Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ORIGINAL ARTICLE

Generalized single index modeling of longitudinal data with multiple binary responses

Zibo Tian and Peihua Qiu

Department of Biostatistics, University of Florida, Florida, USA

Correspondence
Peihua Qiu
Department of Biostatistics
University of Florida
Gainesville FL, 32611, USA
Email: pqiu@ufl.edu

Abstract

In health and clinical research, medical indices (e.g., BMI) are commonly used for monitoring and/or predicting health outcomes of interest. While single-index modeling can be used to construct such indices, methods to use single-index models for analyzing longitudinal data with multiple correlated binary responses are underdeveloped, although there are abundant applications with such data (e.g., prediction of multiple medical conditions based on longitudinally observed disease risk factors). This paper aims to fill the gap by proposing a generalized single-index model that can incorporate multiple single indices and mixed effects for describing observed longitudinal data of multiple binary responses. Compared to the existing methods focusing on constructing marginal models for each response, the proposed method can make use of the correlation information in the observed data about different responses when estimating different single indices for predicting response variables. Estimation of the proposed model is achieved by using a local linear kernel smoothing procedure, together with methods designed specifically for estimating single-index models and traditional methods for estimating generalized linear mixed models. Numerical studies show that the proposed method is effective in various cases considered. It is also demonstrated using a dataset from the English Longitudinal Study of Aging project.

KEYWORDS:
Binary responses; EM algorithm; Local linear kernel smoothing; Mixed-effects modeling; Multiple responses; Single-index model.

1 | INTRODUCTION

A composite index or score can summarize a pool of explanatory variables that are potential predictors of the outcomes of interest. While the composite index may become a pragmatic tool to be used in subsequent studies, it can effectively inform

stakeholders by concealing the complexity of the original data. In health and clinical research, medical indices are commonly used for monitoring and predicting health outcomes of interest. For instance, the visceral adiposity index, a combination of anthropometric and laboratory parameters, can be used to predict the risk of type II diabetes.[1] In practice, many scientifically meaningful response variables are binary, indicating the status of some conditions (e.g., diseases) of interest. In this paper, we focus on constructing composite indices when there are multiple mutually correlated binary response variables and the observed data of these response variables and the related predictors are longitudinal.

Single-index modeling is a commonly used tool for constructing composite indices. This semiparametric modeling approach links the mean of a response variable to a linear combination of the predictors through an unknown nonparametric link function. While combining different predictors into a univariate index helps to circumvent the so-called "curse of dimensionality", the link function allows a more flexible relationship between the response variable and the predictors. In the literature, there is much existing discussion about single-index modeling in cases when there is a single continuous response variable and all observations are assumed to be independent. See, for instance, Härdle and Stoker[2], Ichimura[3], Härdle et al.[4], Xia et al.[5], Xia[6], and Yu and Ruppert[7]. When the response variable is binary, the logit of the mean response can be linked to the single index by an unknown link function, and the resulting model is often called a generalized single-index model in the literature. See related discussions in papers such as Carroll et al.[8] and Cui et al.[9]

In practice, longitudinal data at multiple time points are routinely collected for individual subjects. In addition, there could be multiple response variables of interest in some studies. For instance, we are often concerned about multiple diseases of individual patients in some medical studies, and their disease statuses and the related disease risk factors are usually observed longitudinally over time. To analyze such data, it would definitely be beneficial to model all response variables jointly in order to make use of the information about their association (cf., Diggle et al.[10]). The within-subject data correlation should be accommodated properly in the model as well. In the literature, there is some existing discussion about single-index modeling of longitudinal data, most of which is in cases when the response variables are continuous. For instance, Chen et al.[11] proposed a partially-linear single-index model to analyze longitudinal data with a single continuous response variable, in which a semiparametric generalized estimating equations (SGEE) approach was used to estimate the index coefficients, link function, and parameters in a working correlation matrix. Wu and Tu[12] extended the penalized spline method originally discussed in Yu and Ruppert[7] to cases with multiple continuous response variables and longitudinal observed data. Tian and Qiu[13] proposed a more flexible multivariate single-index model for cases with multiple continuous response variables, which can allow different index coefficients for different response variables and accommodate both the within-subject and between-response-variable correlation. Its model estimation was based on the local linear kernel smoothing procedure for estimating the nonparametric link functions (cf., Qiu[14]) and the expectation-maximization (EM) algorithm (cf., Dempster et al.[15]) for estimating the index coefficients and other parameters. There are a few papers in the literature discussing single-index modeling in cases when there is a single binary response variable. For instance,

Chowdhury and Sinha[16] proposed a partially linear single-index logistic regression model for analyzing longitudinal data with a binary response variable based on the second-order GEE approach.[17] A similar model was studied by Yi et al.[18], where the within-subject association in the observed data was described by marginal odds ratios and the model was estimated by the first-order GEE.

From the above description, it can be seen that there is little existing discussion in the literature about the single-index modeling problem described at the beginning of the paper when there are multiple binary response variables and the observed data are longitudinal. This paper aims to fill the gap by proposing a multivariate single-index longitudinal logistic regression model for analyzing such data that are commonly seen in practice. In the proposed model, the index coefficients can be different for different response variables, and some random-effects terms are used to accommodate both within-subject and between-response-variable data correlation. Estimation of the proposed model is based on the combination of the local linear kernel smoothing procedure, some special methods developed for estimating the single-index models, and some conventional methods for estimating the generalized linear mixed-effect models (GLMM). By integrating the ideas of the EM algorithm and the refined conditional minimum average variance estimation (rMAVE) method proposed by Xia et al.[5], its index coefficients, link functions, and variance components of the random effects can be estimated simultaneously. Both theoretical justifications and numerical studies show that the proposed method provides a powerful analytic tool for constructing composite indices when there are multiple binary response variables that are observed longitudinally.

The rest of the paper is organized as follows. Section 2 describes the proposed model and its estimation in detail, along with some statistical properties of the estimated model. Section 3 presents some numerical results about the finite-sample performance of the proposed method. The proposed method is applied to a dataset from the English Longitudinal Study of Aging (ELSA) in Section 4. Some concluding remarks are given in Section 5. Some technical details are provided in the Web Appendix.

2 | PROPOSED METHODOLOGY

The proposed method is described in several parts in this section. The proposed single-index model for describing the observed longitudinal data of multiple binary response variables is desbribed in Subsection 2.1. Estimation of the index coefficients and the link functions is discussed in Subsections 2.2 and 2.3, respectively.

