Journal of the Indian Society for Probability and Statistics (2024) 25:17–45
https://doi.org/10.1007/s41096-024-00181-0

INVITED ARTICLE

Single-Index Mixed-Effects Model for Asymmetric Bivariate Clustered Data

Weihua Zhao[1] · **Dipankar Bandyopadhyay**[3] · **Heng Lian**[2]

Accepted: 8 February 2024 / Published online: 16 March 2024

Abstract

Studies/trials assessing status and progression of periodontal disease (PD) usually focus on quantifying the relationship between the clustered (tooth within subjects) bivariate endpoints, such as probed pocket depth (PPD), and clinical attachment level (CAL) with the covariates. Although assumptions of multivariate normality can be invoked for the random terms (random effects and errors) under a linear mixed model (LMM) framework, violations of those assumptions may lead to imprecise inference. Furthermore, the response-covariate relationship may not be linear, as assumed under a LMM fit, and the regression estimates obtained therein do not provide an overall summary of the risk of PD, as obtained from the covariates. Motivated by a PD study on Gullah-speaking African-American Type-2 diabetics, we cast the asymmetric clustered bivariate (PPD and CAL) responses into a non-linear mixed model framework, where both random terms follow the multivariate asymmetric Laplace distribution (ALD). In order to provide a one-number risk summary, the possible non-linearity in the relationship is modeled via a single-index model, powered by polynomial spline approximations for index functions, and the normal mixture expression for ALD. To proceed with a maximum-likelihood inferential setup, we devise an elegant EM-type algorithm. Moreover, the large sample theoretical properties are established under some mild conditions. Simulation studies using synthetic data generated under a variety of scenarios were used to study the finite-sample properties of our estimators, and demonstrate that our proposed model and estimation algorithm can efficiently handle asymmetric, heavy-tailed data, with outliers. Finally, we illustrate our proposed methodology via application to the motivating PD study.

Keywords Asymmetric Laplace distribution · Clustered data · EM algorithm · Random-effects · Single-index model

Extended author information available on the last page of the article

1 Introduction

Epidemiological studies in a clustered, or longitudinal data setting often generate multivariate (repeated) outcomes that are analyzed under the ubiquitous multivariate normal (MVN) assumptions of the random terms (random effects, and within-subject random errors) via standard software, such as `SAS`, or `R`. However, violations of those assumptions can lead to imprecise parameter estimates (Bandyopadhyay et al. 2010). These non-Gaussian features are usually manifested through skewness of the response vector, and/or thick-tails. Although achieving close-to-normality via suitable data transformations of the responses (such as log, or Box-Cox) for standard linear mixed model (LMM) analysis are possible, they maybe avoided due to their non-universality, and difficulty in covariate interpretation on the original scale (Jara et al. 2008). To address this, various flexible (parametric) alternatives to the MVN density exists, such as the multivariate skew-normal density (Azzalini and Capitanio 1999; Gupta et al. 2004; Azzalini 2010), the heavy-tailed multivariate skew t-density (Azzalini and Capitanio 2003), and others, that can accommodate departures from normality without resorting to ad-hoc data transformations.

In practice, this setup can be further complicated in presence of multiple outcomes recorded at each cluster units/components. The motivating data example in this paper comes from a clinical study of periodontal disease (PD) conducted on Gullah-speaking African-American Type-2 diabetics (henceforth, GAAD). Here, the multiple outcomes of interest are the *tooth-level* (mean) probed pocket depth (PPD) and clinical attachment level (CAL), which are recorded (in mm, via a periodontal probe) simultaneously for each tooth nested/clustered within a subject. While PPD quantifies the current PD status, CAL measures the (past) disease history and progression (Page and Eke 2007). An oral clinician may be interested in studying the *joint* evolution of these outcomes over some features of covariates, and the complexity is induced from two different sources of correlation—(a) Between repeated observations of any given outcome (PPD, or CAL) measured at a cluster unit (tooth), and (b) Between multiple outcomes (PPD and CAL) measured at the same tooth. The existing literature (both classical and Bayesian) in this context of multiple repeated outcomes modeling is also very rich (Luo and Wang 2014; Verbeke et al. 2014; Lin and Wang 2013; Michaelis et al. 2018; Bandyopadhyay et al. 2010). However, a vast majority of these models are developed under the *restrictive* assumption of linearity of the covariate effects over the multivariate responses.

To motivate further, consider Fig. 1, which presents plots of the empirical Bayes' estimates of random effects (panels a and b), corresponding Q-Q plots (panels c and d), and observed versus estimated (non-linear) curve (panels e and f), obtained from fitting a LMM separately to the PPD and CAL responses in the GAAD data, using the `lme` function in `R`. The plots clearly reveal evidence of asymmetryq (departures from the Gaussian assumptions), which cannot be explained by a standard LMM fit. In addition, the predictor space restricted to be linear combinations of covariates may not provide an elegant picture of their cross-sectional association with

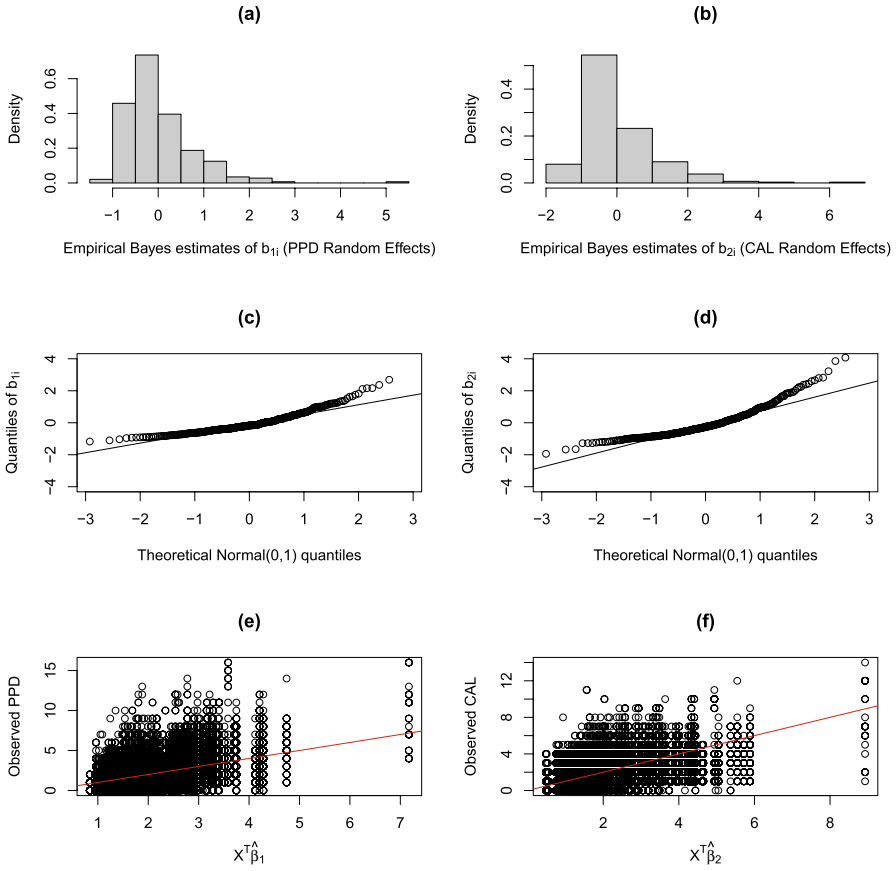

Fig. 1 GAAD Data: Plots of the empirical Bayes' estimates of random effects (panels a and b), corresponding Q-Q plots (panels c and d), and observed versus estimated (non-linear) curve (panels e and f), obtained from fitting a linear mixed model separately to the PPD and CAL responses

the (bivariate) response. Formulating an *index* for PD (that handles possible non-linearity, confounding, and interaction effects between the PD outcomes and the covariates) via a single-index model, or SIM (Hardle et al. 1993) can be a clinically elegant alternative. SIMs are a popular class of semiparametric regression models that relaxes the assumption of linearity, and bypass the 'curse of dimensionality' by reducing the multi-dimensional predictor space $\mathbf{X}$ into an univariate (scalar) index $U = \mathbf{X}^T\boldsymbol{\beta}$. A link function $g(.)$ now connects the covariate space to the response Y, offering a pragmatic compromise between a fully nonparametric (and often non-interpretable) multiple regression, and a restrictive (parametric) linear regression. Here, the magnitude of the index coefficient β_j determine the relative importance of the j-th predictor on the index, and $g(U)$ denotes the location of interest in the response curve at the index U. In biomedical research, the recent work by Wu and Tu (2016) develops an adiposity index via a (multivariate) SIM to efficiently predict multiple longitudinal outcomes (systolic and diastolic blood pressure) in

children. However, their proposal considers the usual MVN assumptions for the random terms (errors and effects), and may not well accommodate heavy tailed and other non-Gaussian features. Furthermore, they did not provide rigorous theoretical justification.

Considering Wu and Tu (2016) as our starting point, we seek to develop an index that can efficiently predict the clustered bivariate (PPD and CAL) PD outcomes. Such a *clinical* index that links both outcomes is vastly absent in the oral health literature. Our bivariate single-index mixed (BV-SIM) model tackles non-Gaussian features in the responses via the multivariate asymmetric Laplace density (ALD; Kotz et al. 2001) assumptions in the random terms. The multivariate ALD can accommodate asymmetric, peaked, and heavy-tailed data using fewer number of parameters than the popular multivariate skew-t density (Gupta 2003). The multivariate symmetric Laplace density (Naik and Plungpongpun 2006), a special case of the ALD, has been applied in other fields, such as speech clustering, classification problems, and image/signal analysis. Under this framework, we consider a polynomial spline approximation to the nonparametric index function, and propose an efficient EM-type algorithm for estimation and inference. The spline approximation, and the mixture normal representation of the multivariate ALD presents a computationally efficient, and intuitively appealing estimation setup, quantifying correlations from both sources.

The rest of the paper is organized as follows. In Sect. 2, we propose the BV-SIM model under the assumptions of a multivariate asymmetric Laplace density. Using the polynomial splines approximation for the nonparametric (index) functions, we derive the maximum likelihood (ML) estimate, and establish the large sample properties of the proposed estimators in Sect. 3, with the detailed technical proofs relegated to the Appendix, where we use the projection method to prove the asymptotic normality of parametric part. In Sect. 4, we develop an efficient MLE procedure based on the EM-algorithm. Simulation studies comparing finite sample performance of our approach to other alternatives appear in Sect. 5, while Sect. 6 illustrates the method via application to the PD dataset. Finally, some concluding remarks are presented in Sect. 7.

2 Statistical Model

We begin with a sketch of the multivariate shifted Laplace density (Kotz et al. 2001), and then develop our SIM mixed effects framework for bivariate clustered data. The multivariate ALD has the density

$$p(\mathbf{y};\boldsymbol{\Sigma},\boldsymbol{\gamma}) = \frac{2\exp\{\mathbf{y}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}\}}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \times \left(\frac{\mathbf{y}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{y}}{2+\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}\right)^{\nu/2} K_{\nu}(u), \tag{2.1}$$

where K_{ν} is the modified Bessel function of the third kind with index ν, $\nu = (2-d)/2$, $u = \sqrt{(2+\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma})(\mathbf{y}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{y})}$, $\boldsymbol{\gamma} \in \mathbb{R}^d$ is a skewness parameter and $\boldsymbol{\Sigma}$ is a positive definite (p.d.) scatter matrix with dimension $d \times d$. We denote (2.1) as

$\text{ALD}_d(\boldsymbol{\Sigma}, \boldsymbol{\gamma})$. Note, the ALD forces each component density to be joined at the same origin. An extension, the multivariate shifted asymmetric Laplace distribution (SALD; Kotz et al. 2001), has the form

$$p(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\gamma}) = \frac{2\exp\{(\mathbf{y}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}\}}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \times \left(\frac{\delta(\mathbf{y},\boldsymbol{\mu},\boldsymbol{\Sigma})}{2+\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}\right)^{\nu/2} K_{\nu}(u), \tag{2.2}$$

where $u = \sqrt{(2+\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma})\delta(\mathbf{y},\boldsymbol{\mu},\boldsymbol{\Sigma})}$, $\delta(\mathbf{y},\boldsymbol{\mu},\boldsymbol{\Sigma}) = (\mathbf{y}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})$, and $\nu, \boldsymbol{\gamma}, \boldsymbol{\Sigma}$ are defined in (2.1). Here, we use the notation $\mathbf{Y} \sim \text{SAL}_d(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\gamma})$ to denote the random variable $\mathbf{y}$ following a d-dimensional SALD. After some calculations, the mean and variance of SALD are given by

$$\text{E}(\mathbf{Y}) = \boldsymbol{\mu} + \boldsymbol{\gamma} \quad \text{and} \quad \text{Var}(\mathbf{Y}) = \boldsymbol{\Sigma} + \boldsymbol{\gamma}\boldsymbol{\gamma}^{\mathrm{T}}.$$

It is clear that the mean depends on the shifted location parameter $\boldsymbol{\mu}$ and skewness parameter $\boldsymbol{\gamma}$, while its variance depends on scatter matrix $\boldsymbol{\Sigma}$ and skewness parameter $\boldsymbol{\gamma}$. Also, $\boldsymbol{\Sigma} + \boldsymbol{\gamma}\boldsymbol{\gamma}^{\mathrm{T}}$ must be p.d. if $\boldsymbol{\Sigma}$ is p.d. The parameter $\boldsymbol{\gamma}$ plays an important role in multivariate asymmetric data analysis, besides the location $\boldsymbol{\mu}$ and scatter matrix $\boldsymbol{\Sigma}$. Note, the multivariate density in (2.2) reduces to (2.1) when $\boldsymbol{\mu} = \mathbf{0}$, and it further reduces to the multivariate symmetry Laplace distribution (Eltoft et al. 2006) when $\boldsymbol{\gamma} = \mathbf{0}$. Moreover, (2.2) reduces to the univariate ALD when dimension $d = 1$, $\gamma = (1-2\tau)/\tau(1-\tau)$ and $\boldsymbol{\Sigma}_{1\times 1} = 2/\tau(1-\tau)$, and is popularly used in the likelihood framework for quantile regression with density $p(y) = \tau(1-\tau)\exp\{-\rho_{\tau}(y-\mu)\}$, where $\rho_{\tau}(u) = u(\tau - I(u<0))$. The SALD in (2.2) has the following stochastic representation

$$\mathbf{Y} = \boldsymbol{\mu} + V\boldsymbol{\gamma} + \sqrt{V}\mathbf{Z}, \tag{2.3}$$

where V is a random variable from an exponential distribution with mean 1 and $\mathbf{Z} \sim \text{N}_d(0,\boldsymbol{\Sigma})$ is generated independent of V. Using Bayes's theorem, the density of V given $\mathbf{Y} = \mathbf{y}$ is generalized inverse Gaussian, with the density

$$p_V(v|\mathbf{Y}=\mathbf{y}) = \frac{v^{\nu-1}}{2K_{\nu}(u)}\left(\frac{\delta(\mathbf{y},\boldsymbol{\mu},\boldsymbol{\Sigma})}{2+\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}\right)^{-\nu/2} \exp\left\{-\frac{1}{2v}\delta(\mathbf{y},\boldsymbol{\mu},\boldsymbol{\Sigma}) - \frac{v}{2}(2+\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma})\right\}, \tag{2.4}$$

where $\nu, \boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \delta(\mathbf{y},\boldsymbol{\mu},\boldsymbol{\Sigma})$ and u are as defined in (2.2). The SALD allows for peakedness, heavy tails, and skewness, and hence provides more flexibility in modeling multivariate data with non-Gaussian features. More properties, extensions and applications of SALD appear in Kozubowski and Podgórski (2001); Franczak et al. (2014); Bouveyron and Brunet-Saumard (2014).

