

By Theorem 6 of Yoon et al. (2020), $h(r)$ is positive and bounded from below by a positive constant independent of r, Δ_j, Δ_k. By Lemma A.19, $|\partial h(r)/\partial r|$ is bounded above by a positive constant. Thus, we have $|\partial^2 G^{-1}(\tau)/\partial\tau^2| < C$ for some $C > 0$. ■

Lemma A.12 *Let $G^{-1}(\tau)$ be the inverse bridge function for TT case, where $\tau = G_{TT}(r, \Delta_j, \Delta_k)$. Under Assumptions 1–2, $|G^{-1}_{\pi_j\pi_j} = \partial^2 G^{-1}(\tau)/\partial\pi_j^2| \leq C$ for some constant $C > 0$ independent of r, Δ_j, Δ_k.*

Proof Let $h(r) = \partial G(r; \Delta_j, \Delta_k)/\partial r$ so that $\partial G^{-1}(\tau)/\partial\tau = (\partial G(r; \Delta_j, \Delta_k)/\partial r)^{-1} = 1/h(r)$. By (A29) and multivariate chain rule, we have

$$
\begin{aligned}
\frac{\partial^2 G^{-1}(\tau)}{\partial\pi_j^2} &= \frac{\partial}{\partial\pi_j}\left(\frac{1}{h(r)}\frac{\partial G(r;\Delta_j,\Delta_k)}{\partial\Delta_j}\frac{\partial\Delta_j}{\partial\pi_j}\right)\\
&= \left[\frac{\partial}{\partial\pi_j}\left\{\frac{1}{h(r)}\right\}\right]\frac{\partial G(r;\Delta_j,\Delta_k)}{\partial\Delta_j}\frac{\partial\Delta_j}{\partial\pi_j} + \frac{1}{h(r)}\left[\frac{\partial^2 G(r;\Delta_j,\Delta_k)}{\partial\pi_j\partial\Delta_j}\right]\frac{\partial\Delta_j}{\partial\pi_j}\\
&\quad + \frac{1}{h(r)}\frac{\partial G(r;\Delta_j,\Delta_k)}{\partial\Delta_j}\left[\frac{\partial^2\Delta_j}{\partial\pi_j^2}\right]\\
&= \left[\frac{\partial\{h(r)\}^{-1}}{\partial\Delta_j}\frac{\partial\Delta_j}{\partial\pi_j}\right]\frac{\partial G(r;\Delta_j,\Delta_k)}{\partial\Delta_j}\frac{\partial\Delta_j}{\partial\pi_j} + \frac{1}{h(r)}\left[\frac{\partial^2 G(r;\Delta_j,\Delta_k)}{\partial\Delta_j^2}\frac{\partial\Delta_j}{\partial\pi_j}\right]\frac{\partial\Delta_j}{\partial\pi_j}\\
&\quad + \frac{1}{h(r)}\frac{\partial G(r;\Delta_j,\Delta_k)}{\partial\Delta_j}\left[\frac{\partial^2\Delta_j}{\partial\pi_j^2}\right]\\
&=\frac{\partial\{h(r)\}^{-1}}{\partial\Delta_j}\frac{\partial G(r;\Delta_j,\Delta_k)}{\partial\Delta_j}\left(\frac{\partial\Delta_j}{\partial\pi_j}\right)^2 + \frac{1}{h(r)}\frac{\partial^2 G(r;\Delta_j,\Delta_k)}{\partial\Delta_j^2}\left(\frac{\partial\Delta_j}{\partial\pi_j}\right)^2\\
&\quad + \frac{1}{h(r)}\frac{\partial G(r;\Delta_j,\Delta_k)}{\partial\Delta_j}\frac{\partial^2\Delta_j}{\partial\pi_j^2}.
\end{aligned}
$$

We next show that each term is bounded.

Consider $\partial\{h(r)\}^{-1}/\partial\Delta_j$. By the multivariate chain rule,

$$
\frac{\partial\{h(r)\}^{-1}}{\partial\Delta_j} = -\frac{1}{h(r)^2}\frac{\partial h(r)}{\partial\Delta_j} = -\frac{1}{h(r)^2}\frac{\partial^2 G(r;\Delta_j,\Delta_k)}{\partial\Delta_j\partial r}.
$$

The term $|\partial^2 G(r; \Delta_j, \Delta_k)/\partial\Delta_j\partial r|$ is bounded from above by Lemma A.16, and $|1/h(r)|$ is bounded from above by Theorem 6 in Yoon et al. (2020). Furthermore, $|\partial G(r; \Delta_j, \Delta_k)/\partial\Delta_j|$ is bounded by Lemma A.15, $|\partial^2 G(r; \Delta_j, \Delta_k)/\partial\Delta_j^2|$ is bounded by Lemma A.17, and $|\partial\Delta_j/\partial\pi_j|$, $|\partial^2\Delta_j/\partial\pi_j^2|$ are both bounded by Lemma A.21. This concludes the proof. ■

Lemma A.13 *Let $G^{-1}(\tau)$ be the inverse bridge function for TT case, where $\tau = G_{TT}(r, \Delta_j, \Delta_k)$. Under Assumptions 1–2, $|\partial^2 G^{-1}(\tau)/\partial\pi_k\pi_j| \leq C$ for some constant $C > 0$ independent of r, Δ_j, Δ_k.*

Proof Let $h(r) = \partial G(r;\Delta_j,\Delta_k)/\partial r$ so that $\partial G^{-1}(\tau)/\partial\tau = (\partial G(r;\Delta_j,\Delta_k)/\partial r)^{-1} = 1/h(r)$. By (A29) and multivariate chain rule, we have

$$\begin{aligned}\frac{\partial^2 G^{-1}(\tau)}{\partial\pi_k\pi_j} =& \frac{\partial}{\partial\pi_k}\left(\frac{1}{h(r)}\frac{\partial G(r;\Delta_j,\Delta_k)}{\partial\Delta_j}\frac{\partial\Delta_j}{\partial\pi_j}\right)\\ &= \left[\frac{\partial}{\partial\pi_k}\left\{\frac{1}{h(r)}\right\}\right]\frac{\partial G(r;\Delta_j,\Delta_k)}{\partial\Delta_j}\frac{\partial\Delta_j}{\partial\pi_j} + \frac{1}{h(r)}\left[\frac{\partial^2 G(r;\Delta_j,\Delta_k)}{\partial\pi_k\partial\Delta_j}\right]\frac{\partial\Delta_j}{\partial\pi_j}\\ &+ \frac{1}{h(r)}\frac{\partial G(r;\Delta_j,\Delta_k)}{\partial\Delta_j}\left[\frac{\partial^2\Delta_j}{\partial\pi_k\partial\pi_j}\right].\end{aligned}$$

As $\partial^2\Delta_j/\partial\pi_k\partial\pi_j = 0$, we further have that

$$\begin{aligned}\frac{\partial^2 G^{-1}(\tau)}{\partial\pi_k\pi_j} &= \left[\frac{\partial\{h(r)\}^{-1}}{\partial\Delta_k}\frac{\partial\Delta_k}{\partial\pi_k}\right]\frac{\partial G(r;\Delta_j,\Delta_k)}{\partial\Delta_j}\frac{\partial\Delta_j}{\partial\pi_j} + \frac{1}{h(r)}\left[\frac{\partial^2 G(r;\Delta_j,\Delta_k)}{\partial\Delta_k\partial\Delta_j}\frac{\partial\Delta_k}{\partial\pi_k}\right]\frac{\partial\Delta_j}{\partial\pi_j}\\ &= \frac{\partial\{h(r)\}^{-1}}{\partial\Delta_k}\frac{\partial G(r;\Delta_j,\Delta_k)}{\partial\Delta_j}\left(\frac{\partial\Delta_k}{\partial\pi_k}\frac{\partial\Delta_j}{\partial\pi_j}\right) + \frac{1}{h(r)}\frac{\partial^2 G(r;\Delta_j,\Delta_k)}{\partial\Delta_k\partial\Delta_j}\left(\frac{\partial\Delta_k}{\partial\pi_k}\frac{\partial\Delta_j}{\partial\pi_j}\right),\end{aligned}$$

where $|\partial\Delta_j/\partial\pi_j|$ and $|\partial\Delta_k/\partial\pi_k|$ are bounded above by some constant $C > 0$ by Lemma A.21. Thus, by the triangle inequality,

$$\begin{aligned}\left|\frac{\partial^2 G^{-1}(\tau)}{\partial\pi_k\pi_j}\right| &\le \left|\frac{\partial\{h(r)\}^{-1}}{\partial\Delta_k}\frac{\partial G(r;\Delta_j,\Delta_k)}{\partial\Delta_j}\left(\frac{\partial\Delta_k}{\partial\pi_k}\frac{\partial\Delta_j}{\partial\pi_j}\right)\right| + \left|\frac{1}{h(r)}\frac{\partial^2 G(r;\Delta_j,\Delta_k)}{\partial\Delta_k\partial\Delta_j}\left(\frac{\partial\Delta_k}{\partial\pi_k}\frac{\partial\Delta_j}{\partial\pi_j}\right)\right|\\ &\le C^2\left|\frac{\partial\{h(r)\}^{-1}}{\partial\Delta_k}\frac{\partial G(r;\Delta_j,\Delta_k)}{\partial\Delta_j}\right| + C^2\left|\frac{1}{h(r)}\frac{\partial^2 G(r;\Delta_j,\Delta_k)}{\partial\Delta_k\partial\Delta_j}\right|\end{aligned}$$

We next show that each term is bounded.