2.1 | Model specification

In a longitudinal study with q binary response variables of interest, assume that there are a total of M subjects involved. For the ith subject, m_i repeated measurements are taken on both the q binary response variables and p predictors at times $\{t_{ij} \in [T_0, T_1], j = 1, \ldots, m_i\}$, where $[T_0, T_1]$ specifies the study period, and the longitudinal observations of the response variables

and predictors are denoted as $\{Y_{ijk}, j = 1, \ldots, m_i, k = 1, \ldots, q\}$ and $\{\mathbf{X}_{ij} = (X_{ij1}, \ldots, X_{ijp})^T, j = 1, \ldots, m_i\}$, respectively, for $i = 1, \ldots, M$. These observed data are assumed to follow the multivariate single-index longitudinal logistic regression model below:

$$\log\left(\frac{\mu_{ijk}}{1-\mu_{ijk}}\right) = \psi_k\left(\boldsymbol{\beta}_k^T\mathbf{X}_{ij}\right) + \mathbf{g}_k^T\mathbf{b}_i, \text{ for } j = 1, \ldots, m_i, i = 1, \ldots, M, k = 1, \ldots, q, \tag{1}$$

where $\mathbf{b}_i = (\mathbf{b}_{i1}^T, \ldots, \mathbf{b}_{iq}^T)^T$ is a vector of random effects for the ith subject, $\mathbf{g}_k$ is a design vector of the random effects, $\mu_{ijk} = \mathbb{E}(Y_{ijk}|\mathbf{X}_{ij}, \mathbf{b}_i)$, $\psi_k(\cdot)$ is an unknown link function, and $\boldsymbol{\beta}_k$ is a p-dimensional vector of index coefficients. In Model (1), $\{\mathbf{b}_i, i = 1, \ldots, M\}$ are assumed to be independently and identically distributed (i.i.d.) with the common distribution $N_q(\mathbf{0}, \boldsymbol{\Sigma}_b)$. For simplicity, it is further assumed that $\mathbf{b}_{ik} = b_{ik}$ is a scalar, for each k, and thus only the random intercept is considered for each response variable. In such cases, $\mathbf{g}_k$ is a q-dimensional vector whose kth element equals 1 and the remaining elements all equal 0. For model identifiability, the index coefficients are assumed to satisfy the conditions that $\boldsymbol{\beta}_k^T\boldsymbol{\beta}_k = 1$ and the first element of $\boldsymbol{\beta}_k$ is positive, for each k. If all link functions $\psi_k(\cdot)$'s are identity functions, then Model (1) reduces to a multivariate GLMM with a logit linkage function. It should be pointed out that inclusion of the random-effects in Model (1) is for accommodating both the within-subject data correlation and the correlation among different response variables. When only random intercepts are considered in the model, it is actually assumed that the log odds ratio can vary among different subjects given the covariate effect.

2.2 | Monte Carlo EM algorithm for estimating model parameters

In this subsection, we describe how to estimate $\{\boldsymbol{\beta}_k\}$ and $\boldsymbol{\Sigma}_b$ in Model (1) by using the Monte Carlo EM algorithm and an extended version of the rMAVE method. The Monte Carlo EM algorithm can be used for estimating mixed-effect models numerically. See, for instance, Laird and Ware[19], McCulloch[20], and Booth and Hobert[21]. Under the mixed-effects modeling framework, the random effects can be regarded as unobserved part of the response, and the log-likelihood for the complete data, denoted as $l_c(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{b})$, can be used for model estimation, where $\mathbf{Y}$ is the vector of the observed responses, $\mathbf{b} = (\mathbf{b}_1^T, \ldots, \mathbf{b}_M^T)^T$ is the vector of the random effects, and $\boldsymbol{\theta}$ is the vector of all unknown parameters in the model. Then, $\boldsymbol{\theta}$ can be estimated iteratively as follows. Let $\boldsymbol{\theta}^*$ be the parameter estimates obtained in the previous iteration. Then, they can be updated iteratively by maximizing $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \mathbb{E}_{\mathbf{b}|\mathbf{Y},\boldsymbol{\theta}^*}\{l_c(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{b})|\mathbf{Y}, \boldsymbol{\theta}^*\}$ until a convergence criterion is met.

For Model (1), let $\mathbf{Y}_{i\bullet k} = (Y_{i1k}, \ldots, Y_{im_ik})^T$ be the observed kth response variable of the ith subject, $\mathbf{Y}_{i\bullet\bullet} = (\mathbf{Y}_{i\bullet 1}^T, \ldots, \mathbf{Y}_{i\bullet q}^T)^T$ be the vector of observations of all response variables of the ith subject, $\mathbf{Y}_{\bullet\bullet k} = (\mathbf{Y}_{1\bullet k}^T, \ldots, \mathbf{Y}_{M\bullet k}^T)^T$ be the vector of observations of the kth response variable for all subjects, and $\mathbf{Y} = (\mathbf{Y}_{\bullet\bullet 1}^T, \ldots, \mathbf{Y}_{\bullet\bullet q}^T)^T$ be the vector of all observed response variables. Let $\mathbf{X}$ be the $(\sum_{i=1}^M m_i) \times p$ matrix of all covariate data where the jth row is $\mathbf{X}_{1j}^T$ for $1 \le j \le m_1$ and the $[\sum_{s=1}^{i-1}(s-1)m_s + j]$-th row is $\mathbf{X}_{ij}^T$, for $1 \le j \le m_i$ and $2 \le i \le M$, and $\mathbf{X}_i$ be the $m_i \times p$ matrix whose jth row is $\mathbf{X}_{ij}^T$. Let $\boldsymbol{\theta}$ denote a collection of all unknown

parameters $(\{\boldsymbol{\beta}_k\}, \boldsymbol{\Sigma}_b)$ as well as the unknown link functions $\{\psi_k(\cdot)\}$ in Model (1). Then, the log-likelihood of the complete data has the following expression:

$$\begin{aligned} l_c(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{b}) &= \log\left\{f(\mathbf{Y}|\mathbf{X}, \mathbf{b}, \{\psi_k\}, \{\boldsymbol{\beta}_k\}) f(\mathbf{b}|\boldsymbol{\Sigma}_b)\right\} \\ &= \sum_{i=1}^{M}\sum_{j=1}^{m_i}\sum_{k=1}^{q}\left[Y_{ijk}\left\{\psi_k(\boldsymbol{\beta}_k^T\mathbf{X}_{ij}) + \mathbf{g}_k^T\mathbf{b}_i\right\} - \log\left[1 + \exp\left\{\psi_k(\boldsymbol{\beta}_k^T\mathbf{X}_{ij}) + \mathbf{g}_k^T\mathbf{b}_i\right\}\right]\right] \\ &\quad + \sum_{i=1}^{M}\left\{-\frac{q}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_b| - \frac{1}{2}\mathbf{b}_i^T\boldsymbol{\Sigma}_b^{-1}\mathbf{b}_i\right\}. \end{aligned} \tag{2}$$