2.1 Single-Index Mixed-Effects Model

Let $\mathbf{y}_{ij} = (y_{ij}^{(1)}, y_{ij}^{(2)})^{\mathrm{T}}$ be the observed values of two response variables (here, mean PPD and CAL) for the ith subject at the jth location (here, tooth), where $i = 1,\ldots,n$ and $j = 1,\ldots,m_i$. We assume

$$\begin{cases} \mathbf{y}_{ij} = \widetilde{\boldsymbol{\mu}}_{ij} + \boldsymbol{\epsilon}_{ij}, \;\; \widetilde{\boldsymbol{\mu}}_{ij} = (\widetilde{\mu}_{ij}^{(1)}, \widetilde{\mu}_{ij}^{(2)})^{\mathrm{T}}, \\ \widetilde{\mu}_{ij}^{(1)} = g_1((\mathbf{x}_{ij}^{(1)})^{\mathrm{T}}\boldsymbol{\beta}_1) + (\mathbf{z}_{ij}^{(1)})^{\mathrm{T}}\mathbf{b}_{i1}, \;\; \widetilde{\mu}_{ij}^{(2)} = g_2((\mathbf{x}_{ij}^{(2)})^{\mathrm{T}}\boldsymbol{\beta}_2) + (\mathbf{z}_{ij}^{(2)})^{\mathrm{T}}\mathbf{b}_{i2}, \\ \boldsymbol{\epsilon}_{ij} \sim \mathrm{SAL}_2(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}), i.i.d. \;\; \forall \;\; i, j, \end{cases} \tag{2.5}$$

where g_1 and g_2 are two unknown nonparametric functions, $\mathbf{x}_{ij}^{(1)} = (x_{ij1}^{(1)}, \ldots, x_{ijp_1}^{(1)})^{\mathrm{T}}$, $\mathbf{x}_{ij}^{(2)} = (x_{ij1}^{(2)}, \ldots, x_{ijp_2}^{(2)})^{\mathrm{T}}$, and $\mathbf{z}_{ij}^{(1)} = (1, z_{ij1}^{(1)}, \ldots, z_{ijq_1}^{(1)})^{\mathrm{T}}$, $\mathbf{z}_{ij}^{(2)} = (1, z_{ij1}^{(2)}, \ldots, z_{ijq_2}^{(2)})^{\mathrm{T}}$, $\boldsymbol{\beta}_k \in \mathbb{R}^{p_k}$ and $\mathbf{b}_{ik} \in \mathbb{R}^{q_k+1}$ are the (fixed) index coefficients and random effect for the k-th response (k=1 or 2), $\boldsymbol{\gamma}$ is a 2×1 vector of skewness parameters, and $\boldsymbol{\Sigma}$ is the scatter matrix with dimension 2×2 for the random error $\boldsymbol{\epsilon}$. To accommodate a robust specification, we also assume the random effects $\mathbf{b}_i = (\mathbf{b}_{i1}{}^{\mathrm{T}}, \mathbf{b}_{i2}{}^{\mathrm{T}})^{\mathrm{T}} \sim \mathrm{SAL}_{(q_1+q_2+2)}(\mathbf{0}, \boldsymbol{\Omega}, \mathbf{0})$, where $\boldsymbol{\Omega}$ is an unstructured covariance matrix with dimension $(q_1 + q_2 + 2) \times (q_1 + q_2 + 2)$. Note, $\boldsymbol{\Omega}$ carries information pertaining to both the clustering correlation within a response found on the two blocks of diagonal sub-matrices, with dimensions $(q_1 + 1) \times (q_1 + 1)$ and $(q_2 + 1) \times (q_2 + 1)$, and the cross-correlations between responses, found on the off-diagonal sub-matrices. In addition, we further assume the joint density of $(\boldsymbol{\epsilon}_{ij}{}^{\mathrm{T}}, \mathbf{b}_i{}^{\mathrm{T}})^{\mathrm{T}}$ is $\mathrm{SAL}_{(q_1+q_2+4)}(\mathbf{0}_{(q_1+q_2+4)}, \mathrm{blockdiag}(\boldsymbol{\Sigma}, \boldsymbol{\Omega}), (\boldsymbol{\gamma}^{\mathrm{T}}, \mathbf{0}_{q_1+q_2+4}^{\mathrm{T}})^{\mathrm{T}})$. We call model (2.5) as the single-index mixed-effects (SIME) model for bivariate clustered data.

For identifiability, we assume both $\|\boldsymbol{\beta}_1\| = 1$ and $\|\boldsymbol{\beta}_2\| = 1$, and their first components are positive, respectively. In this paper, the popular "delete one component" method is used to avoid the equality constraints (Yu and Ruppert 2002; Cui et al. 2011). Specifically, we write $\boldsymbol{\beta}_1 = ((1 - \|\boldsymbol{\beta}_1^{(-1)}\|^2)^{1/2}, \beta_{12}, \ldots, \beta_{1p_1})^{\mathrm{T}}$where, $\boldsymbol{\beta}_1^{(-1)} = (\beta_{12}, \ldots, \beta_{1p_1})^{\mathrm{T}}$. Under this parametrization, $\boldsymbol{\beta}_1$ is a smooth deterministic function of $\boldsymbol{\beta}_1^{(-1)}$, with its Jacobian matrix given by

$$\mathbf{J}_1 = \frac{\partial \boldsymbol{\beta}_1}{\partial \boldsymbol{\beta}_1^{(-1)}} = \begin{pmatrix} -\frac{\boldsymbol{\beta}_1^{(-1)}}{(1-\|\boldsymbol{\beta}_1^{(-1)}\|^2)^{1/2}} \\ \mathbf{I}_{p_1-1} \end{pmatrix},$$

where $\mathbf{I}_{p_1-1}$ is the identity matrix with $p_1 - 1$ rows/columns. The true parameter $\boldsymbol{\beta}_1^{(-1)}$ satisfies the constraint $\boldsymbol{\beta}_1^{(-1)} < 1$, which implies that it is a interior point in a unit ball in $\mathbb{R}^{p_1-1}$. Therefore, $\boldsymbol{\beta}_1$ is infinitely differentiable in a neighborhood of $\boldsymbol{\beta}_1^{(-1)}$. Similarly, we define $\boldsymbol{\beta}_2^{(-1)}$ and $\mathbf{J}_2$, and let $\boldsymbol{\beta}^{(-1)} = ((\boldsymbol{\beta}_1^{(-1)})^{\mathrm{T}}, (\boldsymbol{\beta}_2^{(-1)})^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{J} = \mathrm{blockdiag}(\mathbf{J}_1, \mathbf{J}_2)$. Applying the stochastic representation in (2.3), model (2.5) admits the following hierarchical structure:

$$\begin{cases} \mathbf{y}_i | \mathbf{b}_i, V_i \;\sim\; \mathrm{N}_{2m_i}(\widetilde{\boldsymbol{\mu}}_i + V_i(\mathbf{1}_{m_i} \otimes \boldsymbol{\gamma}), V_i\boldsymbol{\Lambda}_i), \\ \mathbf{b}_i | V_i \sim \mathrm{N}_{2(q+1)}(\mathbf{0}, V_i\boldsymbol{\Omega}), \;\; V_i \sim \mathrm{E}(1), \end{cases} \tag{2.6}$$

where $\mathbf{y}_i = (\mathbf{y}_{i1}{}^{\mathrm{T}}, \ldots, \mathbf{y}_{im_i}{}^{\mathrm{T}})^{\mathrm{T}}$, $\widetilde{\boldsymbol{\mu}}_i = (\widetilde{\boldsymbol{\mu}}_{i1}{}^{\mathrm{T}}, \ldots, \widetilde{\boldsymbol{\mu}}_{im_i}{}^{\mathrm{T}})^{\mathrm{T}}$, E denotes the exponential distribution, $\boldsymbol{\Lambda}_i = \mathbf{I}_{m_i} \otimes \boldsymbol{\Sigma}$, where $\otimes$ denotes the kronecker product, and $\mathbf{1}_{m_i}$ is a m_i column vector with element 1. From (2.5) and (2.6), it is clear that conditional on V_i, $\boldsymbol{\epsilon}_{ij}$ and $\mathbf{b}_i$ are independent. Integrating out $\mathbf{b}_i$ in (2.6), we have the following hierarchical model

$$\mathbf{y}_i | V_i \sim \mathrm{N}_{2m_i}(\boldsymbol{\mu}_i + V_i(\mathbf{1}_{m_i} \otimes \boldsymbol{\gamma}), V_i \mathbf{G}_i), \quad V_i \sim \mathrm{E}(1), \tag{2.7}$$

where $\boldsymbol{\mu}_i = ((\boldsymbol{\mu}_{i1})^{\mathrm{T}}, \ldots, (\boldsymbol{\mu}_{im_i})^{\mathrm{T}})^{\mathrm{T}}$ with $\boldsymbol{\mu}_{ij} = (g_1((\mathbf{x}_{ij}^{(1)})^{\mathrm{T}}\boldsymbol{\beta}_1), g_2((\mathbf{x}_{ij}^{(2)})^{\mathrm{T}}\boldsymbol{\beta}_2))^{\mathrm{T}}$, $\mathbf{Z}_i = (\mathbf{Z}_{i1}, \ldots, \mathbf{Z}_{im_i})$, $\mathbf{Z}_{ij} = \mathrm{blockdiag}(\mathbf{z}_{ij}^{(1)}, \mathbf{z}_{ij}^{(2)})$, $\mathbf{G}_i = \mathbf{Z}_i{}^{\mathrm{T}}\boldsymbol{\Omega}\mathbf{Z}_i + \boldsymbol{\Lambda}_i$. Moreover, it follows from (2.7) that the $\mathbf{y}_i$ are independent and marginally distributed as

$$\mathbf{y}_i \sim \mathrm{SALD}_{2m_i}(\boldsymbol{\mu}_i, \mathbf{G}_i, \boldsymbol{\gamma}_i^*), \quad i = 1, \ldots, n, \tag{2.8}$$

where $\boldsymbol{\gamma}_i^* = \mathbf{1}_{m_i} \otimes \boldsymbol{\gamma}$. From (2.7) and by the properties of the generalized inverse Gaussian distribution in (2.4), we have

$$\mathbb{E}(V_i|\mathbf{y}_i) = \sqrt{\frac{b_i}{a_i}} R_\nu(\sqrt{a_i b_i}) \text{ and } \mathbb{E}(V_i^{-1}|\mathbf{y}_i) = \sqrt{\frac{a_i}{b_i}} R_\nu(\sqrt{a_i b_i}) - \frac{2\nu}{b_i}, \tag{2.9}$$

where $a_i = 2 + (\boldsymbol{\gamma}_i^*)^{\mathrm{T}}\mathbf{G}_i^{-1}\boldsymbol{\gamma}_i^*$, $b_i = (\mathbf{y}_i - \boldsymbol{\mu}_i)^{\mathrm{T}}\mathbf{G}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i)$, $R_\nu(u) = K_{\nu+1}(u)/K_\nu(u)$ and $\nu = 1 - m_i$.

2.2 Modeling the Index Functions

Since the two functions g_1 and g_2 in (2.5) are unknown, we use polynomial splines to approximate them in the subsequent ML estimation. Polynomial splines are simple, yet practical tools with computational tractability and statistical efficiency, and has been proven to be an extremely powerful method for smoothing.

For simplicity, we assume that the covariates $\mathbf{x}_{ij}^{(1)}$ and $\mathbf{x}_{ij}^{(2)}$ are bounded and the supports of $(\mathbf{x}^{(1)})^{\mathrm{T}}\boldsymbol{\beta}_{10}$ and $(\mathbf{x}^{(2)})^{\mathrm{T}}\boldsymbol{\beta}_{20}$ are contained in the finite interval $[a, b]$. Such a compactness assumption is almost always used in nonparametric regression with spline approximation. We use polynomial splines to approximate the nonparametric functions g_1 and g_2. Let $t_0 = a < t_1 < \cdots < t_{K'} < b = t_{K'+1}$ be the partitions of $[a, b]$ into subintervals $[t_k, t_{k+1}), k = 0, \ldots, K'$ with K' internal knots. A polynomial spline of order d is a function whose restriction to each subinterval is a polynomial of degree $d - 1$ and globally $d - 2$ times continuously differentiable on $[a, b]$. The collection of splines with a fixed sequence of knots has a B-spline basis $\{B_1(x), \ldots, B_K(x)\}$, with $K = K' + d$. We assume the B-spline basis is normalized to have $\sum_{k=1}^{K} B_k(x) = \sqrt{K}$, although, any scaling can be used without changing the theoretical results.

Let $\mathbf{B}_1(\cdot) = (B_1(\cdot), \ldots, B_{K_1}(\cdot))^{\mathrm{T}}$ and $\mathbf{B}_2(\cdot) = (B_1(\cdot), \ldots, B_{K_2}(\cdot))^{\mathrm{T}}$, where $K_1 = K_1' + d$ and $K_2 = K_2' + d$ with number of knots K_1' and K_2' for g_1 and g_2. Then, we have $g_k(\cdot) \approx \mathbf{B}_k{}^{\mathrm{T}}(\cdot)\boldsymbol{\theta}_k, k = 1, 2$ where $\boldsymbol{\theta}_k = (\theta_{k1}, \ldots, \theta_{kK_k})^{\mathrm{T}}, k = 1, 2$. As a result, we can write

$$\mu_{ij}^{(1)} \approx \mathbf{B}_1{}^{\mathrm{T}}((\mathbf{x}_{ij}^{(1)})^{\mathrm{T}}\boldsymbol{\beta}_1)\boldsymbol{\theta}_1 \text{ and } \mu_{ij}^{(2)} \approx \mathbf{B}_2{}^{\mathrm{T}}((\mathbf{x}_{ij}^{(2)})^{\mathrm{T}}\boldsymbol{\beta}_2)\boldsymbol{\theta}_2 \tag{2.10}$$

for $i = 1, \ldots, n, j = 1, \ldots, m_i$. By letting the number of knots increase with the sample size at an appropriate rate, the spline estimate of the unknown function can achieve the optimal nonparametric convergence rate.

3 Theoretical Properties

In this section, we will investigate the theoretical properties for the index parameters and the index functions. In the following we establish the large sample properties based on the marginal distribution (2.8) of the proposed BV-SIM model in (2.5). For simplicity, we assume $m_i \equiv m$, with the response viewed as i.i.d. data, $\mathbf{y}_i \sim \mathrm{SALD}_{2m}(\boldsymbol{\mu}_i, \mathbf{G}_i, \boldsymbol{\gamma}^*)$, $i = 1, \ldots, n$. In (2.8), $\boldsymbol{\gamma}^* = \mathbf{1}_m \otimes \boldsymbol{\gamma}$ and $\mathbf{G}_i = \mathbf{Z}_i{}^{\mathrm{T}}\boldsymbol{\Omega}\mathbf{Z}_i + \boldsymbol{\Lambda}$, with $\boldsymbol{\Lambda} = \mathbf{I}_m \otimes \boldsymbol{\Sigma}$. We first introduce some notations.