Consider $\partial\{h(r)\}^{-1}/\partial\Delta_k$. By the multivariate chain rule,

$$\frac{\partial\{h(r)\}^{-1}}{\partial\Delta_k} = -\frac{1}{h(r)^2}\frac{\partial h(r)}{\partial\Delta_k} = -\frac{1}{h(r)^2}\frac{\partial^2 G(r;\Delta_j,\Delta_k)}{\partial\Delta_k\partial r}.$$

The term $|\partial^2 G(r;\Delta_j,\Delta_k)/\partial\Delta_k\partial r|$ is bounded from above by Lemma A.16, and $|1/h(r)|$ is bounded from above by Theorem 6 in Yoon et al. (2020). Furthermore, $|\partial G(r;\Delta_j,\Delta_k)/\partial\Delta_j|$ is bounded by Lemma A.15, $|\partial^2 G(r;\Delta_j,\Delta_k)/\partial\Delta_k\partial\Delta_j|$ is bounded by Lemma A.18. This concludes the proof. ■

Lemma A.14 *Let $G^{-1}(\tau)$ be the inverse bridge function for TT case, where $\tau = G_{TT}(r,\Delta_j,\Delta_k)$. Under Assumptions 1–2, $|\partial^2 G^{-1}(\tau)/\partial\pi_j\partial\tau| \le C$ for some constant $C > 0$ independent of r, Δ_j, Δ_k.*

Proof Let $h(r) = \partial G(r;\Delta_j,\Delta_k)/\partial r$ so that $\partial G^{-1}(\tau)/\partial\tau = (\partial G(r;\Delta_j,\Delta_k)/\partial r)^{-1} = 1/h(r)$. By the multivariate chain rule,

$$\begin{aligned}\frac{\partial^2 G^{-1}(\tau)}{\partial\pi_j\partial\tau} =& \frac{\partial}{\partial\pi_j}\frac{\partial G^{-1}(\tau)}{\partial\tau} = \frac{\partial}{\partial\pi_j}\left\{\frac{1}{h(r)}\right\} = \frac{\partial}{\partial\Delta_j}\left\{\frac{1}{h(r)}\right\}\frac{\partial\Delta_j}{\partial\pi_j}\\ &= -\frac{1}{h(r)^2}\frac{\partial h(r)}{\partial\Delta_j}\frac{\partial\Delta_j}{\partial\pi_j} = -\frac{1}{h(r)^2}\frac{\partial G(r;\Delta_j,\Delta_k)}{\partial\Delta_j\partial r}\frac{\partial\Delta_j}{\partial\pi_j}.\end{aligned}$$

The terms $|\partial\Delta_j/\partial\pi_j|$, $|1/h(r)^2|$, and $|\partial^2 G(r;\Delta_j,\Delta_k)/\partial\Delta_j\partial r|$ are bounded above by constants by Lemma A.21, Theorem 6 of Yoon et al. (2020), and Lemma A.16, respectively. Thus, for some constant $C>0$, we have

$$\left|\frac{\partial^2 G^{-1}(\tau)}{\partial\pi_j\partial\tau}\right| \leq C.$$

■

C.6 Bounds on the partial derivatives of the bridge function

Here we bound partial derivatives of the bridge function $G(r,\Delta_j,\Delta_k)$ for TT case, where

$$G(r,\Delta_j,\Delta_k) = -2\Phi_4(-\Delta_j,-\Delta_k,0,0;\Sigma_{4a}) + 2\Phi_4(-\Delta_j,-\Delta_k,0,0;\Sigma_{4b}).$$

As the bridge function consists of two 4-dimensional Gaussian distribution functions, we will show that, whether $\Sigma_4=\Sigma_{4a}$ or $\Sigma_4=\Sigma_{4b}$, the absolute values of partial derivatives of $\Phi_4(-\Delta_j,-\Delta_k,0,0;\Sigma_4)$ are bounded from above.

Lemma A.15 *Under Assumptions 1–2, $|\partial G(r;\Delta_j,\Delta_k)/\partial\Delta_j|$ and $|\partial G(r;\Delta_j,\Delta_k)/\partial\Delta_k|$ are bounded above by some constant $C>0$ independent from r, Δ_j, Δ_k.*

Proof By the Leibniz rule,

$$\begin{aligned}\frac{\partial}{\partial\Delta_j}\Phi_4(-\Delta_j,-\Delta_k,0,0;\Sigma_4) &= \frac{\partial}{\partial\Delta_j}\int_{-\infty}^{0}\int_{-\infty}^{0}\int_{-\infty}^{-\Delta_k}\int_{-\infty}^{-\Delta_j}\phi(z_1,z_2,z_3,z_4)\mathrm{d}z_1\mathrm{d}z_2\mathrm{d}z_3\mathrm{d}z_4\\ &= (-1)\int_{-\infty}^{0}\int_{-\infty}^{0}\int_{-\infty}^{-\Delta_k}\phi(-\Delta_j,z_2,z_3,z_4)\mathrm{d}z_2\mathrm{d}z_3\mathrm{d}z_4.\end{aligned}$$

Thus, regardless of $\Sigma_4=\Sigma_{4a}$ or $\Sigma_4=\Sigma_{4b}$,

$$\begin{aligned}\left|\frac{\partial}{\partial\Delta_j}\Phi_4(-\Delta_j,-\Delta_k,0,0;\Sigma_4)\right| &= \int_{-\infty}^{0}\int_{-\infty}^{0}\int_{-\infty}^{-\Delta_k}\phi(-\Delta_j,z_2,z_3,z_4)\mathrm{d}z_2\mathrm{d}z_3\mathrm{d}z_4\\ &= \phi(-\Delta_j)\int_{-\infty}^{0}\int_{-\infty}^{0}\int_{-\infty}^{-\Delta_k}\phi(z_2,z_3,z_4\mid -\Delta_j)\mathrm{d}z_2\mathrm{d}z_3\mathrm{d}z_4,\end{aligned}$$

where $\phi(z_2,z_3,z_4\mid -\Delta_j)$ is the conditional pdf given $Z_1=-\Delta_j$. Therefore, the three-dimensional integral above corresponds to a probability (and is bounded by one), leading to

$$\left|\frac{\partial}{\partial\Delta_j}\Phi_4(-\Delta_j,-\Delta_k,0,0;\Sigma_4)\right| \leq \phi(-\Delta_j) \leq \phi(0) = 1/\sqrt{2\pi}.$$

The proof for Δ_k follows analogously. ■

Lemma A.16 *Under Assumptions 1–2, $|\partial^2 G(r;\Delta_j,\Delta_k)|/\partial r\partial\Delta_j|$ and $|\partial^2 G(r;\Delta_j,\Delta_k)|/\partial r\partial\Delta_k|$ are bounded above by some constant $C>0$.*

Proof We start from the partial derivative with respect to Δ_j given in Lemma A.15 as

$$\frac{\partial}{\partial \Delta_j}\Phi_4(-\Delta_j,-\Delta_k,0,0;\Sigma_4)=(-1)\int_{-\infty}^{0}\int_{-\infty}^{0}\int_{-\infty}^{-\Delta_k}\phi(-\Delta_j,z_2,z_3,z_4;\Sigma_4)\mathrm{d}z_2\mathrm{d}z_3\mathrm{d}z_4.$$

Let $\Sigma_4=[\rho_{jk}]_{1\le j,k\le 4}$ and consider the following multivariate chain rule:

$$\begin{aligned}\frac{\partial}{\partial r}&\int_{-\infty}^{0}\int_{-\infty}^{0}\int_{-\infty}^{-\Delta_k}\phi(-\Delta_j,z_2,z_3,z_4;\Sigma_4)\mathrm{d}z_2\mathrm{d}z_3\mathrm{d}z_4\\ &=\sum_{j<k}\left\{\frac{\partial}{\partial\rho_{jk}}\int_{-\infty}^{0}\int_{-\infty}^{0}\int_{-\infty}^{-\Delta_k}\phi(-\Delta_j,z_2,z_3,z_4;\Sigma_4)\mathrm{d}z_2\mathrm{d}z_3\mathrm{d}z_4\frac{\partial\rho_{jk}}{\partial r}\right\}\\ &=\sum_{j<k}\left\{\int_{-\infty}^{0}\int_{-\infty}^{0}\int_{-\infty}^{-\Delta_k}\frac{\partial\phi(-\Delta_j,z_2,z_3,z_4;\Sigma_4)}{\partial\rho_{jk}}\mathrm{d}z_2\mathrm{d}z_3\mathrm{d}z_4\frac{\partial\rho_{jk}}{\partial r}\right\}.\end{aligned}$$

In the above, we only consider partial derivatives with respect to ρ_{12}, ρ_{14}, ρ_{23}, and ρ_{34} because ρ_{13}, ρ_{24} do not involve r whether $\Sigma_4=\Sigma_{4a}$ or $\Sigma_4=\Sigma_{4b}$, i.e., $\partial\rho_{jk}/\partial r=0$.