Since $\{\boldsymbol{\beta}_k\}$, $\{\psi_k\}$ and $\boldsymbol{\Sigma}_b$ appear in different terms of $l_c(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{b})$, the EM algorithm updates the estimate of $\boldsymbol{\theta}$ by separately maximizing $\mathbb{E}_{\mathbf{b}|\mathbf{Y},\boldsymbol{\theta}^*}\{\log f(\mathbf{Y}|\mathbf{X}, \mathbf{b}, \{\psi_k\}, \{\boldsymbol{\beta}_k\})|\mathbf{Y}, \boldsymbol{\theta}^*\}$ and $\mathbb{E}_{\mathbf{b}|\mathbf{Y},\boldsymbol{\theta}^*}\{\log f(\mathbf{b}|\boldsymbol{\Sigma}_b)|\mathbf{Y}, \boldsymbol{\theta}^*\}$ with respect to $(\{\boldsymbol{\beta}_k\}, \{\psi_k\})$ and $\boldsymbol{\Sigma}_b$. While the link functions $\{\psi_k\}$ do not have any parametric forms, some nonparametric techniques should be used for estimating them. By following the idea of rMAVE, for any given k and $\mathbf{X}_{i'j'}$ such that $\boldsymbol{\beta}_k^T\mathbf{X}_{i'j'}$ is close to $\boldsymbol{\beta}_k^T\mathbf{X}_{ij}$, the local linear approximation of $\psi_k(\boldsymbol{\beta}_k^T\mathbf{X}_{ij})$ in a neighborhood of $\boldsymbol{\beta}_k^T\mathbf{X}_{i'j'}$ can be expressed as $\psi_k(\boldsymbol{\beta}_k^T\mathbf{X}_{ij}) = a_{i'j'k} + c_{i'j'k}\boldsymbol{\beta}_k^T(\mathbf{X}_{ij} - \mathbf{X}_{i'j'})$, where $a_{i'j'k}$ and $c_{i'j'k}$ denote the values of ψ_k and ψ_k' evaluated at $\boldsymbol{\beta}_k^T\mathbf{X}_{i'j'}$, respectively. Then, $\log f(\mathbf{Y}|\mathbf{X}, \mathbf{b}, \{\psi_k\}, \{\boldsymbol{\beta}_k\})$ can be approximated by

$$\log f(\mathbf{Y}|\mathbf{X}, \mathbf{b}, \{\boldsymbol{\beta}_k\}, \{\mathbf{a}_k\}, \{\mathbf{c}_k\}) = \sum_{k=1}^{q}\sum_{i,i'=1}^{M}\sum_{j=1}^{m_i}\sum_{j'=1}^{m_{i'}}\left[Y_{ijk}\eta_{iji'j'k} - \log\left\{1 + \exp(\eta_{iji'j'k})\right\}\right] w_{iji'j'k}, \tag{3}$$

where

$$\begin{aligned} \eta_{iji'j'k} &= a_{i'j'k} + c_{i'j'k}\boldsymbol{\beta}_k^T(\mathbf{X}_{ij} - \mathbf{X}_{i'j'}) + \mathbf{g}_k^T\mathbf{b}_i, \\ w_{iji'j'k} &= \frac{K_{h_k}[\boldsymbol{\beta}_k^T(\mathbf{X}_{ij} - \mathbf{X}_{i'j'})]}{\sum_{i=1}^{M}\sum_{j=1}^{m_i} K_{h_k}[\boldsymbol{\beta}_k^T(\mathbf{X}_{ij} - \mathbf{X}_{i'j'})]}, \\ \mathbf{a}_k &= (a_{11k}, a_{12k}, \ldots, a_{1m_1k}, a_{21k}, \ldots, a_{Mm_Mk})^T, \\ \mathbf{c}_k &= (c_{11k}, c_{12k}, \ldots, c_{1m_1k}, c_{21k}, \ldots, c_{Mm_Mk})^T, \end{aligned}$$

$K_{h_k}(\cdot) = K(\cdot/h_k)/h_k$, for $k = 1, \ldots, q$, $K(\cdot)$ is a density kernel function, and $\{h_k\}$ are q bandwidths.

Based on the approximation given by (3), the EM algorithm that iteratively updates the parameter estimates can be summarized by the following formulas: for $k = 1, \ldots, q$,

$$(\widehat{\boldsymbol{\beta}}_k, \widehat{\mathbf{a}}_k^T, \widehat{\mathbf{c}}_k^T)^T = \underset{\boldsymbol{\beta}_k, \mathbf{a}_k, \mathbf{c}_k}{\operatorname{argmax}}\, \mathbb{E}_{\mathbf{b}|\mathbf{Y}, \widehat{\boldsymbol{\Sigma}}_b, \{\widetilde{\boldsymbol{\beta}}_k\}, \{\widetilde{\mathbf{a}}_k\}, \{\widetilde{\mathbf{c}}_k\}}\left\{\log f(\mathbf{Y}|\mathbf{X}, \mathbf{b}, \{\boldsymbol{\beta}_k\}, \{\mathbf{a}_k\}, \{\mathbf{c}_k\})\middle|\mathbf{Y}, \widehat{\boldsymbol{\Sigma}}_b, \{\widetilde{\boldsymbol{\beta}}_k\}, \{\widetilde{\mathbf{a}}_k\}, \{\widetilde{\mathbf{c}}_k\}\right\}, \tag{4}$$

$$\widehat{\boldsymbol{\Sigma}}_b = \underset{\boldsymbol{\Sigma}_b}{\operatorname{argmax}}\, \mathbb{E}_{\mathbf{b}|\mathbf{Y}, \widetilde{\boldsymbol{\Sigma}}_b, \{\widehat{\boldsymbol{\beta}}_k\}, \{\widehat{\mathbf{a}}_k\}, \{\widehat{\mathbf{c}}_k\}}\left\{\log f(\mathbf{b}|\boldsymbol{\Sigma}_b)\middle|\mathbf{Y}, \widetilde{\boldsymbol{\Sigma}}_b, \{\widehat{\boldsymbol{\beta}}_k\}, \{\widehat{\mathbf{a}}_k\}, \{\widehat{\mathbf{c}}_k\}\right\}, \tag{5}$$

where $\{\widehat{\boldsymbol{\beta}}_k\}$, $\widehat{\boldsymbol{\Sigma}}_b$, $\{\widehat{\mathbf{a}}_k\}$ and $\{\widehat{\mathbf{c}}_k\}$ are the parameter estimates in the current iteration, and $\{\widetilde{\boldsymbol{\beta}}_k\}$, $\widetilde{\boldsymbol{\Sigma}}_b$, $\{\widetilde{\mathbf{a}}_k\}$ and $\{\widetilde{\mathbf{c}}_k\}$ are the parameter estimates obtained in the previous iteration.