Let $\boldsymbol{\beta}_{01}$ and $\boldsymbol{\beta}_{02}$ be the true index parameters, and g_{01} and g_{02} the corresponding true index functions. Let $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}{}^{\mathrm{T}}, \boldsymbol{\beta}_{02}{}^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\beta}_0^{(-1)} = ((\boldsymbol{\beta}_{01}^{(-1)})^{\mathrm{T}}, (\boldsymbol{\beta}_{02}^{(-1)})^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\mu}_i^0 = ((\boldsymbol{\mu}_{i1}^0)^{\mathrm{T}}, \ldots, (\boldsymbol{\mu}_{im_i}^0)^{\mathrm{T}})^{\mathrm{T}}$ with $\boldsymbol{\mu}_{ij}^0 = (g_{01}((\mathbf{x}_{ij}^{(1)})^{\mathrm{T}}\boldsymbol{\beta}_{01}), g_{02}((\mathbf{x}_{ij}^{(2)})^{\mathrm{T}}\boldsymbol{\beta}_{02}))^{\mathrm{T}}$. Denote the support of $\{\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0\}$ as $[a, b]$, where $a = \min_i\{\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0\}$ and $b = \max_i\{\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0\}$, $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{im_i})$ with $\mathbf{X}_{ij} = \mathrm{blockdiag}(\mathbf{x}_{ij}^{(1)}, \mathbf{x}_{ij}^{(2)})$. Let $\mathcal{H}_s$ be the collection of all functions on the support $[a, b]$ whose l-th order derivative satisfies the Hölder condition of the order r with $s = l + r$. Then, for each $g \in \mathcal{H}_s$, there exists a positive constant C_0 such that $|g^{(l)}(u) - g^{(l)}(v)| \le C_0|u - v|^r$, $\forall u, v \in [a, b]$. From De Boor (2001), there exists a constant C (see page 149) such that

$$\sup_{u\in[a,b]} |g_k(u) - \mathbf{B}_k^T(u)\boldsymbol{\theta}_{0k}| \le CK_k^{-s}, \tag{3.1}$$

if $g_k \in \mathcal{H}_s$, where $\boldsymbol{\theta}_{0k} = (\theta_{0k1}, \ldots, \theta_{0kK_k})^T, k = 1, 2$ are the true value of spline coefficients, which can be viewed as the best approximation coefficient vectors for g_k.

Denote $\boldsymbol{\delta} = (\boldsymbol{\gamma}^{\mathrm{T}}, \mathrm{vech}(\boldsymbol{\Omega})^{\mathrm{T}}, \mathrm{vech}(\boldsymbol{\Sigma})^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\Theta}$ as the parameter space of $\boldsymbol{\zeta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\delta}^{\mathrm{T}})^{\mathrm{T}}$. Given the covariates $\mathbf{X}_i$ and $\mathbf{Z}_i$, let $\ell_m(\boldsymbol{\mu}_i, \boldsymbol{\delta}, \mathbf{y}_i)$ be the log-likelihood of the marginal distribution for response $\mathbf{y}_i$ in (2.8) and $\ell_m(\boldsymbol{\zeta}, \mathbf{y}_i) \triangleq \ell_m(\mathbf{W}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}, \boldsymbol{\delta}, \mathbf{y}_i)$ be the corresponding spline-approximated log-likelihood. Let $\boldsymbol{\delta}_0$ be the true value of $\boldsymbol{\delta}$ and $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{01}{}^{\mathrm{T}}, \boldsymbol{\theta}_{02}{}^{\mathrm{T}})^{\mathrm{T}}$. Define $\widehat{\boldsymbol{\zeta}} = (\widehat{\boldsymbol{\beta}}^{\mathrm{T}}, \widehat{\boldsymbol{\theta}}^{\mathrm{T}}, \widehat{\boldsymbol{\delta}}^{\mathrm{T}})^{\mathrm{T}}$ as the MLE, given by

$$\widehat{\boldsymbol{\zeta}} = \mathrm{argmax}_{\zeta} \sum_{i=1}^{n} \ell_m(\mathbf{W}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}, \boldsymbol{\delta}, \mathbf{y}_i), \tag{3.2}$$

where $\mathbf{W}_i(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}) = (\mathbf{W}_{i1}, \ldots, \mathbf{W}_{im_i})$, $\mathbf{W}_{ij} = \mathrm{blockdiag}(\mathbf{B}_{ij}^{(1)}, \mathbf{B}_{ij}^{(2)})$ with $\mathbf{B}_{ij}^{(k)} = \mathbf{B}_k((\mathbf{x}_{ij}^{(k)})^{\mathrm{T}}\boldsymbol{\beta}_k), k = 1, 2$. Define the space of square integrable single-index functions $\mathcal{G} = \{\mathbf{g} : \mathbb{E}\|\mathbf{g}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\|^2 < \infty\}$, where $\mathbf{g}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}) = (\mathbf{g}^{\mathrm{T}}(\mathbf{X}_{i1}{}^{\mathrm{T}}\boldsymbol{\beta}), \ldots, \mathbf{g}^{\mathrm{T}}(\mathbf{X}_{im_i}{}^{\mathrm{T}}\boldsymbol{\beta}))^{\mathrm{T}}$

with $\mathbf{g}(\mathbf{X}_{ij}{}^{\mathrm{T}}\boldsymbol{\beta}) = (g_1((\mathbf{x}_{ij}^{(1)})^{\mathrm{T}}\boldsymbol{\beta}_1), g_2((\mathbf{x}_{ij}^{(2)})^{\mathrm{T}}\boldsymbol{\beta}_2))^{\mathrm{T}}$. Denote $\mathbf{C}_i(\boldsymbol{\mu}_i, \boldsymbol{\delta}) = -\partial^2 \ell_m(\boldsymbol{\mu}_i, \boldsymbol{\delta}, \mathbf{y}_i)/\partial\boldsymbol{\mu}_i\partial\boldsymbol{\mu}_i{}^{\mathrm{T}}$ and $\mathbf{C}_i^0 = \mathbf{C}_i(\boldsymbol{\mu}_i^0, \boldsymbol{\delta}_0)$. Then, the projection of a $2m$-dimensional random vector $\boldsymbol{\Gamma}$ onto $\mathcal{G}$ (defined as $\mathbb{E}[\boldsymbol{\Gamma}] = \mathbf{g}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)$) is the minimizer of

$$\min_{\mathbf{g}\in\mathcal{G}} \mathbb{E}\left[(\boldsymbol{\Gamma} - \mathbf{g}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0))^{\mathrm{T}}\mathbf{C}_i^0(\boldsymbol{\Gamma} - \mathbf{g}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0))\right].$$

Note, the definition of projection involves the distributions of both $\mathbf{X}_i, \mathbf{Z}_i$ and $\boldsymbol{\Gamma}$ since we take the expectation over these random variables. This definition can be extended to any $2m \times L$ matrix by column-wise projection. In the following, we list the regularity conditions (Wang et al. 2014; Lian and Liang 2013; Zhao et al. 2017) that are necessary to study the asymptotic behavior of the MLEs.

(A1) Both $g_1(\cdot) \in \mathcal{H}_s$ and $g_2(\cdot) \in \mathcal{H}_s$ for some $s \geq 2$.

(A2) Both $\mathbf{x}_{ij}^{(1)}$ and $\mathbf{x}_{ij}^{(2)}, i = 1, \ldots, n, j = 1, \ldots, m_i$, are bounded, with density supported on a convex set.

(A3) The true parameter point $\boldsymbol{\zeta}_0$ is an interior point of the parameter space $\boldsymbol{\Theta}$.

(A4) The log-likelihood $\ell_m(\boldsymbol{\zeta}, \mathbf{y}_i)$ is at least thrice differentiable on parameters $\boldsymbol{\zeta}$. Furthermore, the second derivatives of the likelihood function satisfy the equations

$$\mathbb{E}\left\{\left(\frac{\partial \ell_m(\boldsymbol{\zeta}, \mathbf{y}_i)}{\partial \boldsymbol{\zeta}}\right)\left(\frac{\partial \ell_m(\boldsymbol{\zeta}, \mathbf{y}_i)}{\partial \boldsymbol{\zeta}}\right)^{\mathrm{T}}\right\} = -\mathbb{E}\left\{\frac{\partial^2 \ell_m(\boldsymbol{\zeta}, \mathbf{y}_i)}{\partial \boldsymbol{\zeta}\partial \boldsymbol{\zeta}^{\mathrm{T}}}\right\}.$$

Also, there exists functions $M_{jkl}(\mathbf{y}_i)$, such that

$$\left|\frac{\partial^3 \ell_m(\boldsymbol{\zeta}, \mathbf{y}_i)}{\partial \boldsymbol{\zeta}_j \partial \boldsymbol{\zeta}_k \partial \boldsymbol{\zeta}_l}\right| \leq M_{jkl}(\mathbf{y}_i)$$

for $\boldsymbol{\zeta} \in \boldsymbol{\Theta}$, and $\mathbb{E}[M_{jkl}(\mathbf{y}_i)] < C_3 < +\infty$. Here $\boldsymbol{\zeta}_j$ denotes the j-th component of $\boldsymbol{\zeta}$.

(A5) The Fisher information matrix $\mathcal{I}(\boldsymbol{\zeta}_0) = -\mathbb{E}\left\{\frac{\partial^2 \ell_m(\boldsymbol{\zeta}, \mathbf{y}_i)}{\partial \boldsymbol{\zeta}\partial \boldsymbol{\zeta}^{\mathrm{T}}}\right\}\Big|_{\boldsymbol{\zeta}_0}$ satisfies the conditions

$$0 < C_1 < \lambda_{\min}\{\mathcal{I}(\boldsymbol{\zeta}_0)\} \leq \lambda_{\max}\{\mathcal{I}(\boldsymbol{\zeta}_0)\} < C_2 < +\infty,$$

where $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest and largest eigenvalues of a matrix.

(A6) Suppose $\mathbb{E}_{\mathcal{G}}[\mathbf{X}_{ij}\mathrm{diag}\{\dot{\mathbf{g}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\}] = (h_1(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0), \ldots, h_{p_1+p_2}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0))^{\mathrm{T}}$. Assume all $h_j \in \mathcal{H}_{s'}$ with $s' > 1$. We also assume that

$$\mathbb{E}\left[(\mathbf{J}^{\mathrm{T}}\mathbf{X}_i\mathrm{diag}\{\dot{\mathbf{g}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\} - \mathbb{E}_{\mathcal{G}}[\mathbf{J}^{\mathrm{T}}\mathbf{X}_i\mathrm{diag}\{\dot{\mathbf{g}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\}])^{\otimes 2}\right]$$

is positive definite, where $\mathbf{J}$ is evaluated at $\boldsymbol{\beta}_0$.

Remark 1 The smoothness condition in (A1) is a requirement to attain the best convergence rate for single-index functions approximated in the spline space. Condition (A2) is widely used in the single-index modeling literature, ensuring that the index functions are defined in a compact set and thus facilitates the technical derivations. Conditions (A3) and (A4) are two common assumptions in the literature of maximum likelihood estimation with spline approximations (Wang et al. 2011, 2014), implying that the information matrix of the likelihood function is positive definite. Condition (A5) is slightly stronger than that used in the usual asymptotic likelihood theory, however, widely used in high-dimensional likelihood estimation literature Fan and Peng (2004). Finally, Condition (A6) is related to the 'projection', or the 'orthogonalization' technique common in a semiparametric setup, which includes partially linear model (Li 2000), partially linear additive model (Lian and Liang 2013), and single-index models (Cui et al. 2011; Zhao et al. 2017).

Denote $K = \max\{K_1, K_2\}$, and let $r_n = \sqrt{K/n} + K^{-s}$. Then, we have the following result.

Theorem 1 *Under the Conditions* (A1)–(A5), *suppose that* $K^4/n \to 0$, $\sqrt{n}K^{-2s+1} \to 0$, *then we have*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| + \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p(r_n).$$

As an immediate implication of Theorem 1, we have $\|\widehat{g}_1 - g_1\| = O_p(r_n)$ and $\|\widehat{g}_2 - g_2\| = O_p(r_n)$.

Remark 2 Note that the rate of convergence for nonparametric functions is $O_p(n^{-s/(2s+1)})$ if the optimal $K \sim n^{1/(2s+1)}$, which is the same as that found in the nonparametric and semiparametric literature.

Theorem 2 *Under Conditions* (A1)–(A6), *suppose that* $K^4/n \to 0$, $\sqrt{n}K^{-2s+1} \to 0$ *and* $\sqrt{n}K^{-s-s'} \to 0$. *Then, we have*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}^{(-1)} - \boldsymbol{\beta}_0^{(-1)}) \xrightarrow{\mathrm{d}} N(\mathbf{0}, \boldsymbol{\Psi}^{-1}),$$

where

$$\begin{aligned}\boldsymbol{\Psi} = \mathbb{E}\big[&(\mathbf{J}^{\mathrm{T}}\mathbf{X}_i \mathrm{diag}\{\dot{\mathbf{g}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\} - \mathbf{J}^{\mathrm{T}}\mathbb{E}_{\mathcal{G}}[\mathbf{X}_i \mathrm{diag}\{\dot{\mathbf{g}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\}]) \cdot \mathbf{C}_i^0 \cdot \\ &(\mathbf{J}^{\mathrm{T}}\mathbf{X}_i \mathrm{diag}\{\dot{\mathbf{g}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\} - \mathbf{J}^{\mathrm{T}}\mathbb{E}_{\mathcal{G}}[\mathbf{X}_i \mathrm{diag}\{\dot{\mathbf{g}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\}])^{\mathrm{T}}\big]\end{aligned}$$

and $\mathbf{J}$ *is evaluated at the true* $\boldsymbol{\beta}_0$.

Following Theorem 2 and invoking the Delta method, we have

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{\mathrm{d}} N(\mathbf{0}, \mathbf{J}\boldsymbol{\Psi}^{-1}\mathbf{J}^{\mathrm{T}}).$$

4 Maximum Likelihood Estimation

In this section, we develop the ML estimation for our BV-SIM model. We utilize EM-type algorithms for obtaining the MLE, based on two types of missing data structures in (2.6). The EM algorithm is a popular iterative algorithm for MLE in models with incomplete data (Dempster et al. 1977), where each iteration of the EM algorithm consists of two steps, the expectation (E) step and the maximization (M) step. Despite desirable features, the M-step in the EM algorithm is often difficult to implement for complicated models, and is replaced with a sequence of computationally simple conditional maximization (CM) steps, i.e. maximizing over one parameter with the other parameters held fixed. This leads to a simple extension of the EM algorithm, called the ECM algorithm (Meng and Rubin 1993).