Consider the case $(j,k)=(2,3)$. By Plackett (1954),

$$\begin{aligned}\int_{-\infty}^{0}\int_{-\infty}^{0}\int_{-\infty}^{-\Delta_k}&\frac{\partial\phi(-\Delta_j,z_2,z_3,z_4;\Sigma_4)}{\partial\rho_{23}}\mathrm{d}z_2\mathrm{d}z_3\mathrm{d}z_4\\ &=\int_{-\infty}^{0}\int_{-\infty}^{0}\int_{-\infty}^{-\Delta_k}\frac{\partial^2\phi(-\Delta_j,z_2,z_3,z_4;\Sigma_4)}{\partial z_2\partial z_3}\mathrm{d}z_2\mathrm{d}z_3\mathrm{d}z_4\\ &=\int_{-\infty}^{0}\phi(-\Delta_j,-\Delta_k,0,z_4;\Sigma_4)\mathrm{d}z_4\\ &=\int_{-\infty}^{0}\phi(z_4\mid-\Delta_j,-\Delta_k,0)\phi(-\Delta_j,-\Delta_k,0)\mathrm{d}z_4\\ &=\phi(-\Delta_j,-\Delta_k,0)\int_{-\infty}^{0}\phi(z_4\mid\Delta_j,-\Delta_k,0)\mathrm{d}z_4,\end{aligned}$$

where $\phi(z_4|-\Delta_j,-\Delta_k,0)$ is the conditional pdf given $Z_1=-\Delta_j$, $Z_2=-\Delta_k$, and $Z_3=0$. Therefore, above integral corresponds to a probability (and is bounded by one), leading to

$$\int_{-\infty}^{0}\int_{-\infty}^{0}\int_{-\infty}^{-\Delta_k}\frac{\partial\phi(-\Delta_j,z_2,z_3,z_4;\Sigma_4)}{\partial\rho_{23}}\mathrm{d}z_2\mathrm{d}z_3\mathrm{d}z_4\le\phi(-\Delta_j,-\Delta_k,0)\le|\Sigma_4|^{-1/2},$$

where above inequalities hold because $\phi(-\Delta_j,-\Delta_k,0)\le\phi(0,0,0)\le|\Sigma_3|^{-1/2}\le|\Sigma_4|^{-1/2}$ and $\Sigma_3=\mathrm{var}\{(Z_1,Z_2,Z_3)\}$. As Lemma A.20 provides that $|\Sigma_4|^{-1/2}\le C$ for some constant $C>0$, we have the desired result. The case $(j,k)=(3,4)$ is similar with the same bound.

For $(j,k)=(1,2)$, again by Plackett (1954), we have

$$\int_{-\infty}^{0}\int_{-\infty}^{0}\int_{-\infty}^{-\Delta_k}\frac{\partial\phi(-\Delta_j,z_2,z_3,z_4)}{\partial\rho_{12}}\mathrm{d}z_2\mathrm{d}z_3\mathrm{d}z_4=-\int_{-\infty}^{0}\int_{-\infty}^{0}\frac{\partial\phi(-\Delta_j,-\Delta_k,z_3,z_4)}{\partial\Delta_j}\mathrm{d}z_3\mathrm{d}z_4.$$

For notational convenience, let $y = (-\Delta_j, -\Delta_k, z_3, z_4)^\top = (y_1, y_2, y_3, y_4)^\top$ and write

$$\int_{-\infty}^{0}\int_{-\infty}^{0} \frac{\partial\phi(-\Delta_j, -\Delta_k, z_3, z_4)}{\partial\Delta_j}\mathrm{d}z_3\mathrm{d}z_4 = \int_{-\infty}^{0}\int_{-\infty}^{0} \frac{\partial\phi(y_1, y_2, y_3, y_4)}{\partial y_1}\mathrm{d}y_3\mathrm{d}y_4$$
$$= \int_{-\infty}^{0}\int_{-\infty}^{0} (-\omega_1^\top y)\phi(y_1, y_2, y_3, y_4)\mathrm{d}y_3\mathrm{d}y_4, \quad \text{(A30)}$$

where $\omega_i^\top$ is the ith row of Σ_4^{-1}. Then, by extending the range of integrations, the absolute value of (A30) is bounded above as

$$\left|\int_{-\infty}^{0}\int_{-\infty}^{0} (-\omega_1^\top y)\phi(y_1, y_2, y_3, y_4)\mathrm{d}y_3\mathrm{d}y_4\right| \le \int_{-\infty}^{0}\int_{-\infty}^{0} \left|\omega_1^\top y\right| \phi(y_1, y_2, y_3, y_4)\mathrm{d}y_3\mathrm{d}y_4$$
$$\le \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left|\omega_1^\top y\right| \phi(y_1, y_2, y_3, y_4)\mathrm{d}y_3\mathrm{d}y_4.$$

By the triangle inequality,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left|\omega_1^\top y\right|\phi(y_1, y_2, y_3, y_4)\mathrm{d}y_3\mathrm{d}y_4$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \sum_{i'=1}^{4} |\omega_{1i'}y_{i'}|\, \phi(y_1, y_2, y_3, y_4)\mathrm{d}y_3\mathrm{d}y_4$$
$$\le \sum_{i'=1}^{2} |\omega_{1i'}y_{i'}|\phi(y_1, y_2) + \sum_{i'=3}^{4} |\omega_{1i'}| \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |y_{i'}|\phi(y_1, y_2, y_3, y_4)\mathrm{d}y_3\mathrm{d}y_4$$
$$\le |\Sigma_4|^{-1/2}\left\{\sum_{i'=1}^{2} |\omega_{1i'}y_{i'}| + \sum_{i'=3}^{4} |\omega_{1i'}| \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |y_{i'}|\phi(y_3, y_4 \mid y_1, y_2)\mathrm{d}y_3\mathrm{d}y_4\right\},$$

where the last inequality holds as $\phi(y_1, y_2) \le |\Sigma_2|^{-1/2} \le |\Sigma_4|^{-1/2}$ and $\Sigma_2 = \mathrm{var}\{(Y_1, Y_2)\}$. Under Assumption 2, $|y_1| = |\Delta_j| \le M$,$|y_2| = |\Delta_k| \le M$. By Lemma A.20, whether $\Sigma_4 = \Sigma_{4a}$ or $\Sigma_4 = \Sigma_{4b}$, $|\Sigma_4|^{-1/2}$ is bounded above and all elements of $\Sigma_4^{-1} = [\omega_{\ell\ell'}]_{1\le\ell\ell'\le 4}$ are all bounded above. Thus, we have

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left|\omega_1^\top y\right|\phi(y_1, y_2, y_3, y_4)\mathrm{d}y_3\mathrm{d}y_4 \le C_0 + C_1\sum_{i'=3}^{4}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |y_{i'}|\phi(y_3, y_4 \mid y_1, y_2)\mathrm{d}y_3\mathrm{d}y_4.$$

By Lemma A.23, for some constant $C > 0$,

$$\sum_{i'=3}^{4}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |y_{i'}|\phi(y_3, y_4 \mid y_1, y_2)\mathrm{d}y_3\mathrm{d}y_4 = \sum_{i'=3}^{4} \mathrm{E}(Y_{i'} \mid Y_1 = y_1, Y_2 = y_2) \le C.$$

This concludes the proof and the proof for Δ_k is analogous. ■

Lemma A.17 *Under Assumptions 1 and 2,* $|\partial^2 G(r; \Delta_j, \Delta_k)/\partial\Delta_j^2|$ *and* $|\partial^2 G(r; \Delta_j, \Delta_k)/\partial\Delta_k^2|$ *are bounded above by some constant* $C > 0$*.*

Proof From the proof of Lemma A.15, we have

$$\frac{\partial}{\partial \Delta_j}\Phi_4(-\Delta_j,-\Delta_k,0,0;\Sigma_4)=(-1)\int_{-\infty}^{0}\int_{-\infty}^{0}\int_{-\infty}^{-\Delta_k}\phi(-\Delta_j \mid z_2,z_3,z_4)\phi(z_2,z_3,z_4)\mathrm{d}z_2\mathrm{d}z_3\mathrm{d}z_4.$$

By interchanging differentiation and integration,

$$\begin{aligned}\frac{\partial^2}{\partial \Delta_j^2}\Phi_4(-\Delta_j,-\Delta_k,0,0;\Sigma_4)=&(-1)\int_{-\infty}^{0}\int_{-\infty}^{0}\int_{-\infty}^{-\Delta_k}\frac{\partial}{\partial \Delta_j}\phi(-\Delta_j \mid z_2,z_3,z_4)\phi(z_2,z_3,z_4)\mathrm{d}z_2\mathrm{d}z_3\mathrm{d}z_4\\ =&\int_{-\infty}^{0}\int_{-\infty}^{0}\int_{-\infty}^{-\Delta_k}\frac{\Delta_j+\mu}{v^2}\phi(-\Delta_j \mid z_2,z_3,z_4)\phi(z_2,z_3,z_4)\mathrm{d}z_2\mathrm{d}z_3\mathrm{d}z_4,\end{aligned}$$

where $\mathrm{E}(z_1 \mid z_2,z_3,z_4)=\mu$ and $\mathrm{var}(z_1 \mid z_2,z_3,z_4)=v^2$ as in Lemma A.22. Thus

$$\left|\frac{\partial^2}{\partial \Delta_j^2}\Phi_4(-\Delta_j,-\Delta_k,0,0;\Sigma_4)\right| \le \int_{-\infty}^{0}\int_{-\infty}^{0}\int_{-\infty}^{-\Delta_k}\left|\frac{\Delta_j}{v^2}\right|\phi(-\Delta_j \mid z_2,z_3,z_4)\phi(z_2,z_3,z_4)\mathrm{d}z_2\mathrm{d}z_3\mathrm{d}z_4 \tag{A31}$$

$$+\int_{-\infty}^{0}\int_{-\infty}^{0}\int_{-\infty}^{-\Delta_k}\left|\frac{\mu}{v^2}\right|\phi(-\Delta_j \mid z_2,z_3,z_4)\phi(z_2,z_3,z_4)\mathrm{d}z_2\mathrm{d}z_3\mathrm{d}z_4. \tag{A32}$$

Consider the first term (A31). Following the proof of Lemma A.15,

$$\begin{aligned}&\int_{-\infty}^{0}\int_{-\infty}^{0}\int_{-\infty}^{-\Delta_k}\left|\frac{\Delta_j}{v^2}\right|\phi(-\Delta_j \mid z_2,z_3,z_4)\phi(z_2,z_3,z_4)\mathrm{d}z_2\mathrm{d}z_3\mathrm{d}z_4\\ &\qquad=\left|\frac{\Delta_j}{v^2}\right|\int_{-\infty}^{0}\int_{-\infty}^{0}\int_{-\infty}^{-\Delta_k}\phi(-\Delta_j,z_2,z_3,z_4)\mathrm{d}z_2\mathrm{d}z_3\mathrm{d}z_4\\ &\qquad\le\left|\frac{\Delta_j}{v^2}\right|\frac{1}{\sqrt{2\pi}}\le C,\end{aligned} \tag{A33}$$

where the last inequality holds as $|\Delta_j|\le M$ under Assumption 2, and v^2 is bounded below by Lemma A.22.