Since the conditional expectations in (4) and (5) do not have closed-form expressions, they can be replaced respectively by the following estimates:

$$\frac{1}{B}\sum_{l=1}^{B}\log f(\mathbf{Y}|\mathbf{X},\mathbf{b}^{(l)},\{\widetilde{\boldsymbol{\beta}}_k\},\{\widetilde{\mathbf{a}}_k\},\{\widetilde{\mathbf{c}}_k\}) \quad \text{and} \quad \frac{1}{B}\sum_{l=1}^{B}\log f(\mathbf{b}^{(l)}|\widetilde{\boldsymbol{\Sigma}}_b),$$

where $\{\mathbf{b}^{(l)}, l = 1,\ldots,B\}$ are the values sampled from the posterior distribution of $\mathbf{b}$ given the current estimates of other parameters. To this end, we refer to McCulloch[20] for generating the posterior samples of the random effects using the Metropolis-Hastings algorithm. By this algorithm, the probability of accepting a new value of the random-effects vector, say $\mathbf{b}^*$, is the minimum of one and

$$\frac{f(\mathbf{b}^*|\mathbf{Y},\mathbf{X},\widehat{\boldsymbol{\theta}})f(\mathbf{b}|\widehat{\boldsymbol{\Sigma}}_b)}{f(\mathbf{b}|\mathbf{Y},\mathbf{X},\boldsymbol{\theta})f(\mathbf{b}^*|\widehat{\boldsymbol{\Sigma}}_b)}.$$

In addition, the above ratio can be simplified to the ratio of conditional likelihoods as follows.

$$\begin{aligned}\frac{f(\mathbf{b}^*|\mathbf{Y},\mathbf{X},\widehat{\boldsymbol{\theta}})f(\mathbf{b}|\widehat{\boldsymbol{\Sigma}}_b)}{f(\mathbf{b}|\mathbf{Y},\mathbf{X},\boldsymbol{\theta})f(\mathbf{b}^*|\widehat{\boldsymbol{\Sigma}}_b)} &= \frac{f(\mathbf{Y}|\mathbf{X},\mathbf{b}^*,\widehat{\boldsymbol{\theta}})f(\mathbf{b}^*|\widehat{\boldsymbol{\Sigma}}_b)f(\mathbf{b}|\widehat{\boldsymbol{\Sigma}}_b)}{f(\mathbf{Y}|\mathbf{X},\mathbf{b},\widehat{\boldsymbol{\theta}})f(\mathbf{b}|\widehat{\boldsymbol{\Sigma}}_b)f(\mathbf{b}^*|\widehat{\boldsymbol{\Sigma}}_b)} = \frac{f(\mathbf{Y}|\mathbf{X},\mathbf{b}^*,\widehat{\boldsymbol{\theta}})}{f(\mathbf{Y}|\mathbf{X},\mathbf{b},\widehat{\boldsymbol{\theta}})}\\ &= \frac{\prod_{i=1}^{M} f(\mathbf{Y}_{i\bullet\bullet}|\mathbf{X}_i,\mathbf{b}_i^*,\widehat{\boldsymbol{\theta}})}{\prod_{i=1}^{M} f(\mathbf{Y}_{i\bullet\bullet}|\mathbf{X}_i,\mathbf{b}_i,\widehat{\boldsymbol{\theta}})}.\end{aligned}$$

This simplification avoids calculating the unknown marginal density of $\mathbf{Y}$ that involves complicated integrals without analytical forms. In practice, updating $\mathbf{b}$ would induce low acceptance rates and intensive computation given the high dimensionality of the random-effects vector. Because of the independence of the subject-specific random-effects $\mathbf{b}_i$'s, we can parallelly generate the Monte Carlo samples from the posterior distribution for each subject by using the single-component Metropolis-Hastings algorithm suggested by Gilks et al.[22] More Specifically, based on the current parameter estimates $\widehat{\boldsymbol{\theta}}$, the sampling process for the ith subject, for $i = 1,\ldots,M$, can be realized as follows:

1) Set the initial values $\mathbf{b}_i^{(l)} = (b_{i1}^{(l)},\ldots,b_{iq}^{(l)})^T = (0,\ldots,0)^T$ with $l = 0$.

2) Let $l = l + 1$. For $k = 1,\ldots,q$, implement the following steps

   i) Generate a random value $b_{ik}^{*(l)}$ from the conditional distribution of $b_{ik}^{(l)}$ given

   $$\mathbf{b}_{i,-k}^{(l)} = (b_{i1}^{(l)},\ldots,b_{i,k-1}^{(l)},b_{i,k+1}^{(l)},\ldots,b_{iq}^{(l)})^T \text{ and } \widehat{\boldsymbol{\Sigma}}_b.$$

   ii) Independently sample u_k from the Uniform[0,1] distribution.

   iii) If $u_k < \min\{1, f(\mathbf{Y}_{i\bullet k}|\mathbf{X}_i,b_{ik}^*,\widehat{\boldsymbol{\theta}})/f(\mathbf{Y}_{i\bullet k}|\mathbf{X}_i,b_{ik}^{(l)},\widehat{\boldsymbol{\theta}})\}$, then accept $b_{ik}^{*(l)}$ and update $\mathbf{b}_i^{(l)} = (b_{i1}^{(l)},\ldots,b_{i,k-1}^{(l)},b_{ik}^{*(l)},b_{i,k+1}^{(l)},\ldots,b_{iq}^{(l)})^T$. Otherwise, set $\mathbf{b}_i^{(l)} = (b_{i1}^{(l)},\ldots,b_{i,k-1}^{(l)},b_{ik}^{(l)},b_{i,k+1}^{(l)},\ldots,b_{iq}^{(l)})^T$.

3) Repeat Step 2) until B sampled values of $\mathbf{b}_i$, say $\mathbf{b}_i^{(1)},\ldots,\mathbf{b}_i^{(B)}$, have been obtained.

Practically, we often generate more than B sets of $\{\mathbf{b}_i^{(l)}\}$ and discard the first several burn-in samples in order to guarantee the quality of the remaining B samples. In part i) of Step 2) above, since the elements of $\mathbf{b}_i$ may be correlated with each other, the

proposal distribution in this single-component Metropolis-Hastings algorithm should be modified into a conditional distribution given the current values of all parameter estimates except the one to be sampled. Since the subject-specific random effects are assumed to be i.i.d. with a multivariate Normal distribution, Appendix A shows that $b_{ik}^{*(l)}$ can be generated from a Normal distribution with mean $-\sum_{k'\neq 1} S_{1k'}S_{11}^{-1}b_{ik'}^{(l)}$ and variance S_{11}^{-1}, where $S_{kk'}$ is the (k,k')-th element of the inverse of a re-arranged version of $\widehat{\boldsymbol{\Sigma}}_b$ that corresponds to the vector $(b_{ik}^{(l)}, \mathbf{b}_{i,-k}^{(l)})$. Once we are able to obtain the posterior samples of the random effects, the approximated version in (4) and (5) can be obtained by the iteratively reweighted least square procedure and a closed-form formula, respectively. The entire iterative estimation procedure for estimating the parameters in our main model is summarized below.