Consider the hierarchical multivariate Laplace model in (2.6), where both V_i and $\mathbf{b}_i$ are missing data. Let $\mathbf{y} = (\mathbf{y}_1{}^{\mathrm{T}}, \ldots, \mathbf{y}_n{}^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{b} = (\mathbf{b}_1{}^{\mathrm{T}}, \ldots, \mathbf{b}_n{}^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{V} = (V_1, \ldots, V_n)^{\mathrm{T}}$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1{}^{\mathrm{T}}, \boldsymbol{\theta}_2{}^{\mathrm{T}})^{\mathrm{T}}$. The log-likelihood for the complete data in the multivariate Laplace single-index mixed-effects model up to an additive constant can be written as

$$\ell(\boldsymbol{\beta}, \theta, \boldsymbol{\gamma}, \boldsymbol{\Sigma}, \boldsymbol{\Omega} | \mathbf{y}, \mathbf{b}, \mathbf{V}) = \ell_1(\boldsymbol{\beta}, \theta, \boldsymbol{\gamma}, \boldsymbol{\Sigma} | \mathbf{y}, \mathbf{b}, \mathbf{V}) + \ell_2(\boldsymbol{\Omega} | \mathbf{b}, \mathbf{V}), \tag{4.1}$$

where

$$\ell_1(\boldsymbol{\beta}, \theta, \boldsymbol{\gamma}, \boldsymbol{\Sigma} | \mathbf{y}, \mathbf{b}, \mathbf{V}) = -\frac{N}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{m_i} V_i^{-1} (\mathbf{y}_{ij} - \widetilde{\boldsymbol{\mu}}_{ij} - V_i \boldsymbol{\gamma})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{y}_{ij} - \widetilde{\boldsymbol{\mu}}_{ij} - V_i \boldsymbol{\gamma})$$

and

$$\ell_2(\boldsymbol{\Omega} | \mathbf{b}, \mathbf{V}) = -\frac{n}{2} \log |\boldsymbol{\Omega}| - \frac{1}{2} \mathrm{trace} \left(\boldsymbol{\Omega}^{-1} \sum_{i=1}^{n} V_i^{-1} \mathbf{b}_i \mathbf{b}_i{}^{\mathrm{T}} \right),$$

where $\widetilde{\boldsymbol{\mu}}_{ij}$ is defined in (2.5) and $N = \sum_{i=1}^{n} m_i$. Note that ℓ_1 can be further written as

$$\begin{aligned} \ell_1 = & -\frac{N}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^{n} V_i^{-1} (\mathbf{y}_i - \mathbf{W}_i{}^{\mathrm{T}} \theta)^{\mathrm{T}} \boldsymbol{\Lambda}_i^{-1} (\mathbf{y}_i - \mathbf{W}_i{}^{\mathrm{T}} \theta) - \frac{1}{2} \sum_{i=1}^{n} V_i^{-1} \mathbf{b}_i{}^{\mathrm{T}} \mathbf{Z}_i \boldsymbol{\Lambda}_i^{-1} \mathbf{Z}_i{}^{\mathrm{T}} \mathbf{b}_i \\ & + \sum_{i=1}^{n} V_i^{-1} (\mathbf{y}_i - \mathbf{W}_i{}^{\mathrm{T}} \theta)^{\mathrm{T}} \boldsymbol{\Lambda}_i^{-1} \mathbf{Z}_i{}^{\mathrm{T}} \mathbf{b}_i - \sum_{i=1}^{n} (\boldsymbol{\gamma}_i^*)^{\mathrm{T}} \boldsymbol{\Lambda}_i^{-1} \mathbf{Z}_i{}^{\mathrm{T}} \mathbf{b}_i + \sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{W}_i{}^{\mathrm{T}} \theta)^{\mathrm{T}} \boldsymbol{\Lambda}_i^{-1} \boldsymbol{\gamma}_i^* \\ & - \frac{1}{2} \sum_{i=1}^{n} V_i (\boldsymbol{\gamma}_i^*)^{\mathrm{T}} \boldsymbol{\Lambda}_i^{-1} \boldsymbol{\gamma}_i^*. \end{aligned}$$

Denote $\boldsymbol{\eta}$ as the full parameter vector to be estimated. We firstly compute the conditional posterior mean and variance of $\mathbf{b}_i$ at the current estimate $\widehat{\boldsymbol{\eta}}$, leading to

$$\mathrm{Cov}(\mathbf{b}_i | \boldsymbol{\eta} = \widehat{\boldsymbol{\eta}}, \mathbf{y}, \mathbf{V}) = V_i \left(\widehat{\boldsymbol{\Omega}}^{-1} + \mathbf{Z}_i \widehat{\boldsymbol{\Lambda}}_i^{-1} \mathbf{Z}_i{}^{\mathrm{T}} \right)^{-1} \triangleq V_i \cdot \widehat{\boldsymbol{\Delta}}_i,$$

$$\mathbb{E}(\mathbf{b}_i | \boldsymbol{\eta} = \widehat{\boldsymbol{\eta}}, \mathbf{y}, \mathbf{V}) = \widehat{\boldsymbol{\Delta}}_i \mathbf{Z}_i \widehat{\boldsymbol{\Lambda}}_i^{-1} (\mathbf{y}_i - \mathbf{W}_i{}^{\mathrm{T}} \widehat{\boldsymbol{\theta}} - V_i \widehat{\boldsymbol{\gamma}}_i^*) \triangleq \widehat{\mathbf{R}}_{i1} - V_i \widehat{\mathbf{R}}_{i2},$$

for $i = 1, \ldots, n$, where

$$\widehat{\boldsymbol{\Delta}}_i = \left(\widehat{\boldsymbol{\Omega}}^{-1} + \mathbf{Z}_i\widehat{\boldsymbol{\Lambda}}_i^{-1}\mathbf{Z}_i^{\mathrm{T}}\right)^{-1}, \ \widehat{\mathbf{R}}_1 = \widehat{\boldsymbol{\Delta}}_i\mathbf{Z}_i\widehat{\boldsymbol{\Lambda}}_i^{-1}(\mathbf{y}_i - \mathbf{W}_i^{\mathrm{T}}\widehat{\boldsymbol{\theta}}) \text{ and } \widehat{\mathbf{R}}_2 = \widehat{\boldsymbol{\Delta}}_i\mathbf{Z}_i\widehat{\boldsymbol{\Lambda}}_i^{-1}\widehat{\boldsymbol{\gamma}}_i^*. \tag{4.2}$$

After obtaining the estimates of the conditional mean and conditional covariance of the random effect $\mathbf{b}_i$, we proceed to calculate the expectation of $\mathbb{E}(\ell(\cdot)) = \mathbb{E}_{\mathbf{V}}\{\mathbb{E}_{\mathbf{b}}[\ell(\cdot)|\mathbf{V}]\}$. Define the quantities

$$\widehat{c}_i = \mathbb{E}(V_i|\boldsymbol{\eta} = \widehat{\boldsymbol{\eta}}, \mathbf{y}) \text{ and } \widehat{d}_i = \mathbb{E}(V_i^{-1}|\boldsymbol{\eta} = \widehat{\boldsymbol{\eta}}, \mathbf{y}), \tag{4.3}$$

which can be computed from (2.9), using the current estimate $\widehat{\boldsymbol{\eta}}$. After some simple calculations, we have

$$\begin{aligned} Q_1 \ &\triangleq \ \mathbb{E}\left[\ell_1(\cdot|\mathbf{y}, \mathbf{b}, \mathbf{V})|\mathbf{y}, \boldsymbol{\eta} = \widehat{\boldsymbol{\eta}}\right] \\ &= \ -\tfrac{N}{2}\log|\boldsymbol{\Sigma}| - \tfrac{1}{2}\textstyle\sum_{i=1}^n \widehat{d}_i(\mathbf{y}_i - \mathbf{W}_i^{\mathrm{T}}\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{\Lambda}_i^{-1}(\mathbf{y}_i - \mathbf{W}_i^{\mathrm{T}}\boldsymbol{\theta}) - \tfrac{1}{2}\textstyle\sum_{i=1}^n \widehat{c}_i(\boldsymbol{\gamma}_i^*)^{\mathrm{T}}\boldsymbol{\Lambda}_i^{-1}\boldsymbol{\gamma}_i^* \\ &\quad - \tfrac{1}{2}\textstyle\sum_{i=1}^n \mathrm{trace}\left\{\mathbf{Z}_i\boldsymbol{\Lambda}_i^{-1}\mathbf{Z}_i^{\mathrm{T}}\left[\widehat{d}_i\widehat{\mathbf{R}}_{i1}\widehat{\mathbf{R}}_{i1}^{\mathrm{T}} - \widehat{\mathbf{R}}_{i1}\widehat{\mathbf{R}}_{i2}^{\mathrm{T}} - \widehat{\mathbf{R}}_{i2}\widehat{\mathbf{R}}_{i1}^{\mathrm{T}} + \widehat{c}_i\widehat{\mathbf{R}}_{i2}\widehat{\mathbf{R}}_{i2}^{\mathrm{T}} + \widehat{\boldsymbol{\Delta}}_i\right]\right\} \\ &\quad + \textstyle\sum_{i=1}^n \widehat{d}_i(\mathbf{y}_i - \mathbf{W}_i^{\mathrm{T}}\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{\Lambda}_i^{-1}\mathbf{Z}_i^{\mathrm{T}}\widehat{\mathbf{R}}_{i1} - \textstyle\sum_{i=1}^n (\mathbf{y}_i - \mathbf{W}_i^{\mathrm{T}}\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{\Lambda}_i^{-1}[\mathbf{Z}_i^{\mathrm{T}}\widehat{\mathbf{R}}_{i2} - \boldsymbol{\gamma}_i^*] \\ &\quad - \textstyle\sum_{i=1}^n (\boldsymbol{\gamma}_i^*)^{\mathrm{T}}\boldsymbol{\Lambda}_i^{-1}\mathbf{Z}_i^{\mathrm{T}}\widehat{\mathbf{R}}_{i1} + \textstyle\sum_{i=1}^n \widehat{c}_i(\boldsymbol{\gamma}_i^*)^{\mathrm{T}}\boldsymbol{\Lambda}_i^{-1}\mathbf{Z}_i^{\mathrm{T}}\widehat{\mathbf{R}}_{i2}, \end{aligned} \tag{4.4}$$

and

$$\begin{aligned} Q_2 \ &\triangleq \ \mathbb{E}\left[\ell_2(\cdot|\mathbf{y}, \mathbf{b}, \mathbf{V})|\mathbf{y}, \boldsymbol{\eta} = \widehat{\boldsymbol{\eta}}\right] \\ &= \ -\tfrac{n}{2}\log|\boldsymbol{\Omega}| - \tfrac{1}{2}\textstyle\sum_{i=1}^n \mathrm{trace}\left\{\boldsymbol{\Omega}^{-1}\left[\widehat{d}_i\widehat{\mathbf{R}}_{i1}\widehat{\mathbf{R}}_{i1}^{\mathrm{T}} - \widehat{\mathbf{R}}_{i1}\widehat{\mathbf{R}}_{i2}^{\mathrm{T}} - \widehat{\mathbf{R}}_{i2}\widehat{\mathbf{R}}_{i1}^{\mathrm{T}} + \widehat{c}_i\widehat{\mathbf{R}}_{i2}\widehat{\mathbf{R}}_{i2}^{\mathrm{T}} + \widehat{\boldsymbol{\Delta}}_i\right]\right\} + C, \end{aligned} \tag{4.5}$$

Next, maximizing Q_1 over parameters $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, and maximizing Q_2 over $\boldsymbol{\Omega}$, we can obtain their estimates, which constitutes the CM-steps 1-5 in the following ECM algorithm:

E-step Given current parameter estimates, for $i = 1, \ldots, n$, update c_i and d_i using (4.3), and update $\widehat{\boldsymbol{\Delta}}_i$, $\widehat{\mathbf{R}}_{i1}$ and $\widehat{\mathbf{R}}_{i2}$ by (4.2).

CM-step 1 Fix $\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\Sigma}}$, and update $\widehat{\boldsymbol{\theta}}$ by maximizing (4.4) over $\boldsymbol{\theta}$, which gives

$$\widehat{\boldsymbol{\theta}} = \left(\sum_{i=1}^n\sum_{j=1}^{m_i} \widehat{d}_i\mathbf{W}_{ij}\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{W}_{ij}^{\mathrm{T}}\right)^{-1} \sum_{i=1}^n\sum_{j=1}^{m_i} \mathbf{W}_{ij}\widehat{\boldsymbol{\Sigma}}^{-1}\left[\widehat{d}_i(\mathbf{y}_{ij} - \mathbf{W}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\theta}} - \mathbf{Z}_{ij}^{\mathrm{T}}\widehat{\mathbf{R}}_{i1}) + \mathbf{Z}_{ij}^{\mathrm{T}}\widehat{\mathbf{R}}_{i2} - \widehat{\boldsymbol{\gamma}}\right].$$

CM-step 2 Fix $\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\Sigma}}$, update $\widehat{\boldsymbol{\gamma}}$ by maximizing (4.4) over $\boldsymbol{\gamma}$, i.e.,

$$\widehat{\boldsymbol{\gamma}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}(\mathbf{y}_{ij} - \mathbf{W}_{ij}{}^{\mathrm{T}}\widehat{\boldsymbol{\theta}} - \mathbf{Z}_{ij}{}^{\mathrm{T}}\widehat{\mathbf{R}}_{i1} + \widehat{c}_i\mathbf{Z}_{ij}{}^{\mathrm{T}}\widehat{\mathbf{R}}_{i2})}{\sum_{i=1}^{n} m_i\widehat{c}_i}.$$

CM-step 3 Fix $\widehat{\boldsymbol{\theta}}$, $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\Sigma}}$, and update $\widehat{\boldsymbol{\beta}}$ by maximizing (4.4) over $\boldsymbol{\beta}$. Since there is no explicit expression for the estimate of the index parameter $\boldsymbol{\beta}$, we use the Newton–Raphson method to obtain $\widehat{\boldsymbol{\beta}}$, leading to the following iterative formula

$$\left(\widehat{\boldsymbol{\beta}}^{(-1)}\right)^{\text{new}} = \left(\widehat{\boldsymbol{\beta}}^{(-1)}\right)^{\text{old}} + \left(\sum_{i=1}^{n}\sum_{j=1}^{m_i}\widehat{d}_i\mathbf{H}_{ij}\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{H}_{ij}{}^{\mathrm{T}}\right)^{-1}\times$$

$$\times\sum_{i=1}^{n}\sum_{j=1}^{m_i}\mathbf{H}_{ij}\widehat{\boldsymbol{\Sigma}}^{-1}\left[\widehat{d}_i(\mathbf{y}_{ij} - \mathbf{W}_{ij}{}^{\mathrm{T}}\widehat{\boldsymbol{\theta}} - \mathbf{Z}_{ij}{}^{\mathrm{T}}\widehat{\mathbf{R}}_{i1}) + \mathbf{Z}_{ij}{}^{\mathrm{T}}\widehat{\mathbf{R}}_{i2} - \widehat{\boldsymbol{\gamma}}\right]$$

where $\mathbf{H}_{ij} = \begin{bmatrix} \mathbf{J}_1{}^{\mathrm{T}}\mathbf{x}_{ij}^{(1)}\{\dot{\mathbf{B}}_1{}^{\mathrm{T}}((\mathbf{x}_{ij}^{(1)})^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_1^{\text{old}})\widehat{\boldsymbol{\theta}}_1\} & \mathbf{0}_{(p_1-1)\times 1} \\ \mathbf{0}_{(p_2-1)\times 1} & \mathbf{J}_2{}^{\mathrm{T}}\mathbf{x}_{ij}^{(2)}\{\dot{\mathbf{B}}_2{}^{\mathrm{T}}((\mathbf{x}_{ij}^{(2)})^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_2^{\text{old}})\widehat{\boldsymbol{\theta}}_2\} \end{bmatrix}$, and $\dot{\mathbf{B}}(\cdot)$ denotes the first derivative of the spline basis $\mathbf{B}(\cdot)$.