Consider the second term (A32). Let $z_{-1}=(z_2,z_3,z_4)^\top$ and write $\mu=rz_2+z_3/2^{1/2}-rz_4/2^{1/2}=u^\top z_{-1}$ as in Lemma A.22. Then, since $\phi(-\Delta_j \mid z_2,z_3,z_4)\le 1/\sqrt{2\pi v^2}$,

$$\begin{aligned}&\int_{-\infty}^{0}\int_{-\infty}^{0}\int_{-\infty}^{-\Delta_k}\left|\frac{\mu}{v^2}\right|\phi(-\Delta_j \mid z_2,z_3,z_4)\phi(z_2,z_3,z_4)\mathrm{d}z_2\mathrm{d}z_3\mathrm{d}z_4\\ &\qquad\le\frac{1}{v^3\sqrt{2\pi}}\int_{-\infty}^{0}\int_{-\infty}^{0}\int_{-\infty}^{-\Delta_k}|u^\top \boldsymbol{z}_{-1}|\phi(z_2,z_3,z_4)\mathrm{d}z_2\mathrm{d}z_3\mathrm{d}z_4.\end{aligned}$$

Since $u^\top z_{-1}\sim \mathrm{N}(0,\frac{1+r^2}{2})$ and $|u^\top z_{-1}|$ follows the folded Gaussian with mean $\mathrm{E}|u^\top z_{-1}|=\sqrt{(1+r^2)/\pi}$, we further have that

$$\begin{aligned}\int_{-\infty}^{0}\int_{-\infty}^{0}\int_{-\infty}^{-\Delta_k}\left|\frac{\mu}{v^2}\right|\phi(-\Delta_j \mid z_2,z_3,z_4)\phi(z_2,z_3,z_4)\mathrm{d}z_2\mathrm{d}z_3\mathrm{d}z_4 &\le\frac{1}{v^3\sqrt{2\pi}}\mathrm{E}|u^\top z_{-1}|\\ &=\frac{1}{v^3\pi}\left(\frac{1+r^2}{2}\right)^{1/2}\\ &\le C,\end{aligned}$$

where the last inequality holds as $|r| \leq 1 - \varepsilon_r$ and v^3 is bounded below by Assumption 1. ∎

Lemma A.18 *Under Assumptions 1 and 2, $|\partial^2 G(r; \Delta_j, \Delta_k)/\partial\Delta_k\partial\Delta_j|$ is bounded above by some constant $C > 0$.*

Proof By the Leibniz rule,

$$\begin{aligned}\frac{\partial^2}{\partial\Delta_k\partial\Delta_j}\Phi_4(-\Delta_j, -\Delta_k, 0, 0; \Sigma_4) &= \int_{-\infty}^{0}\int_{-\infty}^{0} \phi(z_3, z_4 \mid -\Delta_j, -\Delta_k)\phi(-\Delta_j, -\Delta_k)\mathrm{d}z_3\mathrm{d}z_4 \\ &= \phi(-\Delta_j, -\Delta_k)\int_{-\infty}^{0}\int_{-\infty}^{0} \phi(z_3, z_4 \mid -\Delta_j, -\Delta_k)\mathrm{d}z_3\mathrm{d}z_4,\end{aligned}$$

where $\phi(z_3, z_4 \mid -\Delta_j, -\Delta_k)$ is the conditional pdf given $Z_1 = -\Delta_j$ and $Z_2 = -\Delta_k$. Thus, the two-dimensional integral above corresponds to a probability (and is bounded by one), leading to

$$\left|\frac{\partial^2}{\partial\Delta_k\partial\Delta_j}\Phi_4(-\Delta_j, -\Delta_k, 0, 0; \Sigma_4)\right| \leq \phi(-\Delta_j, -\Delta_k) \leq \phi(0, 0) = |\Sigma_2|^{-1/2} \leq |\Sigma_4|^{-1/2},$$

where $\Sigma_2 = \mathrm{var}\{(Z_1, Z_2)\}$. As Lemma A.20 provides that $|\Sigma_4|^{-1/2} \leq C$ for some constant $C > 0$, this concludes the proof. ∎

Lemma A.19 *Under Assumptions 1 and 2, $|\partial^2 G(r; \Delta_j, \Delta_k)/\partial r^2|$ is bounded above by some constant $C > 0$.*

Proof We start from the partial derivative with respect to r given in Theorem 6 of Yoon et al. (2020) as

$$\begin{aligned}&\frac{\partial G(r, \Delta_j, \Delta_k)}{\partial r} = -2\frac{\partial\Phi_4\{\Delta_j, \Delta_k, 0, 0; \Sigma_{4a}(r)\}}{\partial r} + 2\frac{\partial\Phi_4\{\Delta_j, \Delta_k, 0, 0; \Sigma_{4b}(r)\}}{\partial r} \\ &= \ 2^{1/2}h_{14a}(r) + 2^{1/2}h_{23a}(r) + 2h_{23a}(r) + 2h_{12b}(r) + 2^{1/2}h_{14b}(r) + 2^{1/2}h_{23b}(r) + 2h_{34b}(r),\end{aligned}$$

where $h_{14a}(r)$ is defined as

$$h_{14a}(r) = \frac{\partial\Phi(a_1, \ldots, a_4; \Sigma_{4a})}{\partial\rho_{14}(r)} = \int_{-\infty}^{a_3}\int_{-\infty}^{a_2} \phi(a_1, y_2, y_3, a_4; \Sigma_4)\mathrm{d}y_2\mathrm{d}y_3$$

and the rest of $h_{ij}(r)$'s are analogously defined.

As $\partial G(r; \Delta_j, \Delta_k)/\partial r$ is a sum of $h_{ij}(r)$'s, we show that $|\partial h_{ij}(r)/\partial r|$ is bounded above for all i and j whether $\Sigma_4 = \Sigma_{4a}$ and $\Sigma_4 = \Sigma_{4b}$. Using the multivariate chain rule and triangle inequality,

$$\left|\frac{\partial h_{ij}(r)}{\partial r}\right| = \left|\sum_{k<\ell}\frac{\partial h_{ij}(r)}{\partial\rho_{k\ell}}\frac{\partial\rho_{k\ell}}{\partial r}\right| \leq \sum_{k<\ell}\left|\frac{\partial h_{ij}(r)}{\partial\rho_{k\ell}}\right|\left|\frac{\partial\rho_{k\ell}}{\partial r}\right|.$$

By Lemma A.26, for all $1 \leq i < j \leq 4$ and $1 \leq k < \ell \leq 4$, $|\partial h_{ij}(r)/\partial\rho_{k\ell}| \leq C$ for some constant $C > 0$. Also, as $\rho_{k\ell}$'s are linear in r, $|\partial\rho_{k\ell}/\partial r|$'s are bounded above some positive constant. This concludes the proof. ∎

C.7 Auxiliary lemmas

From Theorem 4 in Yoon et al. (2020), the bridge function for TT case takes the following form

$$G(r, \Delta_j, \Delta_k) = -2\Phi_4(-\Delta_j, -\Delta_k, 0, 0, ; \Sigma_{4a}) + 2\Phi_4(-\Delta_j, -\Delta_k, 0, 0, ; \Sigma_{4b})$$

with $\Delta_j = f_j(D_j)$, $\Delta_k = f_k(D_k)$,

$$\Sigma_{4a} = \begin{pmatrix} 1 & 0 & 1/\sqrt{2} & -r/\sqrt{2} \\ 0 & 1 & -r/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -r/\sqrt{2} & 1 & -r \\ -r/\sqrt{2} & 1/\sqrt{2} & -r & 1 \end{pmatrix}, \quad \Sigma_{4b} = \begin{pmatrix} 1 & r & 1/\sqrt{2} & r/\sqrt{2} \\ r & 1 & r/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & r/\sqrt{2} & 1 & r \\ r/\sqrt{2} & 1/\sqrt{2} & r & 1 \end{pmatrix}. \tag{A34}$$

Lemma A.20 *Let $\Sigma_4 = \Sigma_{4a}$ or $\Sigma_4 = \Sigma_{4b}$ from above, and let its inverse be $\Sigma_4^{-1} = [\omega_{\ell\ell'}]_{1\le\ell,\ell'\le 4}$. Under Assumption 1, $|\omega_{\ell\ell'}| \le C_1$, $1 \le \ell, \ell' \le 4$, for some constant $C_1 > 0$. Also, $|\Sigma_4|^{-1} \le C_2$ for some constant $C_2 > 0$ regardless of $\Sigma_4 = \Sigma_{4a}$ or $\Sigma_4 = \Sigma_{4b}$.*

Proof Computing determinants gives $|\Sigma_{4a}| = |\Sigma_{4b}| = (1-r^2)^2/4$, and thus by Assumption 1, $|\Sigma_4| \ge \{1 - (1-\varepsilon_r^2)^2\}/4$, whether $\Sigma_4 = \Sigma_{4a}$ or $\Sigma_4 = \Sigma_{4b}$. Also, the inverses of Σ_{4a} and Σ_{4b} are

$$\Sigma_{4a}^{-1} = \frac{1}{r^2-1}\begin{pmatrix} -2 & 2r & \sqrt{2} & -\sqrt{2}r \\ 2r & -2 & -\sqrt{2}r & \sqrt{2} \\ \sqrt{2} & -\sqrt{2}r & -2 & 0 \\ -\sqrt{2}r & \sqrt{2} & 0 & -2 \end{pmatrix}, \quad \Sigma_{4b}^{-1} = \frac{1}{r^2-1}\begin{pmatrix} -2 & 2r & \sqrt{2} & -\sqrt{2}r \\ 2r & -2 & -\sqrt{2}r & \sqrt{2} \\ \sqrt{2} & -\sqrt{2}r & -2 & 2r \\ -\sqrt{2}r & \sqrt{2} & 2r & -2 \end{pmatrix},$$

respectively. Under Assumption 1, $|\omega_{\ell\ell'}|$'s are all bounded above by $2/\{1-(1-\varepsilon_r)^2\}$. ■

Lemma A.21 *Let $\Delta = \Phi^{-1}(\pi)$. Then, under Assumption 2,*

$$\left|\frac{\partial\Delta}{\partial\pi}\right| \le C_1 \quad \textit{and} \quad \left|\frac{\partial^2\Delta}{\partial\pi^2}\right| \le C_2$$

for some constants $C_1, C_2 > 0$.