Monte Carlo EM Algorithm for Estimating Model (1)

1) Choose the initial values of $\boldsymbol{\beta}_k$, for $k = 1, \ldots, q$, and $\boldsymbol{\Sigma}_b$, denoted as $\widehat{\boldsymbol{\beta}}_k^{(0)}$ and $\widehat{\boldsymbol{\Sigma}}_b^{(0)}$. Initialize $\mathbf{b}^{(l)}$, for $l = 1, \ldots, B$. Set the iteration index $r = 0$.

2) For $r \geq 1$ and $k = 1, \ldots, q$, obtain $\widehat{\mathbf{a}}_k^{(r)}$ and $\widehat{\mathbf{c}}_k^{(r)}$ based on $\{\widehat{\boldsymbol{\beta}}_k^{(r-1)}\}$ and the sampled values of $\{\mathbf{b}^{(l)}\}$ obtained in the $(r-1)$-th iteration as follows. For $i' = 1, \ldots, M$ and $j' = 1, \ldots, m$, iteratively update $(\widehat{a}_{i'j'k}, \widehat{c}_{i'j'k})^T$ by the following formula until convergence:

$$\begin{aligned}\begin{pmatrix}\widehat{a}_{i'j'k}\\ \widehat{c}_{i'j'k}\end{pmatrix} = &\left[\frac{1}{B}\sum_{l=1}^{B}\sum_{i=1}^{M}\sum_{j=1}^{m_i} K_{h_k}\{\widehat{\boldsymbol{\beta}}_k^{(r-1)T}(\mathbf{X}_{ij}-\mathbf{X}_{i'j'})\}\begin{pmatrix}1\\ \widehat{\boldsymbol{\beta}}_k^{(r-1)T}(\mathbf{X}_{ij}-\mathbf{X}_{i'j'})\end{pmatrix}\times\begin{pmatrix}1\\ \widehat{\boldsymbol{\beta}}_k^{(r-1)T}(\mathbf{X}_{ij}-\mathbf{X}_{i'j'})\end{pmatrix}^T \widehat{\mu}_{iji'j'k}^{(l)}\left(1-\widehat{\mu}_{iji'j'k}^{(l)}\right)\right]^{-1}\\ &\times\left[\frac{1}{B}\sum_{l=1}^{B}\sum_{i=1}^{M}\sum_{j=1}^{m_i} K_{h_k}\{\widehat{\boldsymbol{\beta}}_k^{(r-1)T}(\mathbf{X}_{ij}-\mathbf{X}_{i'j'})\}\begin{pmatrix}1\\ \widehat{\boldsymbol{\beta}}_k^{(r-1)T}(\mathbf{X}_{ij}-\mathbf{X}_{i'j'})\end{pmatrix}\right.\\ &\left.\times\left[\widehat{\mu}_{iji'j'k}^{(l)}\left(1-\widehat{\mu}_{iji'j'k}^{(l)}\right)\{\widehat{a}_{i'j'k}+\widehat{c}_{i'j'k}\widehat{\boldsymbol{\beta}}_k^{(r-1)T}(\mathbf{X}_{ij}-\mathbf{X}_{i'j'})\}+\left(Y_{ijk}-\widehat{\mu}_{iji'j'k}^{(l)}\right)\right]\right],\end{aligned}$$

where

$$\widehat{\mu}_{iji'j'k}^{(l)} = \frac{\exp\{\widehat{a}_{i'j'k}+\widehat{c}_{i'j'k}\widehat{\boldsymbol{\beta}}_k^{(r-1)T}(\mathbf{X}_{ij}-\mathbf{X}_{i'j'})+\mathbf{g}_k^T\mathbf{b}_i^{(l)}\}}{1+\exp\{\widehat{a}_{i'j'k}+\widehat{c}_{i'j'k}\widehat{\boldsymbol{\beta}}_k^{(r-1)T}(\mathbf{X}_{ij}-\mathbf{X}_{i'j'})+\mathbf{g}_k^T\mathbf{b}_i^{(l)}\}}.$$

Then, $\widehat{\mathbf{a}}_k^{(r)}$ and $\widehat{\mathbf{c}}_k^{(r)}$ are obtained by permuting the convergent values of $\{\widehat{a}_{i'j'k}\}$ and $\{\widehat{c}_{i'j'k}\}$.

3) For $k = 1, \ldots, q$, update the estimate of $\boldsymbol{\beta}_k$ through

$$\widehat{\boldsymbol{\beta}}_k^{(r)} = \left[\frac{1}{B} \sum_{l=1}^{B} \sum_{i,i'=1}^{M} \sum_{j=1}^{m_i} \sum_{j'=1}^{m_{i'}} K_{h_k}\{\widehat{\boldsymbol{\beta}}_k^{(r-1)T}(\mathbf{X}_{ij} - \mathbf{X}_{i'j'})\}(\mathbf{X}_{ij} - \mathbf{X}_{i'j'})(\mathbf{X}_{ij} - \mathbf{X}_{i'j'})^T \right.$$
$$\left. \times (\widehat{c}_{i'j'k}^{(r)})^2 \times \widehat{\mu}_{iji'j'k}^{(l)*} \left(1 - \widehat{\mu}_{iji'j'k}^{(l)*}\right) / \widehat{f}_{\boldsymbol{\beta}_k^T \mathbf{x}}(\widehat{\boldsymbol{\beta}}_k^{(r-1)T} \mathbf{X}_{i'j'}) \right]^{-1}$$
$$\left[\frac{1}{B} \sum_{l=1}^{B} \sum_{i,i'=1}^{M} \sum_{j=1}^{m_i} \sum_{j'=1}^{m_{i'}} K_{h_k}\{\widehat{\boldsymbol{\beta}}_k^{(r-1)T}(\mathbf{X}_{ij} - \mathbf{X}_{i'j'})\}(\mathbf{X}_{ij} - \mathbf{X}_{i'j'}) \widehat{c}_{i'j'k}^{(r)} \left\{ \widehat{\mu}_{iji'j'k}^{(l)*} \left(1 - \widehat{\mu}_{iji'j'k}^{(l)*}\right) \right.\right.$$
$$\left.\left. \times \widehat{\boldsymbol{\beta}}_k^{(r-1)T} \widehat{c}_{i'j'k}^{(r)} (\mathbf{X}_{ij} - \mathbf{X}_{i'j'}) + \left(Y_{ijk} - \widehat{\mu}_{iji'j'k}^{(l)*}\right) \right\} / \widehat{f}_{\boldsymbol{\beta}_k^T \mathbf{x}}(\widehat{\boldsymbol{\beta}}_k^{(r-1)T} \mathbf{X}_{i'j'}) \right],$$