CM-step 4 Fix $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\gamma}}$, and update $\widehat{\boldsymbol{\Sigma}}$ by maximizing (4.4) over $\boldsymbol{\Sigma}$. Denote

$$\widehat{\mathbf{D}} = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left\{\left[\widehat{d}_i(\mathbf{y}_{ij} - \mathbf{W}_{ij}{}^{\mathrm{T}}\widehat{\boldsymbol{\theta}} - 2\mathbf{Z}_{ij}{}^{\mathrm{T}}\widehat{\mathbf{R}}_{i1}) + 2(\mathbf{Z}_{ij}{}^{\mathrm{T}}\widehat{\mathbf{R}}_{i2} - \widehat{\boldsymbol{\gamma}})\right](\mathbf{y}_{ij} - \mathbf{W}_{ij}{}^{\mathrm{T}}\widehat{\boldsymbol{\theta}})^{\mathrm{T}} + \widehat{c}_i\widehat{\boldsymbol{\gamma}}\widehat{\boldsymbol{\gamma}}^{\mathrm{T}}\right\} +$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}\mathbf{Z}_{ij}{}^{\mathrm{T}}\left[\widehat{d}_i\widehat{\mathbf{R}}_{i1}\widehat{\mathbf{R}}_{i1}{}^{\mathrm{T}} - \widehat{\mathbf{R}}_{i1}\widehat{\mathbf{R}}_{i2}{}^{\mathrm{T}} - \widehat{\mathbf{R}}_{i2}\widehat{\mathbf{R}}_{i1}{}^{\mathrm{T}} + \widehat{c}_i\widehat{\mathbf{R}}_{i2}\widehat{\mathbf{R}}_{i2}{}^{\mathrm{T}} + \widehat{\boldsymbol{\Delta}}_i\right]\mathbf{Z}_{ij}$$

$$+\sum_{i=1}^{n}\sum_{j=1}^{m_i}(\mathbf{Z}_{ij}{}^{\mathrm{T}}\widehat{\mathbf{R}}_{i1} - \widehat{c}_i\mathbf{Z}_{ij}{}^{\mathrm{T}}\widehat{\mathbf{R}}_{i2})\widehat{\boldsymbol{\gamma}}^{\mathrm{T}}.$$

Applying the result in Lemma 1, we obtain $\widehat{\boldsymbol{\Sigma}} = \frac{1}{N}\widehat{\mathbf{D}}$.

CM-step 5 Update $\widehat{\boldsymbol{\Omega}}$ by maximizing (4.5) over $\boldsymbol{\Omega}$, which gives

$$\widehat{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{d}_i\widehat{\mathbf{R}}_{i1}\widehat{\mathbf{R}}_{i1}{}^{\mathrm{T}} - \widehat{\mathbf{R}}_{i1}\widehat{\mathbf{R}}_{i2}{}^{\mathrm{T}} - \widehat{\mathbf{R}}_{i2}\widehat{\mathbf{R}}_{i1}{}^{\mathrm{T}} + \widehat{c}_i\widehat{\mathbf{R}}_{i2}\widehat{\mathbf{R}}_{i2}{}^{\mathrm{T}} + \widehat{\boldsymbol{\Delta}}_i\right].$$

Repeat the above E-step and CM-steps, until all parameters achieve the desired convergence criterion. Since our estimation procedure requires initial values, we set $\widehat{\boldsymbol{\gamma}}^{(0)} = (0,0)^{\mathrm{T}}$, $\widehat{\boldsymbol{\Sigma}}^{(0)} = \mathbf{I}_2$, and the estimates of $\widehat{\boldsymbol{\beta}}_1^{(0)}$, $\widehat{\boldsymbol{\beta}}_2^{(0)}$ and $\widehat{\boldsymbol{\Omega}}^{(0)}$ are obtained from fitting a linear mixed model via the `R` package `lmer`, where $\mathbf{X}_{ij} = \text{blockdiag}(\mathbf{x}_{ij}^{(1)}, \mathbf{x}_{ij}^{(2)})$ and $\mathbf{Z}_{ij}$ are the design matrices corresponding to the fixed effects and random effects, respectively. Simulation studies (in Sect. 5) show that the above strategy works well.

5 Simulation Studies

In this section, we conduct extensive simulation studies using synthetic data to study the finite-sample performance of the model parameters in our proposed method (Simulation 1), and the robustness of our method when compared to existing alternatives, under data generated under various settings (Simulation 2).

5.1 Knots Selection

It is well-known that the performance of any spline estimation depends on the knots selection. Here, we employed Schwartz information criteria (SIC) for adaptive know selection (Ma and Song 2015; Lu 2017; Zhao et al. 2017). In view of the order $n^{1/(2s+1)}$ (of knots) to attain optimal convergence rate of nonparametric functions in 1, a sequence of knots are selected in a neighborhood of $n^{1/(2s+1)}$, such as $\left[0.5N_s, \min(5N_s, n^{1/2})\right]$, where $N_s = \lfloor n^{1/(2\,s+1)} \rfloor$, and s is the smoothing parameter. We choose $s = 2$ in both simulation studies and real data application. For simplicity, we use cubic polynomial splines and the number of interior knots $K_1 = K_2 \equiv K$ are the same for the two nonparametric link functions. The number K_{opt} corresponding to the minimum SIC value is defined as the optimal number of knots $\text{SIC}(K) = -\sum_{i=1}^{n} \log\hat{L}_i^K + \log n \times 2K$, where $\log\hat{L}_i^K$ denotes the estimated value of the log-likelihood function obtained from(2.8), with the given K knots.

5.2 Simulation 1: Assessing Finite-Sample Properties

Here, data is generated from the model (2.5), where the two nonparametric functions are $g_1(u) = 2\sin(\pi u)$ and $g_2(u) = 8u(1-u)$, with the true index parameters $\boldsymbol{\beta}_1 = (1/\sqrt{3}, -1/\sqrt{3}, 1/\sqrt{3})^{\mathrm{T}}$ and $\boldsymbol{\beta}_2 = (2/\sqrt{6}, 1/\sqrt{6}, 1/\sqrt{6})^{\mathrm{T}}$, respectively. Both covariates $\mathbf{x}_{ij}^{(1)}$ and $\mathbf{x}_{ij}^{(2)}$ are generated independently from the trivariate uniform distribution $U^3(0,1)$. The random effects $\mathbf{b}_i = (\mathbf{b}_{i1}{}^{\mathrm{T}}, \mathbf{b}_{i2}{}^{\mathrm{T}})^{\mathrm{T}}$ are generated from $\text{SAL}_4(\mathbf{0}, \boldsymbol{\Omega}, \mathbf{0})$, with covariance matrix

$$\boldsymbol{\Omega} = \begin{pmatrix} 9 & 4.8 & 3.6 & 0.6 \\ 4.8 & 4 & 2 & 1.2 \\ 3.6 & 2 & 4 & 1 \\ 0.6 & 1.2 & 1 & 1 \end{pmatrix},$$

and the corresponding covariates $\mathbf{z}_{ij}^{(1)} = (1, z_{ij1}^{(1)})^{\mathrm{T}}$ and $\mathbf{z}_{ij}^{(2)} = (1, z_{ij1}^{(2)})^{\mathrm{T}}$, where $z_{ij1}^{(1)}$ and $z_{ij1}^{(2)}$ are generated from the standard normal distribution. The random error $\boldsymbol{\epsilon}_{ij}$ is generated from $\text{SAL}_2(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\gamma})$ with $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}$ and $\boldsymbol{\gamma} = (2, 1.5)^{\mathrm{T}}$. The sample size n is set to be 50, 100 and 200, and the number of cluster members m_i in each subject is generated from the discrete uniform distribution on $5, 6, \ldots, 10$. Table 1 presents the

Table 1 Table entries are the average bias (BIAS), average absolute bias (ABIAS), and empirical standard error (ESE) estimates for $n = 50, 100, 200$, calculated over 400 replications, corresponding to Simulation 1

Parameters		β_{11}	β_{12}	β_{13}	β_{21}	β_{22}	β_{23}	γ_1	γ_2
$n = 50$	BIAS	0.0011	0.0007	−0.0011	0.0007	−0.0013	−0.0005	−0.0294	−0.0429
	ABIAS	0.0130	0.0125	0.0134	0.0061	0.0100	0.0096	0.2862	0.2479
	ESE	0.0174	0.0163	0.0169	0.0080	0.0128	0.0123	0.3596	0.3093
$n = 100$	BIAS	−0.0014	−0.0005	0.0005	−0.0001	0.0000	0.0000	−0.0417	−0.0144
	ABIAS	0.0093	0.0093	0.0085	0.0043	0.0066	0.0068	0.2299	0.1931
	ESE	0.0119	0.0120	0.0109	0.0053	0.0084	0.0085	0.2838	0.2424
$n = 200$	BIAS	0.0004	−0.0002	−0.0007	−0.0005	0.0004	0.0006	−0.0211	−0.0158
	ABIAS	0.0056	0.0055	0.0052	0.0029	0.0040	0.0041	0.1592	0.1309
	ESE	0.0072	0.0070	0.0067	0.0038	0.0052	0.0052	0.2084	0.1675

averages of bias, absolute bias, and the empirical standard error estimates for the index parameters and the skewness parameter, over 400 replications.

From Table 1, all biases are close to zero for all sample sizes, implying our proposed estimators are consistent. Moreover, the absolute biases and the standard errors are smaller with increasing sample sizes, with the estimation performance of index parameters significantly better than the skewness parameters. To further assess the estimation results, we calculate the integrated mean squared error (IMSE), defined as

$$\mathrm{IMSE}(g_l) = \frac{1}{400}\sum_{s=1}^{400}\sqrt{\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\{\widehat{g}_l^{(s)}((\mathbf{x}_{ij}^{(1)})^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_l) - g_l((\mathbf{x}_{ij}^{(1)})^{\mathrm{T}}\boldsymbol{\beta}_l)\}^2}, \quad l = 1, 2,$$

where $\widehat{g}_l^{(s)}(\cdot)$ is the spline approximation to $g_l(\cdot)$ in the sth simulation run. We report the average of the IMSE as $\mathrm{AIMSE} = \frac{1}{2}\sum_{l=1}^{2}\mathrm{IMSE}(g_l)$ in Table 2. For evaluating the estimation performances of the scatter matrix $\boldsymbol{\Sigma}$ (corresponding to the bivariate responses) and the covariance matrix $\boldsymbol{\Omega}$ (for the random effects), we use the Frobenius-norm of the matrix of differences between the estimated and true values, i.e. $\|\mathbf{A}\|_F = \sqrt{\mathrm{trace}(\mathbf{A}^{\mathrm{T}}\mathbf{A})}$, where $\mathbf{A}$ is either $\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}$ or $\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}$. Simulation results, together with the root of mean square error (RMSE) for $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\boldsymbol{\gamma}$ are listed in Table 2, where the RMSE for an arbitrary parameter $\boldsymbol{\delta}$ is defined as

Table 2 Table entries are the averages of the IMSE (AIMSE), the Frobenius-norms for $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$, and the root of mean squared errors (RMSE) of the model parameters, under various sample sizes ($n = 50, 100, 200$), calculated over 400 replications, corresponding to Simulation 1

	AIMSE	$\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_F$	$\|\widehat{\Omega} - \Omega\|_F$	$\mathrm{RMSE}_{\boldsymbol{\beta}_1}$	$\mathrm{RMSE}_{\boldsymbol{\beta}_2}$	$\mathrm{RMSE}_{\boldsymbol{\gamma}}$
$n = 50$	0.1397	0.2157	3.9259	0.0250	0.0169	0.4002
$n = 100$	0.0976	0.1894	2.4522	0.0173	0.0115	0.3172
$n = 200$	0.0616	0.1276	1.9059	0.0105	0.0071	0.2194

$\text{RMSE}_{\delta} = \sqrt{(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta})^{\mathrm{T}}(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta})}$. It is clear from Table 2 that the finite-sample performances of our proposed estimation procedures are satisfactory, with increasing sample sizes. In sum, the simulation results show that both index parameters, the nonparametric functions, and other parameters associated with the mixed effect models are reliably estimated, thereby confirming that our proposed algorithm works well in synthetic data settings.

5.3 Simulation 2: Assessing Robustness, in Light of Competing Methods

Here, the data is generated similar to Simulation 1 (from a BV-SIM), except that the random effects and errors are independently generated under the following four distributional assumptions:

Case 1: $\mathbf{b}_i \sim N(\mathbf{0}, \boldsymbol{\Omega})$, $\boldsymbol{\epsilon}_{ij} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$;
Case 2: $\mathbf{b}_i \sim t(\mathbf{0}, \boldsymbol{\Omega}, \nu)$, $\boldsymbol{\epsilon}_{ij} \sim t(\mathbf{0}, \boldsymbol{\Sigma}, \nu)$;
Case 3: $\mathbf{b}_i \sim \text{SAL}_4(\mathbf{0}, \boldsymbol{\Omega}, \mathbf{0})$, $\boldsymbol{\epsilon}_{ij} \sim \text{SAL}_2(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{0})$;
Case 4: $\mathbf{b}_i \sim 0.8N(\mathbf{0}, \boldsymbol{\Omega}) + 0.2N(\mathbf{0}, 10\boldsymbol{\Omega})$, $\boldsymbol{\epsilon}_{ij} \sim 0.8N(\mathbf{0}, \boldsymbol{\Sigma}) + 0.2N(\mathbf{0}, 10\boldsymbol{\Sigma})$,

for $i = 1, \cdots, n$, $j = 1, \cdots, m_i$,

Here, Case 1 corresponds to random effects and errors independently generated from the multivariate normal distribution. For Case 2, both are generated from the multivariate t-distribution with degree of freedom ν (setting $\nu = 5$). For Case 3, the random effects and errors are generated from the multivariate symmetric Laplace distribution with covariance matrix $\boldsymbol{\Omega}$ and $\boldsymbol{\Sigma}$, respectively. Finally, Case 4 corresponds to generating both the random terms (effects and errors) from multivariate normal mixtures. Note, for the above four cases, the bivariate clustered response is symmetric, since both the random effects and errors are generated from symmetric distributions. This is to make our approach comparable to the following two existing alternatives, (a) The bivariate normal mixed effect single-index model of Wu and Tu (2016), and (b) The bivariate mixed effect single-index model using the multivariate t-distribution, which extends the univariate linear mixed model proposal of (Pinheiro et al. 2001). In (a), penalized splines were used to approximate the nonparametric index function, whereas we use polynomial splines. At each replication, we use the same dataset to obtain the estimates from these three competing methods. We focus on the estimation of the index parameters and the index functions for the fixed effect part, with the same interpretation for all cases.

The results are summarized in Table 3. For all cases, RMSEs and AIMSEs decrease quickly as the sample size increases for all three methods. That said, our proposed method performs well for all four cases, and is significantly better than both the alternatives for Cases 3 and 4. The advantages of our method appears more prominent if we further reduce the mixing proportion of the mixture distribution in Case 4 from 0.8 to 0.7, 0.6 or 0.5 (results not reported here). In Cases 1 and 2, the performances of our method is comparable to the two others. In particular, our method performs almost similar to Pinheiro's t-distribution method in Case 2 when $n = 200$, while they are both better than the normal mixed-effects method of Wu and Tu (2016). To summarize,

Table 3 Table entries are the root of mean squared errors (RMSE) of β_1 and β_2, and the Average Integrated Mean Squared Error (AIMSE) from our model and the 2 competing models (Wu and Pinheiro), for $n = 50, 100, 200$, with data generated from the 4 cases described in Sect. 5.2

		Our proposal			Wu			Pinheiro		
		RMSE_{β_1}	RMSE_{β_2}	AIMSE	RMSE_{β_1}	RMSE_{β_2}	AIMSE	RMSE_{β_1}	RMSE_{β_2}	AIMSE
$n = 50$	Case 1	0.0554	0.0358	0.0847	0.0504	0.0325	0.0780	0.0551	0.0335	0.0798
	Case 2	0.0598	0.0408	0.1333	0.0618	0.0468	0.1434	0.0593	0.0402	0.1135
	Case 3	0.0229	0.0173	0.0237	0.0439	0.0329	0.0763	0.0376	0.0265	0.0511
	Case 4	0.0673	0.0438	0.1648	0.0861	0.0571	0.2145	0.0686	0.0444	0.1666
$n = 100$	Case 1	0.0352	0.0232	0.0451	0.0308	0.0219	0.0422	0.0345	0.0225	0.0436
	Case 2	0.0382	0.0267	0.0559	0.0424	0.0313	0.0578	0.0379	0.0266	0.0450
	Case 3	0.0129	0.0109	0.0115	0.0303	0.0243	0.0374	0.0257	0.0184	0.0250
	Case 4	0.0399	0.0277	0.0776	0.0542	0.0379	0.1113	0.0434	0.0279	0.0827
$n = 200$	Case 1	0.0238	0.0167	0.0197	0.0205	0.0154	0.0188	0.0234	0.0162	0.0194
	Case 2	0.0268	0.0187	0.0253	0.0282	0.0221	0.0315	0.0267	0.0188	0.0251
	Case 3	0.0087	0.0065	0.0075	0.0207	0.0157	0.0193	0.0187	0.0122	0.0127
	Case 4	0.0312	0.0211	0.0390	0.0406	0.0278	0.0564	0.0339	0.0213	0.0414

the performance of our proposed method appears to be satisfactory in all cases, and is robust to misspecified (non-Gaussian) random effects and errors, under a bivariate mixed model framework.