Proof Since $|\Delta| \le M$ and $\phi(|x|)$ is a decreasing function,

$$\frac{\partial\Delta}{\partial\pi} = \left(\frac{\partial\pi}{\partial\Delta}\right)^{-1} = \left\{\frac{\partial\Phi(\Delta)}{\partial\Delta}\right\}^{-1} = \frac{1}{\phi(\Delta)} \le \frac{1}{\phi(M)}.$$

Furthermore, as the second derivative is

$$\frac{\partial^2\Delta}{\partial\pi^2} = \frac{\partial}{\partial\pi}\frac{1}{\phi(\Delta)} = \frac{\partial\Delta}{\partial\pi}\frac{\partial}{\partial\Delta}\frac{1}{\phi(\Delta)} = -\frac{1}{\{\phi(\Delta)\}^3}\frac{\partial\phi(\Delta)}{\partial\Delta} = \frac{\Delta}{\{\phi(\Delta)\}^2},$$

we have

$$\left|\frac{\partial^2 \Delta}{\partial \pi^2}\right| \le \frac{M}{\{\phi(M)\}^2}.$$

■

Lemma A.22 *Let* $(Z_1, Z_2, Z_3, Z_4)^\top \sim \mathrm{N}_4(0, \Sigma_4)$*. Then, it follows that regardless of* $\Sigma_4 = \Sigma_{4a}$ *or* $\Sigma_4 = \Sigma_{4b}$*, the conditional distribution of* Z_1 *given* Z_2, Z_3, Z_4 *is* $\mathrm{N}(\mu, v^2)$*, where*

$$\begin{aligned} \mu &:= \mathrm{E}(Z_1 \mid Z_2, Z_3, Z_4) = rZ_2 + Z_3/2^{1/2} - rZ_4/2^{1/2} \\ v^2 &:= \mathrm{var}(Z_1 \mid Z_2, Z_3, Z_4) = (1 - r^2)/2. \end{aligned}$$

Proof The results follow from the properties of conditional Gaussian distribution using the form of Σ_{4a} and Σ_{4b} (A34). ■

Lemma A.23 *Let* $Y \sim \mathrm{N}_4(0, \Sigma_4)$*, where* $\Sigma_4 = \Sigma_{4a}$ *or* $\Sigma_4 = \Sigma_{4b}$*. Then, under Assumptions 1 and 2, for any* $1 \le k < \ell \le 4$ *and* $1 \le i \le 4$*,*

$$\mathrm{E}\left(|Y_i| \mid Y_k = y_k, Y_\ell = y_\ell\right) \le C_0, \quad 0 < C_1 \le \mathrm{var}\left(Y_i \mid Y_k = y_k, Y_\ell = y_\ell\right) \le C_2$$

for some $C_0, C_1, C_2 > 0$*, where*

$$y_m = \begin{cases} -\Delta_j, & \textit{if} \quad m = 1; \\ -\Delta_k, & \textit{if} \quad m = 2; \\ 0, & \textit{otherwise}. \end{cases}$$

Proof We first calculate the conditional means and covariance matrices using the properties of multivariate Gaussian distribution to obtain:

$$\mathrm{E}(Y_1, Y_2 \mid Y_3 = 0, Y_4 = 0; \Sigma_{4a}) = \mathrm{E}(Y_1, Y_2 \mid Y_3 = 0, Y_4 = 0; \Sigma_{4b}) = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

$$\mathrm{E}(Y_1, Y_3 \mid Y_2 = -\Delta_k, Y_4 = 0; \Sigma_{4a}) = \mathrm{E}(Y_1, Y_3 \mid Y_2 = -\Delta_k, Y_4 = 0; \Sigma_{4a}) = \begin{pmatrix} -\Delta_k r \\ 0 \end{pmatrix},$$

$$\mathrm{E}(Y_1, Y_4 \mid Y_2 = -\Delta_k, Y_3 = 0; \Sigma_{4a}) = \mathrm{E}(Y_1, Y_4 \mid Y_2 = -\Delta_k, Y_3 = 0; \Sigma_{4b}) = \frac{1}{2 - r^2} \begin{pmatrix} -\Delta_k r \\ -\sqrt{2}\Delta_k(1 - r^2) \end{pmatrix},$$

$$\mathrm{E}(Y_2, Y_3 \mid Y_1 = -\Delta_j, Y_4 = 0; \Sigma_{4a}) = \mathrm{E}(Y_2, Y_3 \mid Y_1 = -\Delta_j, Y_4 = 0; \Sigma_{4a}) = \frac{1}{2 - r^2} \begin{pmatrix} -\Delta_j r \\ -\sqrt{2}\Delta_j(1 - r^2) \end{pmatrix},$$

$$\mathrm{E}(Y_2, Y_4 \mid Y_1 = -\Delta_j, Y_3 = 0; \Sigma_{4a}) = \mathrm{E}(Y_2, Y_4 \mid Y_1 = -\Delta_j, Y_3 = 0; \Sigma_{4b}) = \begin{pmatrix} -\Delta_j r \\ 0 \end{pmatrix}$$

$$\mathrm{E}(Y_3, Y_4 \mid Y_1 = -\Delta_j, Y_2 = -\Delta_k; \Sigma_{4a}) = \frac{1}{\sqrt{2}} \begin{pmatrix} \Delta_k r - \Delta_j \\ \Delta_j r - \Delta_k \end{pmatrix},$$

$$\mathrm{E}(Y_3, Y_4 \mid Y_1 = -\Delta_j, Y_2 = -\Delta_k; \Sigma_{4b}) = \frac{1}{\sqrt{2}} \begin{pmatrix} -\Delta_j \\ -\Delta_k \end{pmatrix},$$

and

$$
\begin{aligned}
&\operatorname{var}(Y_1, Y_2|Y_3, Y_4; \Sigma_{4a}) = \operatorname{var}(Y_1, Y_2|Y_3, Y_4; \Sigma_{4b}) = \frac{1}{2}\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}, \\
&\operatorname{var}(Y_1, Y_3|Y_2, Y_4; \Sigma_{4a}) = \operatorname{var}(Y_1, Y_3|Y_2, Y_4; \Sigma_{4b}) \\
&= \operatorname{var}(Y_2, Y_4|Y_1, Y_3; \Sigma_{4a}) = \operatorname{var}(Y_2, Y_4|Y_1, Y_3; \Sigma_{4b}) = (1 - r^2)\begin{pmatrix} 1 & 1/\sqrt{2} \\ 1/\sqrt{2} & 1 \end{pmatrix}, \\
&\operatorname{var}(Y_1, Y_4 \mid Y_2, Y_3; \Sigma_{4a}) = \operatorname{var}(Y_2, Y_3 \mid Y_1, Y_4; \Sigma_{4a}) \\
&= \operatorname{var}(Y_1, Y_4 \mid Y_2, Y_3; \Sigma_{4b}) = \operatorname{var}(Y_2, Y_3 \mid Y_1, Y_4; \Sigma_{4b}) = \left(1 - \frac{1}{2 - r^2}\right)\begin{pmatrix} 1 & -r/\sqrt{2} \\ -r/\sqrt{2} & 1 \end{pmatrix}, \\
&\operatorname{var}(Y_3, Y_4 \mid Y_1, Y_2; \Sigma_{4a}) = \frac{2}{1 - r^2}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \\
&\operatorname{var}(Y_3, Y_4 \mid Y_1, Y_2; \Sigma_{4b}) = \frac{1}{2}\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}.
\end{aligned}
$$

Consider $\operatorname{var}(Y_i \mid Y_k = y_k, Y_\ell = y_\ell)$. From the above, it is clear that all conditional variances are bounded below by some positive constant. It can be also seen that all conditional variances are bounded above as long as $1 - r^2 \geq C$ for some constant $C > 0$. Under Assumptions 1, $|r| \leq 1 - \varepsilon_r$ and thus $1 - r^2 \geq 1 - (1 - \varepsilon_r)^2 > 0$.