where

$$\widehat{\mu}_{iji'j'k}^{(l)*} = \frac{\exp\{\widehat{a}_{i'j'k}^{(r)} + \widehat{c}_{i'j'k}^{(r)} \widehat{\boldsymbol{\beta}}_k^{(r-1)T}(\mathbf{X}_{ij} - \mathbf{X}_{i'j'}) + \mathbf{g}_k^T \mathbf{b}_i^{(l)}\}}{1 + \exp\{\widehat{a}_{i'j'k}^{(r)} + \widehat{c}_{i'j'k}^{(r)} \widehat{\boldsymbol{\beta}}_k^{(r-1)T}(\mathbf{X}_{ij} - \mathbf{X}_{i'j'}) + \mathbf{g}_k^T \mathbf{b}_i^{(l)}\}},$$

$$\widehat{f}_{\boldsymbol{\beta}_k^T \mathbf{x}}(\widehat{\boldsymbol{\beta}}_k^{(r-1)T} \mathbf{X}_{i'j'}) = \frac{1}{\sum_{i=1}^{M} m_i} \sum_{i=1}^{M} \sum_{j=1}^{m_i} K_{h_k}\{\widehat{\boldsymbol{\beta}}_k^{(r-1)T}(\mathbf{X}_{ij} - \mathbf{X}_{i'j'})\}.$$

Then, standardize $\widehat{\boldsymbol{\beta}}_k^{(r)}$ such that its L_2-norm is one and its first element is positive.

4) Update the variance-covariance matrix estimate through

$$\widehat{\boldsymbol{\Sigma}}_b^{(r)} = \frac{1}{MB} \sum_{i=1}^{M} \sum_{l=1}^{B} \mathbf{b}_i^{(l)} \mathbf{b}_i^{(l)T}.$$

5) Implement the single-component Metropolis-Hastings algorithm for each subject to obtain the posterior samples $\{\mathbf{b}_i^{(l)}\}$ used in the next iteration.

6) Repeat Steps 2 to 5 until the convergence of $\{\widehat{\boldsymbol{\beta}}_k^{(r)}\}$ and $\widehat{\boldsymbol{\Sigma}}_b^{(r)}$.

In practice, the initial values for $\{\boldsymbol{\beta}_k\}$ and $\boldsymbol{\Sigma}_b$ can be obtained by fitting q separate generalized linear mixed-effects models with logit link functions and random intercepts. Then, $\widehat{\boldsymbol{\beta}}_k^{(0)}$ is chosen to be the standardized version of the fitted coefficient vector, for each k, and $\widehat{\boldsymbol{\Sigma}}_b^{(0)}$ is chosen to be a diagonal matrix of the q fitted variances of the random intercepts. The updating formula for $\boldsymbol{\Sigma}_b$ in Step 4) is just the Monte Carlo approximation of the averaged posterior mean of $\mathbf{b}_i \mathbf{b}_i^T$, which can be derived analytically from (5). Here, we follow the suggestion of Booth and Hobert[21] to use an automatically increased simulation size B so that the algorithm is more computationally efficient. Namely, we set $B = 100$ for the first 10 iterations. Starting from the 11th iteration, B is increased to 200. After the 31th iterations, B is set to be 500.

2.3 | Estimation of $\psi_k(\cdot)$

In Step 2) of the Monte Carlo EM algorithm discussed at the end of Subsection 2.2, $\widehat{a}_{i'j'k}$ and $\widehat{c}_{i'j'k}$ are estimates of ψ_k and ψ_k' evaluated at $\widehat{\boldsymbol{\beta}}_k\mathbf{X}_{i'j'}$, respectively, for $i' = 1, \ldots, M$, $j' = 1, \ldots, m_i$, and $k = 1, \ldots, q$. In this section, we first formally define the estimators of the link functions in the presence of random effects, and then derive some statistical properties of the estimators.

Let us first consider the case when $\{\boldsymbol{\beta}_k\}$ and $\boldsymbol{\Sigma}_b$ are assumed known. For $k = 1, \ldots, q$, let a_{xk} and c_{xk} denote the values of ψ_k and ψ_k' evaluated at $\boldsymbol{\beta}_k^T\mathbf{x}$. Then, similar to the exposition in Cui et al.[9], the estimators $\widehat{\psi}_k(\boldsymbol{\beta}_k^T\mathbf{x})$ and $\widehat{\psi}_k'(\boldsymbol{\beta}_k^T\mathbf{x})$ can be obtained by solving the following equation with respect to $\boldsymbol{\nu}_{xk} = (a_{xk}, c_{xk})^T$:

$$\mathbf{S}(\boldsymbol{\nu}_{xk}) := \mathbb{E}_{\mathbf{b}|\mathbf{Y},\boldsymbol{\Sigma}_b}\left[\sum_{i=1}^{M}\sum_{j=1}^{m_i} K_{h_k}\{\boldsymbol{\beta}_k^T(\mathbf{X}_{ij}-\mathbf{x})\}\left(Y_{ijk}-\mu_{ijxk}\right)\begin{pmatrix}1\\ \boldsymbol{\beta}_k^T(\mathbf{X}_{ij}-\mathbf{x})\end{pmatrix}\middle|\mathbf{Y},\boldsymbol{\Sigma}_b\right] = 0, \tag{6}$$

where

$$\mu_{ijxk} = \frac{\exp\{a_{xk} + c_{xk}\boldsymbol{\beta}_k(\mathbf{X}_{ij}-\mathbf{x}) + \mathbf{g}_k^T\mathbf{b}_i\}}{1+\exp\{a_{xk} + c_{xk}\boldsymbol{\beta}_k(\mathbf{X}_{ij}-\mathbf{x}) + \mathbf{g}_k^T\mathbf{b}_i\}}.$$

In practice, as discussed in Subsection 2.2, the conditional expectation can be replaced by its estimate using the posterior samples of the random effects. The following theorem gives the asymptotic conditional bias and variance of each estimated link function defined above.