6 Application: GAAD Dataset

In this section, we illustrate our method via application to the GAAD dataset. Here, the tooth-level mean PPD and CAL measures are non-Gaussian bivariate responses representing PD status, and our objective is to evaluate the distribution of PD status for this population, and quantify the effects of various subject-level covariates such as Age (in years), body mass index (BMI), Gender ($1 = \text{Female}, 0 = \text{Male}$), Smoking status ($1 = \text{Smoker}, 0 = \text{Never Smoker}$) and glycemic level or HbA1c ($1 = \text{High/Uncontrolled}$), $0 = \text{Controlled}$) on the PD status. For our analysis, we have $n = 288$ subjects with complete covariate information. About 30% of the subjects are smokers. The mean age of the subjects is about 54 years with a range from 26–87 years. There is a predominance of female subjects (around 76%) in the data. Around 60% of subjects are obese (BMI ≥ 30), and 59% are with uncontrolled HbA1c. Each subject has varying number of teeth, ranging from 3 to 28, with a total of 5461 observations. A full dentition will constitute 28 teeth, however, missing tooth is very common in any oral health studies, with the actual cause of missingness mostly unknown. Hence, in order to avoid unverifiable missing data assumptions, we did not resort to missing data analysis, and present only complete case analysis.

As part of explanatory analysis, we present the bivariate kernel density estimate of the PPD and CAL responses in Fig. 2 (left panel). The plot reveals significant (right) skewness for both responses. Also, the right panel in Fig. 2 indicates presence of possible outliers. Recent research (Zhao et al. 2018) confirmed possible non-linear relationship between oral health responses, and

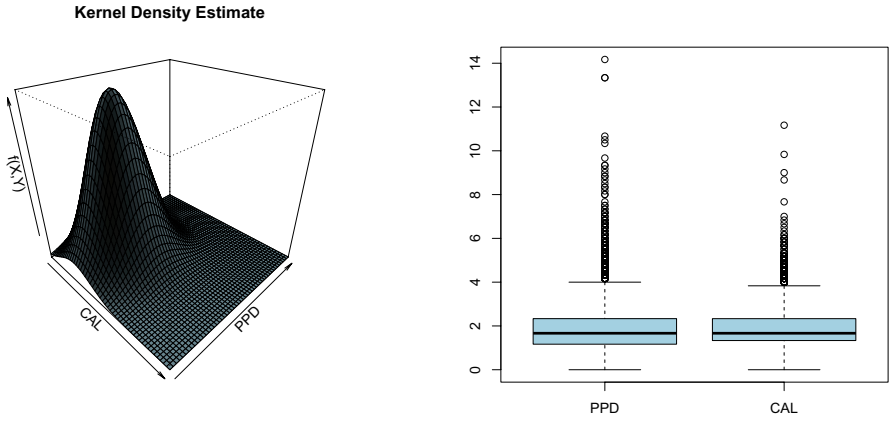

Fig. 2 Bivariate kernel density estimate (left panel) and boxplots (right panel) for PPD and CAL responses, from the GAAD data

Table 4 Estimates of the index parameters, the skewness parameter and their 95% confidence intervals, corresponding to the PPD and CAL responses from the GAAD study

Parameter (PPD)	Estimate	Confidence interval	Parameter (CAL)	Estimate	Confidence interval
β_{11}	0.7987	[0.7448, 0.8273]	β_{21}	0.6129	[0.4571, 0.6983]
β_{12}	0.5312	[0.4841, 0.5923]	β_{22}	0.7411	[0.6492, 0.8388]
β_{13}	−0.1219	[−0.1448, −0.1107]	β_{23}	0.0318	[0.0109, 0.0577]
β_{14}	0.1958	[0.1806, 0.2169]	β_{24}	0.1408	[0.0736, 0.2299]
β_{15}	0.1636	[0.1432, 0.1828]	β_{25}	0.2330	[0.1805, 0.3388]
γ_1	0.7977	[0.7037, 0.8844]	γ_2	0.6589	[0.5821, 0.7427]

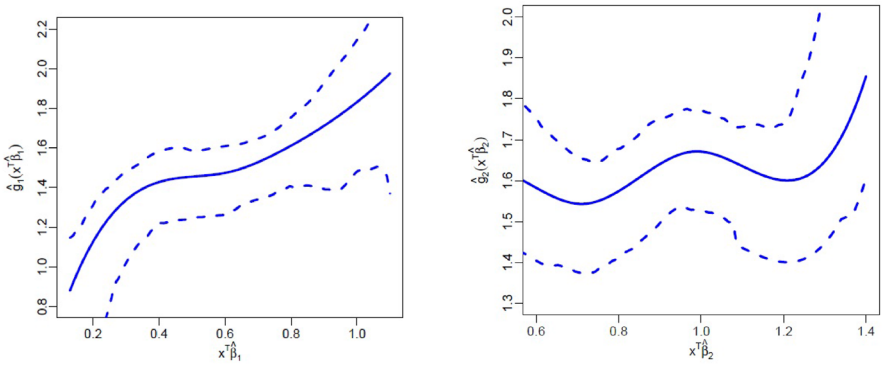

Fig. 3 Estimated curves for the two index functions $\widehat{g}_1$ and $\widehat{g}_2$, along with the 95% confidence bands. The left and right panels correspond to PPD and CAL regressions, respectively

continuous covariates, like Age. Motivated by this, we set forward to estimate a clinically meaningful single-index structure determining PD for the subjects in this database.

We consider fitting the following model to the GAAD data

$$\begin{cases} \text{PPD}_{ij} = g_1(\mathbf{x}_{ij}{}^{\text{T}}\boldsymbol{\beta}_1) + \mathbf{z}_{ij}{}^{\text{T}}\mathbf{b}_{i1} + \epsilon_{ij1}, \\ \text{CAL}_{ij} = g_2(\mathbf{x}_{ij}{}^{\text{T}}\boldsymbol{\beta}_2) + \mathbf{z}_{ij}{}^{\text{T}}\mathbf{b}_{i2} + \epsilon_{ij2}, \end{cases} \quad i = 1, \ldots, 288, j = 1, \ldots, m_i,$$

where $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ij5})^{\text{T}}$ with $x_{ij1} =$ Age, $x_{ij2} =$ BMI, $x_{ij3} =$ Gender, $x_{ij4} =$ Smoker, $x_{ij5} =$ HbA1c and $\mathbf{z}_{ij} = (1, z_{ij1}, z_{ij2}, z_{ij3})^{\text{T}}$ with $z_{ij1} =$ Gender, $z_{ij2} =$ Smoker, $z_{ij3} =$ HbA1c. We further assume $\mathbf{b}_i = (\mathbf{b}_{i1}{}^{\text{T}}, \mathbf{b}_{i2}{}^{\text{T}})^{\text{T}} \sim \text{SAL}_8(\mathbf{0}, \boldsymbol{\Omega}, \mathbf{0})$ and $\boldsymbol{\epsilon}_{ij} = (\epsilon_{ij1}, \epsilon_{ij2})^{\text{T}} \sim \text{SAL}_2(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\gamma})$. The estimates for index parameters, skewness parameter and their 95% confidence intervals are presented in Table 4, where the 95% confidence intervals are obtained by bootstrap resampling with 200 replications. We observe that all parameters (except β_{13} corresponding to Gender for the

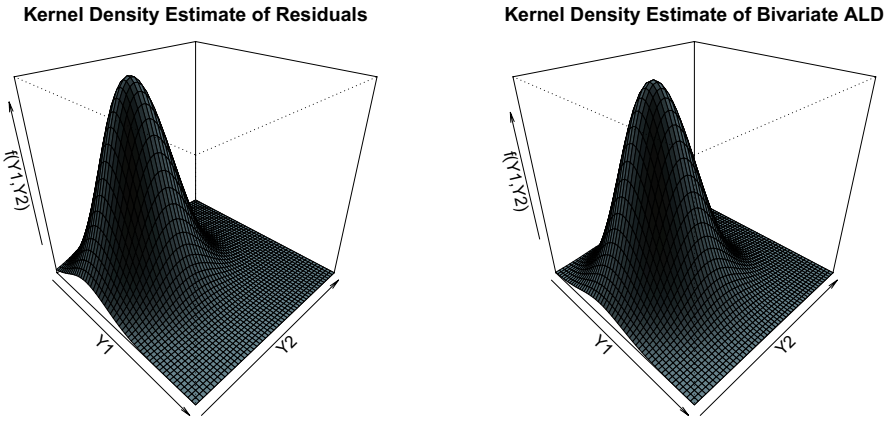

Fig. 4 Plots of bivariate kernel density estimates from model residuals (left panel), and from random draws of $n = 5461$ observations following ALD$(\widehat{\boldsymbol{\Sigma}}, \widehat{\boldsymbol{\gamma}})$

PPD regression) were positive and significant. Interestingly, the estimate of Gender (β_{13}) is negative yet significant for PPD, while, the corresponding estimate (β_{23}) for CAL is positive and significant, implying that Gender is contributing to the index development for the two responses in opposite directions. Figure 3 presents the estimated curves corresponding to the two index functions, along with their 95% confidence bands using bootstrap method. Compared to the CAL, the 95% band is tighter for the PPD.

It is immediate that the correlation between PPD and CAL are significant, implying the need to account for the crosswise correlation between the two responses, and the cluster-wise correlation of the responses within the same subject, while modeling the bivariate clustered responses. Furthermore, Fig. 4 presents the bivariate kernel density surface of the estimated residuals (left panel), and the same from random draws of $n = 5461$ observations from the bivariate ALD density ALD$(\widehat{\boldsymbol{\Sigma}}, \widehat{\boldsymbol{\gamma}})$, where $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\gamma}}$ are plugged-in estimates derived from our fit. We observe that the estimated surfaces are very similar, confirming the adequacy of model fit to the GAAD dataset.

Correlation matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ are estimates as:

$$\widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} 1.2429 & 0.7937 \\ 0.7937 & 0.9024 \end{pmatrix}$$

and

$$\widehat{\boldsymbol{\Omega}} = \begin{pmatrix} 1.6589 & -0.0089 & -0.0461 & -0.2792 & 1.5780 & -0.1832 & -0.1815 & -0.5760 \\ -0.0089 & 0.8797 & -0.4081 & 0.1553 & -0.1685 & 0.5379 & -0.0466 & 0.4289 \\ -0.0461 & -0.4081 & 0.8423 & 0.3273 & -0.0808 & 0.1296 & 0.1931 & 0.1264 \\ -0.2792 & 0.1553 & 0.3273 & 0.7782 & -0.4164 & 0.3802 & 0.1290 & 0.6585 \\ 1.5780 & -0.1685 & -0.0808 & -0.4164 & 2.1987 & -0.8975 & -0.4840 & -0.8462 \\ -0.1832 & 0.5379 & 0.1296 & 0.3802 & -0.8975 & 1.0517 & 0.3364 & 0.6420 \\ -0.1815 & -0.0466 & 0.1931 & 0.1290 & -0.4840 & 0.3364 & 0.2016 & 0.1681 \\ -0.5760 & 0.4289 & 0.1264 & 0.6585 & -0.8462 & 0.6420 & 0.1681 & 0.8158 \end{pmatrix}.$$

To further evaluate the usefulness of our proposed new model, we consider the fitted and prediction errors in light of two alternatives, denoted as "AM1" (bivariate normal, mixed effects SIM) and "AM2" (bivariate, asymmetric Laplace SIM, without random effects). We randomly partition the data into training and testing sets, where the training data is used to fit the 3 models, and the test data to evaluate the prediction errors. Using varying sizes of training and testing data, the average absolute fitted errors (AAFE), and the average absolute prediction errors (AAPE) for the two responses, based on 200 random partitions, are reported in Table 5, where

$$\text{AAFE}_k = \frac{1}{\sum_{i=1}^{nb} m_i} \sum_{i=1}^{nb} \sum_{j=1}^{m_i} |y_{ijk} - \widehat{y}_{ijk}|$$

and

$$\text{AAPE}_k = \frac{1}{\sum_{i=1}^{n-nb} m_i} \sum_{i=1}^{n-nb} \sum_{j=1}^{m_i} |y_{ijk} - \widetilde{y}_{ijk}|,$$

for $k = 1$ and 2, with $\widehat{y}_{ijk}$, the fitted value based on training data, and $\widetilde{y}_{ijk}$, the predicted value based on the test data, and nb denote the number of subjects in the training data.

From Table 5, we observe that our model performs the best in terms of AAFE and AAPE, for various sizes of the training and testing set. More specifically, our proposed mixed-effects SIM model is superior to the bivariate asymmetric Laplace SIM (excluding random effects), implying the necessity to account for the within-subject correlation. Furthermore, our proposed model is also better than the SIM with the usual multivariate normal specification for the random effects, thereby providing evidence of the gain in accounting for data asymmetry during modeling.

Table 5 Average absolute fitted and prediction errors for our model and 2 competing models (AM1 and AM2), for the PPD and CAL responses in the GAAD data, based on 200 random partitions

Size			PPD response				CAL Response		
Training set	Test set		Our Model	AM1	AM2		Our Model	AM1	AM2
100	188	AAFE_1	0.8670	0.9046	0.9297	AAPE_1	0.8884	0.9339	0.9513
		AAFE_2	0.6982	0.7005	0.7236	AAPE_2	0.7159	0.7164	0.7364
150	138	AAFE_1	0.8750	0.9209	0.9390	AAPE_1	0.8813	0.9335	0.9509
		AAFE_2	0.7024	0.7054	0.7274	AAPE_2	0.7091	0.7138	0.7346
200	88	AAFE_1	0.8718	0.9237	0.9406	AAPE_1	0.8785	0.9317	0.9502
		AAFE_2	0.6976	0.7050	0.7259	AAPE_2	0.7126	0.7172	0.7380
250	38	AAFE_1	0.8718	0.9265	0.9442	AAPE_1	0.8633	0.9185	0.9408
		AAFE_2	0.6988	0.7095	0.7309	AAPE_2	0.6994	0.7082	0.7299

7 Conclusions

Derivation of useful medical indices that correlate with multiple health outcomes is an issue of significant practical importance. In this paper, we propose a single-index mixed-effects regression model for bivariate responses, where both the error term and random effect are assumed to follow multivariate asymmetric Laplace distribution. By the polynomial spline smoothing for index functions, we proposed a scalable ML estimation method based on EM-type algorithm, and study the asymptotic properties of the ML estimates under some mild conditions. Simulations and real data analysis reveal the potential of the proposed model under data asymmetry, compared to existing alternatives.