Consider $\mathrm{E}\left(|Y_i| \mid Y_k = y_k, Y_\ell = y_\ell\right)$. If $i = j$ or $i = \ell$, then $\mathrm{E}\left(|Y_i| \mid Y_k = y_k, Y_\ell = y_\ell\right) = |y_i|$ and the result is immediate under Assumption 2. For $i \neq j, k$, let $\mathrm{E}(Y_i \mid Y_k = y_k, Y_\ell = y_\ell) = \mu_i$ and $\operatorname{var}(Y_i \mid Y_k = y_k, Y_\ell = y_\ell) = \sigma_i^2$, where detailed expressions are given above. Then, by Lemma A.27, we have that

$$
\begin{aligned}
\mathrm{E}\left(|Y_i| \mid Y_k = y_k, Y_\ell = y_\ell\right) &= \left[\sigma_i\sqrt{\frac{2}{\pi}}\exp\left(-\frac{\mu_i^2}{2\sigma_i^2}\right) + \mu_i\left\{1 - 2\Phi\left(-\frac{\mu_i}{\sigma_i}\right)\right\}\right] \\
&\leq \sigma_i\sqrt{\frac{2}{\pi}} + |\mu_i|.
\end{aligned}
$$

We can see from the above conditional means that, under Assumptions 1 and 2, $|\mu_i|$ is bounded above by some positive constant. As we already showed that σ_i^2 is bounded above, the proof is complete. ■

Lemma A.24 *Let $Y \sim \mathrm{N}_4(0, \Sigma_4)$, where $\Sigma_4 = \Sigma_{4a}$ or $\Sigma_4 = \Sigma_{4b}$. Also let Y_{-i} be the 3-dimensional random vector without the ith component and $y_{-i} = (y_j, y_k, y_\ell)^\top$ be its realization such that*

$$
y_m = \begin{cases} -\Delta_j, & \textit{if} \quad m = 1; \\ -\Delta_k, & \textit{if} \quad m = 2; \\ 0, & \textit{otherwise}. \end{cases}
$$

Then, under Assumptions 1 and 2, for any $1 \leq i \leq 4$,

$$
\mathrm{E}\left(|Y_i| \mid Y_{-i} = y_{-i}\right) \leq C
$$

for some constant $C > 0$.

Proof It follows by the conditional mean and variance formulas of the multivariate Gaussian distribution that, regardless of $\Sigma_4 = \Sigma_{4a}$ or $\Sigma_4 = \Sigma_{4b}$,

$$\begin{aligned} \operatorname{var}(Y_i \mid Y_{-i} = y_{-i}; \Sigma_4) &= \frac{(1-r^2)}{2}, \quad i = 1, \ldots, 4, \\ \mathrm{E}(Y_1 \mid Y_{-1} = y_{-1}; \Sigma_4) &= -\Delta_k r, \\ \mathrm{E}(Y_2 \mid Y_{-2} = y_{-2}; \Sigma_4) &= -\Delta_j r, \\ \mathrm{E}(Y_3 \mid Y_{-3} = y_{-3}; \Sigma_4) &= -\frac{\Delta_j - \Delta_k r}{\sqrt{2}}, \\ \mathrm{E}(Y_4 \mid Y_{-4} = y_{-4}; \Sigma_4) &= -\frac{\Delta_k - \Delta_j r}{\sqrt{2}}. \end{aligned}$$

Under Assumptions 1 and 2, the absolute values of the conditional means and conditional variances are bounded above by $\sqrt{2}M$ and $1/2$, respectively. Then the result follows by Lemma A.27. ■

Lemma A.25 *Let $y \sim \mathrm{N}_4(\mathbf{0}, \Sigma_4)$, where $\Sigma_4 = \Sigma_{4a}$ or $\Sigma_4 = \Sigma_{4b}$. Then, for any $1 \leq k < \ell \leq 4$ and $1 \leq i < j \leq 4$,*

$$\mathrm{E}\left\{|Y_i Y_j| \;\middle|\; Y_k = y_k, Y_\ell = y_\ell\right\} \leq C$$

for some constant $C > 0$.

Proof Let $I = \{k, \ell\}$ and write $y_I = (y_k, y_\ell)^\top = (y_{I_1}, y_{I_2})^\top$. We prove this lemma by considering the following three cases: $\operatorname{card}(\{i, j\} \cap I) = 0, 1, 2$, namely cases 1, 2, and 3, respectively.

Case 1: Consider the case $\operatorname{card}(\{i, j\} \cap I) = 0$. Let

$$\begin{aligned} \mathrm{E}(Y_i \mid Y_I = y_I) = \mu_i, \quad \operatorname{var}(Y_i \mid Y_I = y_I) = \sigma_i^2, \\ \mathrm{E}(Y_j \mid Y_I = y_I) = \mu_j, \quad \operatorname{var}(Y_j \mid Y_I = y_I) = \sigma_j^2, \end{aligned}$$

whose detailed expressions are provided in Lemma A.23. Also, let $Z_i = Y_i/(2^{1/2}\sigma_i) \sim \mathrm{N}(\mu_i/(2^{1/2}\sigma_i), 1/2)$ and $Z_j = Y_j/(2^{1/2}\sigma_j) \sim \mathrm{N}(\mu_j/(2^{1/2}\sigma_j), 1/2)$ and write

$$\begin{aligned} |Y_i Y_j| = 2\sigma_i\sigma_j \left| \frac{Y_i}{\sqrt{2}\sigma_i} \frac{Y_j}{2^{1/2}\sigma_j} \right| &= 2\sigma_i\sigma_j \left| \frac{1}{4}(Z_i + Z_j)^2 - \frac{1}{4}(Z_i - Z_j)^2 \right| \\ &\leq 2\sigma_i\sigma_j \left\{ \frac{1}{4}(Z_i + Z_j)^2 + \frac{1}{4}(Z_i - Z_j)^2 \right\}, \end{aligned}$$

where the last inequality holds by the triangle inequality. We have that $(Z_i + Z_j)^2$ and $(Z_i - Z_j)^2$ follow non-central $\chi^2_{\mathrm{df}=1}$ distributions with non-centrality parameters $\lambda_+ = \mu_i^2/(2\sigma_i^2) + \mu_j^2/(2\sigma_j^2)$ and $\lambda_- = \mu_i^2/(2\sigma_i^2) - \mu_j^2/(2\sigma_j^2)$, and thus,

$$\mathrm{E}\left\{|Y_i Y_j| \mid Y_{I_1} = y_{I_1}, Y_{I_2} = y_{I_2}\right\} \leq \frac{\sigma_i\sigma_j}{2} \left\{\lambda_+ + \lambda_- + 2\right\}.$$

By Lemma A.23, we have $\mathrm{E}\{|Y_iY_j| \mid Y_{I_1} = y_{I_1}, Y_{I_2} = y_{I_2}\} < C$ for some constant $C > 0$.

Case 2: For the case $\mathrm{card}(\{i,j\} \cap I) = 1$, we assume, without loss of generality that, $\{i,j\} \cap I = \{i\}$ and write

$$\begin{aligned}\mathrm{E}\{|Y_iY_j| \mid Y_{I_1} = y_{I_1}, Y_{I_2} = y_{I_2}\} &= |y_{I_1}|\ \mathrm{E}\{|Y_j| \mid Y_{I_1} = y_{I_1}, Y_{I_2} = y_{I_2}\}\\ &\leq M\ \mathrm{E}\{|Y_j| \mid Y_{I_1} = y_{I_1}, Y_{I_2} = y_{I_2}\},\end{aligned}$$

where $|y_k| \leq M$ by Assumption 2. Then, by Lemma A.23, we have

$$\mathrm{E}\{|Y_iY_j| \mid Y_{I_1} = y_{I_1}, Y_{I_2} = y_{I_2}\} \leq C$$

for some constant $C > 0$.

Case 3: For the case $\{i,j\} \cap I = \{i,j\}$, Assumption 2 gives that

$$\mathrm{E}\{|Y_iY_j| \mid Y_{I_1} = y_{I_1}, Y_{I_2} = y_{I_2}\} = |y_{I_1}y_{I_2}| \leq M^2.$$

This concludes the proof. ■

Lemma A.26 *Let $h_{ij}(r) = \partial\Phi(a_1, \ldots, a_4; \Sigma_4)/\partial\rho_{ij}(r)$, where $\Sigma_4 = [\rho_{ij}(r)]_{1 \leq i,j \leq 4}$. Then, for any $1 \leq i < j \leq 4$ and $1 \leq k < \ell \leq 4$, and some constant $C > 0$,*

$$\left|\frac{\partial h_{ij}(r)}{\partial \rho_{k\ell}}\right| \leq C.$$

Proof Let $I = \{i,j\}$, $K = \{k,\ell\}$. We write $x_I = (x_i, x_j)^\top = (x_{I_1}, x_{I_2})^\top$ and $\mathcal{R}_I = \{(x_i, x_j) \mid x_i < a_i, x_j < a_j\} \subset \mathbb{R}^2$, where.

$$a_m = \begin{cases} -\Delta_j, & \text{if} \quad m = 1;\\ -\Delta_k, & \text{if} \quad m = 2;\\ 0, & \text{otherwise.}\end{cases}$$

We consider three cases where $\mathrm{card}(I \cap K) = 2, 1, 0$.

Consider the case $\mathrm{card}(I \cap K) = 0$, i.e., $K = \{1, \ldots, 4\} - I = I^c$. By Plackett (1954)

$$\frac{\partial h_I(r)}{\partial \rho_{I^c}} = \frac{\partial}{\partial \rho_{I^c}} \int_{\mathcal{R}_{I^c}} \phi(a_I, x_{I^c}; \Sigma_4) dx_{I^c} = \phi(a_I, a_{I^c}; \Sigma_4) \leq |\Sigma_4|^{1/2}$$

because $\phi(a_I, a_{I^c}; \Sigma_4) \leq \phi(0,0,0,0; \Sigma_4) = |\Sigma_4|^{-1/2}$. By Lemma A.20, we have

$$\left|\frac{\partial h_I(r)}{\partial \rho_{I^c}}\right| \leq C$$

for some constant $C > 0$.