Theorem 1. Under the regularity conditions given in Appendix B.1, if $h_k = o(1)$, $1/\sum_{i=1}^{M} m_i = o(1)$, and $1/(h_k\sum_{i=1}^{M} m_i) = o(1)$, for $k = 1, \ldots, q$, then we have

(i) $\mathbb{E}\left[\widehat{\psi}_k(\boldsymbol{\beta}_k^T\mathbf{x}) - \psi_k(\boldsymbol{\beta}_k^T\mathbf{x})|\mathbf{X}\right] = \frac{1}{2}h_k^2\psi_k''(\boldsymbol{\beta}_k^T\mathbf{x})\int t^2K(t)dt + O_p\left[h_k^4 + (h_k\sum_{i=1}^{M} m_i)^{-1}\right]$,

(ii) $\mathrm{Var}\left[\widehat{\psi}_k(\boldsymbol{\beta}_k^T\mathbf{x})|\mathbf{X}\right] = \left[h_k f_{\beta_k^T x}(\boldsymbol{\beta}_k^T\mathbf{x})\mathbb{E}_{\mathbf{b}|\mathbf{Y},\boldsymbol{\Sigma}_b}\left\{\sum_{i=1}^{M} m_i v_k(\mathbf{b}_\mathbf{i})\right\}\right]^{-1}\int K^2(t)dt + O_p\left[h_k^4 + (h_k\sum_{i=1}^{M} m_i)^{-1}\right]$,

where $v_k(\mathbf{b}_\mathbf{i}) = \exp\{\psi_k(\boldsymbol{\beta}_k^T\mathbf{x}) + \mathbf{g}_k^T\mathbf{b}_i\}/[1+\exp\{\psi_k(\boldsymbol{\beta}_k^T\mathbf{x}) + \mathbf{g}_k^T\mathbf{b}_i\}]^2$.

The proof of Theorem 1 is given in Appendix B.2. While the score function in (6) takes a form similar to the local linear kernel estimating functions discussed in Fan et al.[23] and Cui et al.[9] where the logit function is used to account for the binary data, we can see that the asymptotic conditional bias of the proposed estimator in the current problem has a similar asymptotic expression to that in cases with independent binary data discussed in Cui et al.[9] In the asymptotic expression of the conditional variance given above, a term with the conditional expectation is present. Since this term has no explicit analytical form, its estimation may need to be obtained using the Monte Carlo sampling approach similar to the one discussed in Section 2.2.

Given the good theoretical properties of the Epanechnikov kernel function given in the literature,[24] $K(\cdot)$ is chosen to be that kernel function, which takes the form of $K(x) = 0.75(1-x^2)\mathbf{I}(|x| \le 1)$. For the kth response variable, the bandwidth h_k is used in Steps 2) and 3) of the proposed Monte Carlo EM algorithm. Since the asymptotic conditional mean integrated squared error (MISE) of the related link function estimator would involve intractable terms, it is difficult to derive a formula for the optimal

bandwidth that minimizes the asymptotic conditional MISE. Here, we suggest using the N-fold cross-validation method to select the bandwidths for estimating the q link functions. Compared to the conventional leave-one-out CV procedure, the N-fold CV would be computationally more feasible, particularly in cases when we need to update the bandwidths when the estimated single-indices get updated. To account for the within-subject correlation, the CV score needs to incorporate the current estimate of $\mathbf{\Sigma}_b$. After taking all these considerations into account, we suggest choosing h_k by minimizing the following CV score in a given iteration of the proposed Monte Carlo EM algorithm:

$$\mathrm{CV}(h_k) = \frac{1}{B}\sum_{l=1}^{B}\sum_{i=1}^{M}\sum_{j=1}^{m_i}\left[Y_{ijk}\{\widehat{\psi}_{k,-n_{ij}}(\widehat{\boldsymbol{\beta}}_k^T\mathbf{X}_{ij}) + \mathbf{g}_k^T\mathbf{b}_i^{(l)}\} - \log\left[1+\exp\{\widehat{\psi}_{k,-n_{ij}}(\widehat{\boldsymbol{\beta}}_k^T\mathbf{X}_{ij}) + \mathbf{g}_k^T\mathbf{b}_i^{(l)}\}\right]\right], \tag{7}$$

where $\mathbf{b}_i^{(l)}$ denotes the posterior sample of the random effects for the ith subject obtained by the Metropolis-Hastings algorithm based on the current estimate of $\mathbf{\Sigma}_b$, and $\widehat{\psi}_{k,-n_{ij}}(\widehat{\boldsymbol{\beta}}_k^T\mathbf{X}_{ij})$ is the leave-one-fold-out estimate of $\psi_k(\widehat{\boldsymbol{\beta}}_k^T\mathbf{X}_{ij})$ obtained by solving the modified version of (6) with observations in the n_{ij}-th fold being omitted in the calculation, for $n_{ij} \in \{1, 2, \ldots, N\}$. In the fold assignments, since observations within a subject could be correlated with each other, "subject" is used as the basic unit for assignments and observations of the same subject are always assigned to the same fold. In practice, we suggest choosing $N = 5$ or 10, and numerical results presented in Sections 3 and 4 show that the choice of N has a negligible effect on the performance of the estimated model.

3 | SIMULATION STUDY

In this section, two sets of simulations are carried out to assess the numerical performance of our proposed method. The first set of simulations evaluates the finite-sample performance of the proposed method in various cases considered, and the second set compares it with the rMAVE method proposed by Xia[6] and the EFM method proposed by Cui et al.[9] Throughout this section, it is assumed that all m_i's are the same to be m, $M = 50, 100$, or 200, and $m = 5$, or 10.

3.1 | Finite-sample performance of the proposed method

Suppose we have $q = 2$ correlated binary response variables whose observations are generated from the following model:

$$\begin{cases} \mathrm{logit}\{\mathbb{E}(Y_{ij1}|\mathbf{X}_{ij}, \mathbf{b}_i)\} = 2(\beta_{11}X_{ij1} + \beta_{12}X_{ij2} + \beta_{13}X_{ij3}) - 1 + b_{i1} \\ \mathrm{logit}\{\mathbb{E}(Y_{ij2}|\mathbf{X}_{ij}, \mathbf{b}_i)\} = 3(\beta_{21}X_{ij2} + \beta_{22}X_{ij2} + \beta_{23}X_{ij3})^2 - 2 + b_{i2}, \end{cases} \tag{8}$$

where the single indices are linked to the logit of the conditional mean responses via the link functions $\psi_1(u) = 2u - 1$ and $\psi_2(u) = 3u^2 - 2$, $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \beta_{13})^T = (1, -1, 0)^T/\sqrt{2}$, and $\boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22}, \beta_{23})^T = (2, 1, 3)^T/\sqrt{14}$. The observation times $t_{ij} \in [0, 1]$ are generated independently from the uniform distribution $\mathrm{U}((j-1)/m, j/m)$, for $i = 1, \ldots, M$ and $j = 1, \ldots, m$.