There exists a number of future directions to pursue. To further improve model fit and prediction, we can consider the joint modeling of the location, skewness, and scatter matrix, within a multivariate ALD setup. When the number of covaiates is large in both fixed effects and random effects, it is of interest to select important variables in both parts to obtain a concise model. Some existing variable selection work of linear mixed effects model are available for univariate response case; see, for example, Kinney and Dunson (2010); Bondell et al. (2010); Fan and Li (2012); Schelldorfer and Geer (2011); Pan and Huang (2014), and others. However, for the case of single-index mixed effects models for multivariate responses, there is limited work, and pursuing the variable selection is a non-trivial journey. Another extension is to consider mixed effects quantile regression (Waldmann and Kneib 2015) for bivariate responses. These will be pursued elsewhere.

Appendix

Appendix 1: Lemmas

Lemma 1 *Assume that* $\mathbf{A}$ *is a* $d \times d$ *positive definite matrix, then for any positive definite matrix* $\boldsymbol{\Sigma}$ *with dimension* $d \times d$*, we have*

$$f(\boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-n/2} \exp\left\{-\frac{1}{2}\text{trace}(\boldsymbol{\Sigma}^{-1}\mathbf{A})\right\} \leq \left|\frac{1}{n}\mathbf{A}\right|^{-n/2} \exp\left\{-\frac{nd}{2}\right\},$$

if and only if $\boldsymbol{\Sigma} = \frac{1}{n}\mathbf{A}$.

Proof of Lemma 1 See proofs in Anderson (1984). □

According to the MLE defined in (12), the likelihood estimating equations for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ can be written as

$$\sum_{i=1}^{n} \begin{pmatrix} \mathbf{J}^{\mathrm{T}}\mathbf{X}_i \text{diag}\{\dot{\mathbf{W}}_i^{\mathrm{T}}(\mathbf{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\theta}}\} \\ \mathbf{W}_i(\mathbf{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}) \end{pmatrix} \frac{\partial \ell_m(\boldsymbol{\mu}_i, \widehat{\boldsymbol{\delta}}, \mathbf{y}_i)}{\partial \boldsymbol{\mu}_i}\bigg|_{\boldsymbol{\mu}_i = \mathbf{W}_i^{\mathrm{T}}(\mathbf{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\theta}}} = \mathbf{0}.$$

Denote $\dot{\ell}_m(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\delta},\mathbf{y}_i) \triangleq \frac{\partial \ell_m(\boldsymbol{\mu}_i,\boldsymbol{\delta},\mathbf{y}_i)}{\partial \boldsymbol{\mu}_i}\Big|_{\boldsymbol{\mu}_i=\mathbf{W}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}}$ and $\dot{\ell}_m(\mathbf{g}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}),\boldsymbol{\delta},\mathbf{y}_i) \triangleq \frac{\ell_m(\boldsymbol{\mu}_i,\boldsymbol{\delta},\mathbf{y}_i)}{\partial \boldsymbol{\mu}_i}\Big|_{\boldsymbol{\mu}_i=\mathbf{g}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta})}$. Then we have the following Lemma 2.

Lemma 2 *Assuming Conditions* (*A*1)–(*A*6) *hold*, *we have*

$$\left\|\sum_{i=1}^{n}\left[\begin{pmatrix}\mathbf{J}^{\mathrm{T}}\mathbf{X}_i\mathrm{diag}\{\dot{\mathbf{W}}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}\}\\ \mathbf{W}_i(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta})\end{pmatrix}-\begin{pmatrix}\mathbf{J}^{\mathrm{T}}\mathbf{X}_i\mathrm{diag}\{\dot{\mathbf{W}}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_0\}\\ \mathbf{W}_i(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\end{pmatrix}\right]\right\|\dot{\ell}_m(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\delta},\mathbf{y}_i) = o_p(\sqrt{n}) \tag{18}$$

and

$$\left\|\sum_{i=1}^{n}\left[\begin{pmatrix}\mathbf{J}^{\mathrm{T}}\mathbf{X}_i\mathrm{diag}\{\dot{\mathbf{W}}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}\}\\ \mathbf{W}_i(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta})\end{pmatrix}-\begin{pmatrix}\mathbf{J}^{\mathrm{T}}\mathbf{X}_i\mathrm{diag}\{\dot{\mathbf{W}}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_0\}\\ \mathbf{W}_i(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\end{pmatrix}\right]\right\|\dot{\ell}_m(\mathbf{g}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0),\boldsymbol{\delta}_0,\mathbf{y}_i) = o_p(\sqrt{n}) \tag{19}$$

uniformly over $\|\boldsymbol{\beta}^{(-1)}-\boldsymbol{\beta}_0^{(-1)}\|+\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|+\|\boldsymbol{\delta}-\boldsymbol{\delta}_0\|\le Cr_n$.

Proof of Lemma 2 We firstly prove (19). To obtain the bound, we only need to calculate the conditional variance of the left term in (19) since the conditional expection $\mathbb{E}\Big(\dot{\ell}_m(\mathbf{g}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0),\boldsymbol{\delta}_0,\mathbf{y}_i)\Big|\mathbf{X}_i,\mathbf{Z}_i\Big)=\mathbf{0}$. By the Condition (A4), the eigenvalues of the conditional variance for $\mathbb{V}\mathrm{ar}\Big(\dot{\ell}_m(\mathbf{g}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0),\boldsymbol{\delta}_0,\mathbf{y}_i)\Big|\mathbf{X}_i,\mathbf{Z}_i\Big)$ are bounded, hence we only need to obtain the bound of $\left\|\begin{matrix}\mathbf{J}^{\mathrm{T}}\mathbf{X}_i\mathrm{diag}\{\dot{\mathbf{W}}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}-\dot{\mathbf{W}}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_0\}\\ \mathbf{W}_i(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta})-\mathbf{W}_i(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\end{matrix}\right\|$.

By the properties of spline basis, we have

$$\begin{aligned}|\mathbf{W}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}-\mathbf{W}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_0| &\le |\dot{\mathbf{W}}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}^{*})\boldsymbol{\theta}\mathbf{X}_i{}^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)|+|\mathbf{W}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta})(\boldsymbol{\theta}-\boldsymbol{\theta}_0)|\\ &\le CK^{1/2}(\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|+\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|)\end{aligned}$$

and

$$\begin{aligned}|\dot{\mathbf{W}}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}-\dot{\mathbf{W}}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_0| &\le |\ddot{\mathbf{W}}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}^{**})\boldsymbol{\theta}\mathbf{X}_i{}^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)|+|\dot{\mathbf{W}}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta})(\boldsymbol{\theta}-\boldsymbol{\theta}_0)|\\ &\le CK^{3/2}(\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|+\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|),\end{aligned}$$

where both $\boldsymbol{\beta}^{*}$ and $\boldsymbol{\beta}^{**}$ lies on the line segment connecting $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$. As a result,

$$\left\|\begin{matrix}\mathbf{J}^{\mathrm{T}}\mathbf{X}_i\mathrm{diag}\{\dot{\mathbf{W}}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}-\dot{\mathbf{W}}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_0\}\\ \mathbf{W}_i(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta})-\mathbf{W}_i(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\end{matrix}\right\| \le CK^{3/2}r_n. \tag{20}$$

Then the order of (19) is $O_p(\sqrt{n}K^{3/2}r_n)=o_p(\sqrt{n})$ since $d>2$.

We next prove (18). By the Taylor's expansion and regularity conditions, it is clear that

$$\left\|\dot{\ell}_m(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\delta},\mathbf{y}_i)-\dot{\ell}_m(\mathbf{g}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0),\boldsymbol{\delta}_0,\mathbf{y}_i)\right\|=O_p(r_n).$$

Applying the results of (19) and (20), the order of (18) is $o_p(\sqrt{n})+O_p(nK^{3/2}r_n^2)=o_p(\sqrt{n})$. □

Lemma 3 *Assume that Condition* (*A*1)–(*A*6) *hold, the singular values of the matrix*

$$\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}\mathbf{J}^{\mathrm{T}}\mathbf{X}_i\mathrm{diag}\{\dot{\mathbf{W}}_i^{\mathrm{T}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_0\}\\ \mathbf{W}_i(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\end{pmatrix}\mathbf{C}_i^0\begin{pmatrix}\mathbf{J}^{\mathrm{T}}\mathbf{X}_i\mathrm{diag}\{\dot{\mathbf{W}}_i^{\mathrm{T}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_0\}\\ \mathbf{W}_i(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\end{pmatrix}^{\mathrm{T}}$$

are bounded and bounded away from zero with probability approaching one.

Proof of Lemma 3 Note that we can replace $\dot{\mathbf{W}}_i^{\mathrm{T}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_0$ with $\dot{\mathbf{g}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)$ in above expression with only a difference of $o_p(1)$ since $\|\dot{\mathbf{W}}_i^{\mathrm{T}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_0-\dot{\mathbf{g}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\|\le CK^{-s+1}$. Therefore we next only need to show that the eigenvalues of

$$\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}\mathbf{J}^{\mathrm{T}}\mathbf{X}_i\mathrm{diag}\{\dot{\mathbf{g}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\}\\ \mathbf{W}_i(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\end{pmatrix}\mathbf{C}_i^0\begin{pmatrix}\mathbf{J}^{\mathrm{T}}\mathbf{X}_i\mathrm{diag}\{\dot{\mathbf{g}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\}\\ \mathbf{W}_i(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\end{pmatrix}^{\mathrm{T}}$$

are bounded and bounded away from zero.

By the Condition (A6), there exists a $(p_1+p_2)\times(K_1+K_2)$ matrix $\boldsymbol{\Pi}_0$ such that

$$\|\mathbb{E}_{\mathcal{G}}[\mathbf{X}_i\mathrm{diag}\{\dot{\mathbf{g}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\}]-\boldsymbol{\Pi}_0\dot{\mathbf{W}}_i(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\|\le CK^{-s'}.$$

It is obvious that the singular values of $\begin{pmatrix}\mathbf{I} & -\boldsymbol{\Pi}_0\\ \mathbf{0} & \mathbf{I}\end{pmatrix}$ are bounded and bounded away from zero. Thus, by pre-/post-multiplying this matrix, we only need to prove that the singular values of

$$\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}\mathbf{J}^{\mathrm{T}}\mathbf{X}_i\mathrm{diag}\{\dot{\mathbf{g}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\}-\mathbf{J}^{\mathrm{T}}\boldsymbol{\Pi}_0\dot{\mathbf{W}}_i(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\\ \mathbf{W}_i(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\end{pmatrix}\mathbf{C}_i^0$$
$$\times\begin{pmatrix}\mathbf{J}^{\mathrm{T}}\mathbf{X}_i\mathrm{diag}\{\dot{\mathbf{g}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\}-\mathbf{J}^{\mathrm{T}}\boldsymbol{\Pi}_0\dot{\mathbf{W}}_i(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\\ \mathbf{W}_i(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\end{pmatrix}^{\mathrm{T}}$$

are bounded and bounded away from zero. Apply the approximation of splines again, we only need to show that the singular values of

$$\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}\mathbf{J}^{\mathrm{T}}\mathbf{X}_i\mathrm{diag}\{\dot{\mathbf{g}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\}-\mathbf{J}^{\mathrm{T}}\mathbb{E}_{\mathcal{G}}[\mathbf{X}_i\mathrm{diag}\{\dot{\mathbf{g}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\}]\\ \mathbf{W}_i(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\end{pmatrix}\mathbf{C}_i^0$$
$$\times\begin{pmatrix}\mathbf{J}^{\mathrm{T}}\mathbf{X}_i\mathrm{diag}\{\dot{\mathbf{g}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\}-\mathbf{J}^{\mathrm{T}}\mathbb{E}_{\mathcal{G}}[\mathbf{X}_i\mathrm{diag}\{\dot{\mathbf{g}}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\}]\\ \mathbf{W}_i(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\end{pmatrix}^{\mathrm{T}}$$

are bounded, and bounded away from zero. By the law of large numbers, we only need to show its expectation has eigenvalues bounded and bounded away from zero. This is true by checking Conditions (A5) and (A6). □

Proof of Theorem 1 By Lemma 2 and Taylor's expansion, it is easy to show

$$\begin{aligned}
&\sum_{i=1}^{n}\begin{pmatrix}\mathbf{J}^{\mathrm{T}}\mathbf{X}_i\mathrm{diag}\{\dot{\mathbf{W}}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_0\}\\ \mathbf{W}_i(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\end{pmatrix}\dot{\ell}_m(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\delta},\mathbf{y}_i)\\
=\ &\sum_{i=1}^{n}\begin{pmatrix}\mathbf{J}^{\mathrm{T}}\mathbf{X}_i\mathrm{diag}\{\dot{\mathbf{W}}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_0\}\\ \mathbf{W}_i(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\end{pmatrix}\dot{\ell}_m(\mathbf{g}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0),\boldsymbol{\delta}_0,\mathbf{y}_i)\\
&-\sum_{i=1}^{n}\begin{pmatrix}\mathbf{J}^{\mathrm{T}}\mathbf{X}_i\mathrm{diag}\{\dot{\mathbf{W}}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_0\}\\ \mathbf{W}_i(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\end{pmatrix}\mathbf{C}_i^0\begin{pmatrix}\mathbf{J}^{\mathrm{T}}\mathbf{X}_i\mathrm{diag}\{\dot{\mathbf{W}}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_0\}\\ \mathbf{W}_i(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}\boldsymbol{\beta}^{(-1)}-\boldsymbol{\beta}^{(-1)}\\ \boldsymbol{\theta}-\boldsymbol{\theta}_0\end{pmatrix}\\
&\qquad\qquad+o_p(\sqrt{n})+O_p(nr_n)
\end{aligned}\tag{21}$$

if $\|\boldsymbol{\beta}^{(-1)}-\boldsymbol{\beta}_0^{(-1)}\|+\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|+\|\boldsymbol{\delta}-\boldsymbol{\delta}_0\|=O_p(r_n)$.

By direct variance calculation

$$\sum_{i=1}^{n}\begin{pmatrix}\mathbf{J}^{\mathrm{T}}\mathbf{X}_i\mathrm{diag}\{\dot{\mathbf{W}}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_0\}\\ \mathbf{W}_i(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\end{pmatrix}\dot{\ell}_m(\mathbf{g}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0),\boldsymbol{\delta}_0,\mathbf{y}_i)=O_p(\sqrt{nK}).\tag{22}$$

Moreover, by Lemma 3, the singular values of the matrix

$$\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}\mathbf{J}^{\mathrm{T}}\mathbf{X}_i\mathrm{diag}\{\dot{\mathbf{W}}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_0\}\\ \mathbf{W}_i(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\end{pmatrix}\mathbf{C}_i^0\begin{pmatrix}\mathbf{J}^{\mathrm{T}}\mathbf{X}_i\mathrm{diag}\{\dot{\mathbf{W}}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_0\}\\ \mathbf{W}_i(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\end{pmatrix}^{\mathrm{T}}$$

are bounded and bounded away from zero with probability approaching one.