Consider the case $\mathrm{card}(I \cap K) = 2$, i.e., $I = K$. For notational convenience, we write $a_I = y_I$ and $x_{I^c} = y_{I^c}$. Then,

$$\begin{aligned}\frac{\partial h_I(r)}{\partial \rho_I} &= \frac{\partial}{\partial \rho_I} \int_{\mathcal{R}_{I^c}} \phi(a_I, x_{I^c}; \Sigma_4) \mathrm{d}x_{I^c} = \frac{\partial}{\partial \rho_I} \int_{\mathcal{R}_{I^c}} \phi(y_I, y_{I^c}; \Sigma_4) \mathrm{d}y_{I^c}\\ &= \int_{\mathcal{R}_{I^c}} \frac{\partial}{\partial \rho_I} \phi(y_I, y_{I^c}; \Sigma_4) \mathrm{d}y_{I^c} = \int_{\mathcal{R}_{I^c}} \frac{\partial^2}{\partial y_{I_1} \partial y_{I_2}} \phi(y_I, y_{I^c}; \Sigma_4) \mathrm{d}y_{I^c},\end{aligned}$$

where the last equality is due to Plackett (1954). Let $\omega_j^\top$ be the jth row of $\Sigma_4^{-1} = [\omega_{ij}]_{1\le i,j\le 4}$, $\Sigma_I = \text{var}(y_I)$, and $\Sigma_{I^c|I} = \text{var}(y_{I^c}|y_I)$. By differentiating the multivariate Gaussian density, we have

$$\begin{aligned}\left|\int_{\mathcal{R}_{I^c}} \frac{\partial^2}{\partial y_{I_1}\partial y_{I_2}}\phi(y_I, y_{I^c};\Sigma_4)dy_{I^c}\right| &= \left|\int_{\mathcal{R}_{I^c}} \left\{(\omega_{I_1}^\top y)(\omega_{I_2}^\top y) - \omega_I\right\}\phi(y_I, y_{I^c};\Sigma_4)dy_{I^c}\right| \\ &= \left|\phi(y_I;\Sigma_I)\int_{\mathcal{R}_{I^c}} \left\{(\omega_{I_1}^\top y)(\omega_{I_2}^\top y) - \omega_I\right\}\phi(y_{I^c}|y_I;\Sigma_{I^c|I})dy_{I^c}\right|.\end{aligned}$$

We also have that $\phi(y_I;\Sigma_I) \le \phi(0,0;\Sigma_I) = |\Sigma_I|^{-1/2} \le |\Sigma_4|^{-1/2}$ for any I, and by Lemma A.20, $|\Sigma_4|^{-1/2}| \le C_2^{1/2}$ for some constant $C_2 > 0$. Thus,

$$\left|\int_{\mathcal{R}_{I^c}} \frac{\partial^2}{\partial y_{I_1}\partial y_{I_2}}\phi(y_I, y_{I^c};\Sigma_4)dy_{I^c}\right| \le C_2^{1/2}\left|\int_{\mathcal{R}_{I^c}} \left\{(\omega_{I_1}^\top y)(\omega_{I_2}^\top y) - \omega_I\right\}\phi(y_{I^c}|y_I;\Sigma_{I^c|I})dy_{I^c}\right|.$$

The absolute value of the integral of the last term is bounded as

$$\begin{aligned}\left|\int_{\mathcal{R}_{I^c}} \left\{(\omega_{I_1}^\top y)(\omega_{I_2}^\top y) - \omega_I\right\}\phi(y_{I^c}|y_I;\Sigma_{I^c|I})dy_{I^c}\right| &\le \int_{\mathcal{R}_{I^c}} \left|(\omega_{I_1}^\top y)(\omega_{I_2}^\top y) - \omega_I\right|\phi(y_{I^c}|y_I;\Sigma_{I^c|I})dy_{I^c} \\ &\le \int_{\mathbb{R}^2} \left|(\omega_{I_1}^\top y)(\omega_{I_2}^\top y) - \omega_I\right|\phi(y_{I^c}|y_I;\Sigma_{I^c|I})dy_{I^c} \\ &\le \int_{\mathbb{R}^2} \left|(\omega_{I_1}^\top y)(\omega_{I_2}^\top y)\right|\phi(y_{I^c}|y_I;\Sigma_{I^c|I})dy_{I^c} + |\omega_I|,\end{aligned}$$

where the second inequality is due to expanding the range of integration, and the third inequality is due to the triangle inequality. By Lemma A.20, we know that, whether $\Sigma_4 = \Sigma_{4a}$ or $\Sigma_4 = \Sigma_{4b}$, $|\omega_{jk}| \le C_1$, for all $1 \le j,k \le 4$. Also, by the triangle inequality,

$$\left|(\omega_{I_1}^\top y)(\omega_{I_2}^\top y)\right| = \left|\sum_{i'=1}^4\sum_{j'=1}^4 \omega_{I_1 i'}\omega_{I_2 j'}y_{i'}y_{j'}\right| \le C_1^2\sum_{i'=1}^4\sum_{j'=1}^4 \left|y_{i'}y_{j'}\right|.$$

Hence, for some constant $C > 0$,

$$\begin{aligned}\left|\frac{\partial h_I(r)}{\partial \rho_I}\right| &\le C_2^{1/2}\int_{\mathbb{R}^2}\left|(\omega_{I_1}^\top y)(\omega_{I_2}^\top y)\right|\phi(y_{I^c}|y_I;\Sigma_{I^c|I})dy_{I^c} + C_2^{1/2}C_1 \\ &\le C_2^{1/2}C_1^2\sum_{i'=1}^4\sum_{j'=1}^4\int_{\mathbb{R}^2}\left|y_{i'}y_{j'}\right|\phi(y_{I^c}|y_I;\Sigma_{I^c|I})dy_{I^c} + C_2^{1/2}C_1 \\ &\le C,\end{aligned}$$

where the last inequality holds by Lemma A.25.

Consider the case card$(I\cap K) = 1$. We assume, without loss of generality, that $I = \{i,j\}$ and $K = \{j,\ell\}$, i.e., $I\cap K = \{j\}$. Then, by Plackett (1954),

$$\frac{\partial h_{ij}(r)}{\partial \rho_{j\ell}} = \frac{\partial}{\partial \rho_{j\ell}}\int_{-\infty}^{a_\ell}\int_{-\infty}^{a_k}\phi(a_i,a_j,x_k,x_\ell;\Sigma_4)\mathrm{d}x_k\mathrm{d}x_\ell = \int_{-\infty}^{a_k}\frac{\partial}{\partial a_j}\phi(a_i,a_j,x_k,a_\ell;\Sigma_4)\mathrm{d}x_k.$$

For notational convenience, let $y = (a_i, a_j, x_k, a_\ell)^\top = (y_i, y_j, y_k, y_\ell)^\top$ and write

$$\frac{\partial h_{ij}(r)}{\partial \rho_{j\ell}} = \int_{-\infty}^{a_k} \frac{\partial}{\partial y_j} \phi(y; \Sigma_4) \mathrm{d}y_k = \int_{-\infty}^{a_k} (-\omega_j^\top y) \phi(y; \Sigma_4) \mathrm{d}y_k.$$

Then, we have that

$$\begin{aligned}
\left| \frac{\partial h_{ij}(r)}{\partial \rho_{j\ell}} \right| &= \left| \int_{-\infty}^{a_k} (-\omega_j^\top y) \phi(y; \Sigma_4) \mathrm{d}y_k \right| \\
&\le \int_{-\infty}^{a_k} |\omega_j^\top y| \phi(y; \Sigma_4) \mathrm{d}y_k \\
&\le \int_{-\infty}^{\infty} |\omega_j^\top y| \phi(y; \Sigma_4) \mathrm{d}y_k \quad \text{(by expanding the range of integration)} \\
&\le \sum_{i' \ne k} |\omega_{ji'} y_{i'}| \phi(y_i, y_j, y_\ell; \Sigma_3) + \phi(y_i, y_j, y_\ell; \Sigma_3) \int_{-\infty}^{\infty} |\omega_{jk} y_k| \phi(y_k | y_i, y_j, y_\ell; \Sigma_4) \mathrm{d}y_k,
\end{aligned}$$

where the last inequality holds by the triangle inequality. By Lemma A.20, $|\omega_{jk}| \le C_1$ for all $1 \le j, k \le 4$, and by Assumption 2, $|y_i| \le M$ for all for $1 \le i \le 4$. This gives

$$\begin{aligned}
\left| \frac{\partial h_{ij}(r)}{\partial \rho_{j\ell}} \right| &\le \sum_{i' \ne k} |\omega_{ji'} y_{i'}| \phi(y_i, y_j, y_\ell; \Sigma_3) + \int_{-\infty}^{\infty} |\omega_{jk} y_k| \phi(y_k | y_i, y_j, y_\ell; \Sigma_4) \mathrm{d}y_k \left\{ \phi(y_i, y_j, y_\ell; \Sigma_3) \right\} \\
&\le 3 C_1 M |\Sigma_4|^{-1/2} + C_1 |\Sigma_4|^{-1/2} \int_{-\infty}^{\infty} |y_k| \phi(y_k | y_i, y_j, y_\ell; \Sigma_4) \mathrm{d}y_k.
\end{aligned}$$

Again, by Lemma A.20, $|\Sigma_4|^{-1/2} \le C_2^{1/2}$, and by Lemma A.24, above integral is bounded above by some positive constant. Thus, for some constant $C > 0$

$$\left| \frac{\partial h_{ij}(r)}{\partial \rho_{j\ell}} \right| \le C.$$

■

Lemma A.27 *Let $X \sim \mathrm{N}(\mu, \sigma^2)$. Then* $\mathrm{E}(|X|) \le \sigma(2/\pi)^{1/2} + |\mu|$.

Proof For $X \sim \mathrm{N}(\mu, \sigma^2)$, $|X|$ follows the folded Gaussian distribution with mean

$$\mathrm{E}(|X|) = \sigma \left(\frac{2}{\pi} \right)^{1/2} \exp\left(-\frac{\mu^2}{2\sigma^2} \right) + \mu \left\{ 1 - 2\Phi\left(-\frac{\mu}{\sigma} \right) \right\}.$$

Since $\exp(-a^2) \le 1$ and $0 \le a\{1 - 2\Phi(a)\} \le |a|$, $a \in \mathbb{R}$, we have that

$$\mathrm{E}(|X|) \le \sigma \left(\frac{2}{\pi} \right)^{1/2} + |\mu|.$$

■

References

Constantin Ahlmann-Eltze and Wolfgang Huber. Comparison of transformations for single-cell rna-seq data. *Nature Methods*, pages 1–8, 2023.

Jeongyoun Ahn and JS Marron. The maximal data piling direction for discrimination. *Biometrika*, 97(1):254–259, 2010.

Jeongyoun Ahn, Hee Cheol Chung, and Yongho Jeon. Trace ratio optimization for high-dimensional multi-class discrimination. *Journal of Computational and Graphical Statistics*, 30(1):192–203, 2021.

John Aitchison. Principal component analysis of compositional data. *Biometrika*, 70(1): 57–65, 1983.

James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679, 1993.

Rina Foygel Barber and Mladen Kolar. Rocket: Robust confidence intervals via kendall's tau for transelliptical graphical models. *The Annals of Statistics*, 46(6B):3422–3450, 2018.

Peter J Bickel and Elizaveta Levina. Some theory for fisher's linear discriminant function, naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010, 2004.

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

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

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

Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

Siddhartha Chib and Edward Greenberg. Analysis of multivariate probit models. *Biometrika*, 85(2):347–361, 1998.

Yasuko Chikuse. *Statistics on Special Manifolds*, volume 174. Springer New York, NY, 2003.

Hee Cheol Chung, Irina Gaynanova, and Yang Ni. Phylogenetically informed bayesian truncated copula graphical models for microbial association networks. *The Annals of Applied Statistics*, 16(4):2437–2457, 2022.

Line Clemmensen, Trevor Hastie, Daniela Witten, and Bjarne Ersbøll. Sparse discriminant analysis. *Technometrics*, 53(4):406–413, 2011.

David Roxbee Cox. *Analysis of Binary Data*. Routledge, 2018.

Victor De la Pena and Evarist Giné. *Decoupling: from dependence to independence*. Springer Science & Business Media, 2012.

Debangan Dey and Vadim Zipunnikov. Semiparametric gaussian copula regression modeling for mixed data types (sgcrm). *arXiv preprint arXiv:2205.06868*, 2022.

Kai Dong, Hongyu Zhao, Tiejun Tong, and Xiang Wan. NBLDA: negative binomial linear discriminant analysis for RNA-seq data. *BMC Bioinformatics*, 17(1):1–10, 2016.

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

Augusto Fasano and Daniele Durante. A class of conjugate priors for multinomial probit models which includes the multivariate normal one. *Journal of Machine Learning Research*, 23(30):1–26, 2022.

Huijie Feng and Yang Ning. High-dimensional mixed graphical model with ordinal data - parameter estimation and statistical inference. *AISTATS*, 2019.

Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. doi: 10.18637/jss.v033.i01. URL `https://www.jstatsoft.org/index.php/jss/article/view/v033i01`.

Irina Gaynanova. Prediction and estimation consistency of sparse multi-class penalized optimal scoring. *Bernoulli*, 26(1):286–322, 2020.

Irina Gaynanova. *MGSDA: Multi-Group Sparse Discriminant Analysis*, 2021. URL `https://CRAN.R-project.org/package=MGSDA`. R package version 1.6.

Irina Gaynanova, James G Booth, and Martin T Wells. Simultaneous sparse estimation of canonical vectors in the $p \gg N$ setting. *Journal of the American Statistical Association*, 111(514):696–706, 2016.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. CRC Press, 2013.

Dincer Goksuluk, Gokmen Zararsiz, Selcuk Korkmaz, and Ahmet Ergun Karaagaoglu. *NBLDA: Negative Binomial Linear Discriminant Analysis*, 2022. URL `https://CRAN.R-project.org/package=NBLDA`. R package version 1.0.1.

Yaqian Guo, Trevor Hastie, and Robert Tibshirani. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, 8(1):86–100, 2007.

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

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The Elements of Statistical Learning: Data mining, Inference, and Prediction*, volume 2. Springer New York, NY, 2009.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, 2015. ISBN 9781498712170. URL `https://books.google.com/books?id=f-A_CQAAQBAJ`.

Adolfo Hernández and Santiago Velilla. Dimension reduction in nonparametric kernel discriminant analysis. *Journal of Computational and Graphical Statistics*, 14(4):847–866, 2005.

C. Holmes and L. Knorr-Held. Efficient simulation of bayesian logistic regression models. Technical report, Ludwig Maximilians University, Munich, 2003.

Mingze Huang, Christian L Müller, and Irina Gaynanova. latentcor: An r package for estimating latent correlations from mixed data types. *Journal of Open Source Software*, 6(65):3634, 2021.

Paul Jaccard. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2): 37–50, 1912.

Satkartar K Kinney and David B Dunson. Fixed and random effects selection in linear and logistic models. *Biometrics*, 63(3):690–698, 2007.

Michel Lang. *mlr3measures: Performance Measures for 'mlr3'*, 2022. URL `https://CRAN.R-project.org/package=mlr3measures`. R package version 0.5.0.

Alexander F Lapanowski and Irina Gaynanova. Sparse feature selection in kernel discriminant analysis via optimal scoring. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1704–1713. PMLR, 2019.

Myung Hee Lee, Jeongyoun Ahn, and Yongho Jeon. HDLSS discrimination with adaptive data piling. *Journal of Computational and Graphical Statistics*, 22(2):433–451, 2013.

Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL `https://CRAN.R-project.org/doc/Rnews/`.

Yi Lin and Yongho Jeon. Discriminant analysis through a semiparametric model. *Biometrika*, 90(2):379–392, 2003.

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

Verónica Lloréns-Rico, Sara Vieira-Silva, Pedro J Gonçalves, Gwen Falony, and Jeroen Raes. Benchmarking microbiome transformations favors experimental quantitative approaches to address compositionality and sampling depth biases. *Nature communications*, 12(1): 3562, 2021.

Sugnet Lubbe, Peter Filzmoser, and Matthias Templ. Comparison of zero replacement strategies for compositional data with large numbers of zeros. *Chemometrics and Intelligent Laboratory Systems*, 210:104248, 2021.

Shujie Ma, Liangjun Su, and Yichong Zhang. Detecting latent communities in network formation models. *Journal of Machine Learning Research*, 23(310):1–61, 2022.

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

Qing Mai, Hui Zou, and Ming Yuan. A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, 99(1):29–42, 2012.

Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

Donald T McKnight, Roger Huerlimann, Deborah S Bower, Lin Schwarzkopf, Ross A Alford, and Kyall R Zenger. Methods for normalizing microbiome data: an ecological perspective. *Methods in Ecology and Evolution*, 10(3):389–400, 2019.

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

Sean M O'brien and David B Dunson. Bayesian multivariate logistic regression. *Biometrics*, 60(3):739–746, 2004.

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

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

EA Rakha, ME El-Sayed, AR Green, EC Paish, AHS Lee, and IO Ellis. Breast carcinoma with basal differentiation: a proposal for pathology definition based on basal cytokeratin expression. *Histopathology*, 50(4):434–438, 2007.

Justin D Silverman, Kimberly Roche, Sayan Mukherjee, and Lawrence A David. Naught all zeros in sequence count data are the same. *Computational and Structural Biotechnology Journal*, 18:2789–2798, 2020.

Doris Vandeputte, Gunter Kathagen, Kevin D'hoe, Sara Vieira-Silva, Mireia Valles-Colomer, João Sabino, Jun Wang, Raul Y Tito, Lindsey De Commer, Youssef Darzi, et al. Quantitative microbiome profiling links gut community variation to microbial load. *Nature*, 551(7681):507–511, 2017.

Sophie Weiss, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, Jesse R Zaneveld, Yoshiki Vázquez-Baeza, Amanda Birmingham, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5:1–18, 2017.

Daniela Witten. *PoiClaClu: Classification and Clustering of Sequencing Data Based on a Poisson Model*, 2019. URL `https://CRAN.R-project.org/package=PoiClaClu`. R package version 1.0.2.1.

Daniela M Witten. Classification and clustering of sequencing data using a Poisson model. *The Annals of Applied Statistics*, 5(4):2493–2518, 2011.

Daniela M Witten and Robert Tibshirani. Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):753–772, 2011.

Congrui Yi and Jian Huang. Semismooth newton coordinate descent algorithm for elastic-net penalized huber loss regression and quantile regression. *Journal of Computational and Graphical Statistics*, 26(3):547–557, 2017.

Congrui Yi and Yaohui Zeng. *sparseSVM: Solution Paths of Sparse High-Dimensional Support Vector Machine with Lasso or Elastic-Net Regularization*, 2018. URL `https://CRAN.R-project.org/package=sparseSVM`. R package version 1.1-6.

Grace Yoon and Irina Gaynanova. *Sparse Canonical Correlation Analysis for High-Dimensional Mixed Data*, 2021. R package version 1.4.6.

Grace Yoon, Irina Gaynanova, and Christian L Müller. Microbial networks in spring-semi-parametric rank-based correlation and partial correlation estimation for quantitative microbiome data. *Frontiers in Genetics*, 10:516, 2019.

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