Combining the above results of (21) and (22) together with Lemma 2, if choosing L sufficiently large enough for $\|\boldsymbol{\beta}^{(-1)}-\boldsymbol{\beta}_0^{(-1)}\|+\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|+\|\boldsymbol{\delta}-\boldsymbol{\delta}_0\|=Lr_n$, we have

$$\begin{aligned}
&P\left(\left\|\sum_{i=1}^{n}\begin{pmatrix}\mathbf{J}^{\mathrm{T}}\mathbf{X}_i\mathrm{diag}\{\dot{\mathbf{W}}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta}\}\\ \mathbf{W}_i(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta})\end{pmatrix}\dot{\ell}_m(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\delta},\mathbf{y}_i)\right\|\right.\\
&\qquad\left.>\left\|\sum_{i=1}^{n}\begin{pmatrix}\mathbf{J}^{\mathrm{T}}\mathbf{X}_i\mathrm{diag}\{\dot{\mathbf{W}}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_0\}\\ \mathbf{W}_i(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\end{pmatrix}\dot{\ell}_m(\boldsymbol{\beta}_0,\boldsymbol{\theta}_0,\boldsymbol{\delta}_0,\mathbf{y}_i)\right\|\right)\to 1.
\end{aligned}$$

Thus we can conclude that $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|+\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|=O_p(r_n)$. □

Proof of Theorem 2 Denote $\mathbf{T}=(\mathbf{T}_1,\ldots,\mathbf{T}_n)^{\mathrm{T}}$, $\mathbf{D}=\mathrm{diag}(\mathbf{C}_1^0,\ldots,\mathbf{C}_n^0)$ and define the "projection matrix" $\mathbf{P}=\mathbf{T}(\mathbf{T}^{\mathrm{T}}\mathbf{D}\mathbf{T})^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{D}$, where $\mathbf{T}_i=\mathbf{W}_i(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)$. Let $\mathbf{X}_i^*=\mathbf{J}^{\mathrm{T}}\mathbf{X}_i\mathrm{diag}\{\dot{\mathbf{W}}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\theta}_0\}$, $\mathbf{X}^*=(\mathbf{X}_1^*,\ldots,\mathbf{X}_n^*)^{\mathrm{T}}$ and $\widetilde{\mathbf{X}}^*=(\mathbf{I}-\mathbf{P})\mathbf{X}^*$. Then we can write $\widetilde{\mathbf{X}}_i^*=\mathbf{X}_i^*-\mathbf{A}\mathbf{T}_i$ where $\mathbf{A}=\mathbf{X}^{*\mathrm{T}}\mathbf{D}\mathbf{T}(\mathbf{T}^{\mathrm{T}}\mathbf{D}\mathbf{T})^{-1}$. Let $\widetilde{\mathbf{A}}=\begin{pmatrix}\mathbf{I}&\mathbf{A}\\ \mathbf{0}&\mathbf{I}\end{pmatrix}$. By the Lemma 2 and the proof shown in Theorem 1, we have

$$\sup_{\|\boldsymbol{\beta}^{(-1)}-\boldsymbol{\beta}_0^{(-1)}\|+\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|+\|\boldsymbol{\delta}-\boldsymbol{\delta}_0\|\le Cr_n} \Bigg\| \widetilde{\mathbf{A}} \sum_{i=1}^{n} \begin{pmatrix} \widetilde{\mathbf{X}}_i^* \\ \mathbf{T}_i \end{pmatrix} \dot{\ell}_m(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\delta}, \mathbf{y}_i)$$
$$- \widetilde{\mathbf{A}} \sum_{i=1}^{n} \begin{pmatrix} \widetilde{\mathbf{X}}_i^* \\ \mathbf{T}_i \end{pmatrix} \dot{\ell}_m(\mathbf{g}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0), \boldsymbol{\delta}_0, \mathbf{y}_i)$$
$$+ \widetilde{\mathbf{A}} \sum_{i=1}^{n} \begin{pmatrix} \widetilde{\mathbf{X}}_i^* \\ \mathbf{T}_i \end{pmatrix} \mathbf{C}_i^0 \left[(\widetilde{\mathbf{X}}_i^{*\mathrm{T}}, \mathbf{T}_i{}^{\mathrm{T}}) \begin{pmatrix} \boldsymbol{\beta}^{(-1)} - \boldsymbol{\beta}_0^{(-1)} \\ \boldsymbol{\theta} - \boldsymbol{\theta}_0 + \mathbf{A}^{\mathrm{T}}(\boldsymbol{\beta}^{(-1)} - \boldsymbol{\beta}_0^{(-1)}) \end{pmatrix} + \mathbf{R}_i \right] \Bigg\|$$
$$= o_p(\sqrt{n}),$$

where $\mathbf{R}_i = \mathbf{W}_i{}^{\mathrm{T}}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\theta} - \mathbf{g}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0)$.

By parameter transformation, we can write $\boldsymbol{\theta} - \boldsymbol{\theta}_0 + \mathbf{A}^{\mathrm{T}}(\boldsymbol{\beta}^{(-1)} - \boldsymbol{\beta}_0^{(-1)})$ as $\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0$. Further denote

$$\mathbf{U}(\boldsymbol{\beta}, \boldsymbol{\vartheta}) \triangleq \widetilde{\mathbf{A}} \sum_{i=1}^{n} \begin{pmatrix} \widetilde{\mathbf{X}}_i^* \\ \mathbf{T}_i \end{pmatrix} \dot{\ell}_m(\mathbf{g}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0), \boldsymbol{\delta}_0, \mathbf{y}_i)$$
$$- \sum_{i=1}^{n} \begin{pmatrix} \widetilde{\mathbf{X}}_i^* \\ \mathbf{T}_i \end{pmatrix} \mathbf{C}_i^0 \left[(\widetilde{\mathbf{X}}_i^{*\mathrm{T}}, \mathbf{T}_i{}^{\mathrm{T}}) \begin{pmatrix} \boldsymbol{\beta}^{(-1)} - \boldsymbol{\beta}_0^{(-1)} \\ \boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0 \end{pmatrix} + \mathbf{R}_i \right],$$

and the first $p_1 + p_2 - 2$ and the last $K_1 + K_2$ equations of $\mathbf{U}(\boldsymbol{\beta}, \boldsymbol{\vartheta})$ as $\mathbf{U}_1(\boldsymbol{\beta}, \boldsymbol{\vartheta})$ and $\mathbf{U}_2(\boldsymbol{\beta}, \boldsymbol{\vartheta})$, respectively. Let

$$\widetilde{\boldsymbol{\beta}}^{(-1)} = \boldsymbol{\beta}_0^{(-1)} + \left(\sum_{i=1}^{n} \widetilde{\mathbf{X}}_i^* \mathbf{C}_i^0 \widetilde{\mathbf{X}}_i^{*\mathrm{T}} \right)^{-1} \sum_{i=1}^{n} \widetilde{\mathbf{X}}_i^* \dot{\ell}_m(\mathbf{g}(\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{\beta}_0), \boldsymbol{\delta}_0, \mathbf{y}_i).$$

It is easy to see that

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \widetilde{\mathbf{X}}_i^* \mathbf{C}_i^0 \widetilde{\mathbf{X}}_i^{*\mathrm{T}} - \boldsymbol{\Psi} \right\| = o_p(1),$$

and by the central limit theorem, we have

$$\sqrt{n}(\widetilde{\boldsymbol{\beta}}^{(-1)} - \boldsymbol{\beta}_0^{(-1)}) \xrightarrow{\mathrm{d}} N(\mathbf{0}, \boldsymbol{\Psi}_1^{-1}).$$

In the following, we only need to show $\|\widehat{\boldsymbol{\beta}}^{(-1)} - \widetilde{\boldsymbol{\beta}}_0^{(-1)}\| = o_p(1/\sqrt{n})$.

For any $\boldsymbol{\beta}$ satisfying $\|\boldsymbol{\beta}^{(-1)} - \boldsymbol{\beta}_0^{(-1)}\| = \varepsilon/\sqrt{n}, \forall \varepsilon > 0$, similar to the proof of Lemma A.6 in Zhao et al. (2017), we can show that

$$\| \sum_{i=1}^{n} \widetilde{\mathbf{X}}_i^* \mathbf{C}_i^0 \mathbf{R}_i \| = o_p(\sqrt{n}) \text{ and } \| \sum_{i=1}^{n} \mathbf{T}_i \mathbf{C}_i^0 \mathbf{R}_i \| = o_p(n),$$

which lead to

$$\|\mathbf{U}_2(\boldsymbol{\beta}, \widehat{\boldsymbol{\vartheta}}) - \mathbf{U}_2(\widetilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\vartheta}})\| = o_p(\sqrt{n}).$$

Furthermore, note that

$$\sum_{i=1}^{n} \widetilde{\mathbf{X}}_i^* \mathbf{C}_i^0 \mathbf{T}_i{}^{\mathrm{T}} = \sum_{i=1}^{n} (\mathbf{X}_i^* - \mathbf{X}^{*\mathrm{T}} \mathbf{D}\mathbf{T}(\mathbf{T}^{\mathrm{T}}\mathbf{D}\mathbf{T})^{-1}\mathbf{T}_i)\mathbf{D}_i \mathbf{T}_i{}^{\mathrm{T}} = \mathbf{0},$$

and $\mathbf{U}_1(\boldsymbol{\beta}, \boldsymbol{\vartheta})$ is a linear function of $\boldsymbol{\beta}$ up to a $o_p(\sqrt{n})$ term. Consequently,

$$\|\mathbf{U}_1(\boldsymbol{\beta}, \widehat{\boldsymbol{\vartheta}})\| \geq Cn\|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\| + o_p(\sqrt{n}).$$

As a result, we have

$$\|\widetilde{\mathbf{A}}\mathbf{U}(\boldsymbol{\beta}, \widehat{\boldsymbol{\vartheta}})\| \geq C\varepsilon\sqrt{n} \text{ while } \|\widetilde{\mathbf{A}}\mathbf{U}(\widetilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\vartheta}})\| = o_p(\sqrt{n})$$

since the eigenvalues of $\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^{\mathrm{T}}$ are bounded and bounded away from zero with probability approaching 1. Then we can conclude that $\|\widehat{\boldsymbol{\beta}}^{(-1)} - \widetilde{\boldsymbol{\beta}}_0^{(-1)}\| = o_p(1/\sqrt{n})$ holds. □

Acknowledgements The authors thank the Center for Oral Health Research at the Medical University of South Carolina for providing the motivation, and the context of this work.

Funding The work is partially funded by grants grants R21DE031879 and R01DE031134 awarded by the United States National Institutes of Health.

Declarations

Conflict of interest The authors report no conflict of interest.

References

Anderson TW (1984) An introduction to multivariate statistical analysis. John Wiley & Sons, USA

Azzalini A (2010) The skew-normal distribution and related multivariate families. Scand J Stat 32:159–188

Azzalini A, Capitanio A (1999) Statistical applications of the multivariate skew normal distribution. J Roy Stat Soc 61:579–602

Azzalini A, Capitanio A (2003) Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. J R Stat Soc Series B (Stat Methodol) 65:367–389

Bandyopadhyay D, Lachos VH, Abanto-Valle CA, Ghosh P (2010) Linear mixed models for skew-normal/independent bivariate responses with an application to periodontal disease. Stat Med 29:2643–2655

Bondell HD, Krishna A, Ghosh SK (2010) Joint variable selection for fixed and random effects in linear mixed-effects models. Biometrics 66:1069–1077
Bouveyron C, Brunet-Saumard C (2014) Model-based clustering of high-dimensional data: a review. Comput Stat Data Anal 71:52–78
Cui X, Haerdle WK, Zhu L (2011) The EFM approach for single-index models. Ann Stat 39:1658–1688
De Boor C (2001) A practical guide to splines, 4th edn. Applied Mathematical Sciences. Springer-Verlag, Berlin
Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). J R Stat Soc Series B (Stat Methodol) 39:1–38
Eltoft T, Kim T, Lee TW (2006) On the multivariate Laplace distribution. IEEE Signal Process Lett 13:300–303
Fan JQ, Peng H (2004) Nonconcave penalized likelihood with a diverging number of parameters. Ann Stat 32:928–961
Fan Y, Li R (2012) Variable selection in linear mixed effects models. Ann Stat 40:2043–2068
Franczak BC, Browne RP, Mcnicholas PD (2014) Mixtures of shifted asymmetric Laplace distributions. IEEE Trans Pattern Anal Mach Intell 36:1149–1157
Gupta A (2003) Multivariate skew t-distribution. Stat A J Theor Appl Stat 37:359–363
Gupta AK, González-FaríAs G, DomáNguez-Molina JA (2004) A multivariate skew normal distribution. J Multivar Anal 89:181–190
Hardle W, Hall P, Ichimura H (1993) Optimal smoothing in single-index models. Ann Stat 21:157–178
Jara A, Quintana F, San Martín E (2008) Linear mixed models with skew-elliptical distributions: a Bayesian approach. Comput Stat Data Anal 52:5033–5045
Kinney SK, Dunson DB (2010) Fixed and random effects selection in linear and logistic models. Biometrics 63:690–698
Kotz S, Kozubowski TJ, Podgórski K (2001) The Laplace distribution and generalizations. Birkhauser, Switzerland
Kozubowski TJ, Podgórski K (2001) Asymmetric Laplace laws and modeling financial data. Math Comput Modell 34:1003–1021
Li Q (2000) Efficient estimation of additive partially linear models. Int Econ Rev 41:1073–1092
Lian H, Liang H (2013) Generalized additive partial linear models with high-dimensional covariates. Econometr Theor 29:1136–1161
Lin TI, Wang WL (2013) Multivariate skew-normal at linear mixed models for multi-outcome longitudinal data. Stat Model 13:199–221
Lu M (2017) Efficient estimation of quasi-likelihood models using b-splines. Ann Inst Stat Math 69:1099–1127
Luo S, Wang J (2014) Bayesian hierarchical model for multiple repeated measures and survival data: an application to Parkinson's disease. Stat Med 33:4279–4291
Ma S, Song PX-K (2015) Varying index coefficient models. J Am Stat Assoc 110:341–356
Meng X, Rubin DB (1993) Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika 80:267–278
Michaelis P, Klein N, Kneib T (2018) Bayesian multivariate distributional regression with skewed responses and skewed random effects. J Comput Graph Stat 27:602–611
Naik DN, Plungpongpun K (2006) A Kotz-type distribution for multivariate statistical inference. Birkhäuser Boston, Boston
Page RC, Eke PI (2007) Case definitions for use in population-based surveillance of periodontitis. J Periodontol 78:1387–1399
Pan J, Huang C (2014) Random effects selection in generalized linear mixed models via shrinkage penalty function. Stat Comput 24:725–738
Pinheiro JC, Liu C, Wu YN (2001) Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. J Comput Graph Stat 10:249–276
Schelldorfer J, Geer SVD (2011) Estimation for high-dimensional linear mixed-effects models using ℓ_1 -penalization. Scand J Stat 38:197–214
Verbeke G, Fieuws S, Molenberghs G, Davidian M (2014) The analysis of multivariate longitudinal data: a review. Stat Methods Med Res 23:42–59
Waldmann E, Kneib T (2015) Bayesian bivariate quantile regression. Stat Model 15:326–344
Wang L, Liu X, Liang H, Carroll RJ (2011) Estimation and variable selection for generalized additive partial linear models. Ann Stat 39:1827–1851

Wang L, Xue L, Qu A, Liang H et al (2014) Estimation and model selection in generalized additive partial linear models for correlated data with diverging number of covariates. Ann Stat 42:592–624
Wu J, Tu W (2016) A multivariate single-index model for longitudinal data. Stat Model 16:392–408
Yu Y, Ruppert D (2002) Penalized spline estimation for partially linear single-index models. J Am Stat Assoc 97:1042–1054
Zhao W, Lian H, Bandyopadhyay D (2018) A partially linear additive model for clustered proportion data. Stat Med 37:1009–1030
Zhao W, Lian H, Liang H (2017) GEE analysis for longitudinal single-index quantile regression. J Stat Plann Inference 187:78–102

Authors and Affiliations

Weihua Zhao[1] · Dipankar Bandyopadhyay[3] · Heng Lian[2]

✉ Dipankar Bandyopadhyay
dbandyop@vcu.edu

[1] School of Sciences, Nantong University, Nantong, China

[2] Department of Mathematics, City University of Hong Kong, Kowloon Tong, Hong Kong, China

[3] Department of Biostatistics, School of Population Health, Virginia Commonwealth University, Richmond, VA, USA