Observations of the first two predictors $\{X_{ij1}\}$ and $\{X_{ij2}\}$ are assumed to be time-independent and generated from U(0, 1) and U(−1, 1), respectively. Observations X_{ij3} are generated from U(1, 2), multiplying by the corresponding observation times t_{ij}. Thus, they are time-dependent. For the ith subject, the random-effects $\mathbf{b}_i = (b_{i1}, b_{i2})^T$ are generated from a bivariate normal distribution $N_2(\mathbf{0}, \mathbf{\Sigma}_b)$ with

$$\mathbf{\Sigma}_b = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

In evaluating the finite-sample performance of our proposed method, we compare the estimated parameter values and their true values in various cases considered based on 200 repeated simulation runs. Tables 1 and 2 present the biases of the parameter estimates, their variances (Var), and the mean squared errors (MSE) when $m = 5$ and 10, respectively. From the tables, it can be seen that most biases and all Var and MSE values of the parameter estimates decrease as either the number of subjects (i.e., M) or the number of repeated measurements (i.e., m) within a subject increases. Figure 1 further shows the histograms of the estimated values of β_{11} from the 200 the simulation runs when M and m change. From the figure, it can be seen that while the averaged estimated parameter value gets closer to the true value $\beta_{11} = 0.7071$ shown by the vertical dashed lines in the plots, as either M or m increases, the variability of the estimated values gets smaller and their distribution gets more and more symmetric and bell-shaped. Figures A.1-A.5 in Appendix C present the histograms of the estimates of other parameters, and show the similar patterns. These numerical results demonstrate the asymptotic normality of the estimated index coefficients by the proposed method. Theoretical justification of the proposed method is left for our future research.

[Table 1 about here]

[Table 2 about here]

Next, we examine the performance of the estimated link functions in various cases considered. Figures 2 and 3 compare the true link functions with their estimates in different cases of the sample size. From the figures, it can be seen that the pointwise estimates of the first link function are almost identical with the true function values as the sample size increases, and the pointwise estimates of the second link function are getting closer to the true values as the sample size increases. It is also clear that for both link functions, the empirical 95% pointwise confidence intervals are becoming narrower as M and/or m increase. These figures show that the estimated link functions by our proposed method perform well too.

[Figure 1 about here]

[Figure 2 about here]

[Figure 3 about here]

3.2 | Method comparison

In this part, we compare the numerical performance of our proposed method with the rMAVE and EFM methods, which are two representative existing methods developed to handle cases with independent observed data and a univariate response variable. The comparison is conducted in three scenarios corresponding to three different assumptions on the correlation structure of the observed data. While the rMAVE method was first proposed to estimate a single-index model with a continuous response variable, it can be generalized easily to cases with a univariate binary response variable. Cui et al.[9] compared the EFM method with rMAVE and showed that the EFM method gave more accurate estimates of the index coefficients when the number of the predictors p was large and could be implemented in cases when p was too large for rMAVE to implement. When the rMAVE and EFM methods are used to analyze correlated data with multiple response variables, they are implemented for the observed data of each response variable and all predictors.

Similar to the setup of the simulation study in Section 3.1, assume that there are two binary response variables whose observations are generated from model (8). The within-subject correlation is controlled by the random-effects $\mathbf{b}_i = (b_{i1}, b_{i2})^T$, which are generated in the way described in the following three scenarios:

- Scenario 1: The observed data are assumed to be independent within and between different subjects, i.e., $b_{i1} = b_{i2} = 0$.

- Scenario 2: The two response variables are independent with each other while the repeated measurements on each response variable within a subject are correlated. In such cases, the random-effects terms $\mathbf{b}_i$ are generated from the distribution $\mathrm{N}_2(\mathbf{0}, \mathbf{\Sigma}_b)$, where $\mathbf{\Sigma}_b$ is a 2×2 identity matrix.

- Scenario 3: The response variables are correlated and there is within-subject data correlation as well. In such cases, the random-effects terms $\mathbf{b}_i$ are generated from the distribution $\mathrm{N}_2(\mathbf{0}, \mathbf{\Sigma}_b)$, where

$$\mathbf{\Sigma}_b = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

In this example, the sum of squares of the estimation error, $\sum_{p=1}^{3}(\hat{\beta}_{kp} - \beta_{kp})^2$ for $k = 1, 2$, are considered to measure the estimation error for the index coefficients. Table 3 presents the results under different scenarios described above, where "PROPOSED" denotes our proposed method. The table shows that either rMAVE or EFM has the best performance in Scenario 1 when the observed data are independent over time and among different response variables. This result is reasonable because PROPOSED needs to estimate the parameters in the variance-covariance matrix of the random effects in addition to the index coefficients, which would add some extra variability to the estimates of the index coefficients. In Scenarios 2 and 3 when there is within-subject correlation, it can be seen that PROPOSED generally overperforms the other two methods. There is only one exception

that happens in Scenario 3 when the sample size is relatively small, in which case EFM is the best for estimating β_2 and PROPOSED is close to the best. This example shows it is beneficial to use PROPOSED, in comparison with the existing methods rMAVE or EFM, in cases when there are multiple correlated binary response variables and the observed data are longitudinal.

[Table 3 about here]

4 | CASE STUDY

Changes in the extent of social connection have a profound impact on individual people's lifestyle and health outcomes.[25] Social isolation is an objective and quantifiable reflection of reduced social networks and paucity of social connection. It is a particular problem at older ages, when decreasing economic resources, mobility impairment, and the death of contemporaries conspire to limit social contacts.[26] While loneliness is often seen as the emotional manifestation of social isolation, stemming from dissatisfaction with the frequency and quality of social interactions, previous research has shown a relatively weak correlation between social isolation and loneliness.[27] It is worth noting that some individuals may be content with limited social contact, while others may feel lonely despite frequent interactions. Hence, social isolation and loneliness are distinct concepts with potentially different implications for health.[27] Numerous cross-sectional and longitudinal studies have revealed a significant association between chronic diseases (e.g., cardiovascular disease) and social isolation.[28] Moreover, social isolation consistently exhibits a negative impact on health and well-being, with socially isolated and lonely individuals adopting less favorable lifestyles and experiencing lower quality of life.[29] In this section, we use our proposed method to analyze a dataset from the English Longitudinal Study of Aging (ELSA) to study the relationship between the cardiovascular diseases and the extent of social connection as well as the longitudinal relationship between quality of life and social contact. The ELSA project is an ongoing panel study that focuses on adults aged 50 and over. It was initiated in 2002, with participants being followed up approximately every two years. Most of the raw data were collected through face-to-face interviews and self-completed questionnaires. See Cadar et al.[30] and Steptoe et al.[31] for more detailed information about the ELSA project.

In each wave of data collection, the ELSA personnel interviewed participants to know whether they newly developed any specific cardiovascular diseases or recovered from any previously existing medical conditions. Based on that information, we can derive the observed status of cardiovascular diseases, including heart attack, stroke, heart failure, arrhythmia, and heart valve complications, for individual participants during the period from the previous interview to the interview of the current wave. The well-being or quality of life is quantified by the CAPS-19 index, which is the sum of 19 self-reported items with a common 4-point Likert scale coded as 0 to 3.[32] While higher scores on the CAPS-19 index represent higher levels of positive well-being, we create a dichotomous version of it as our second response variable, which reflects whether a given participant has a quality of life score higher than the median score. Since some of the assessments in the CAPS-19 index are health-related and treatments of

cardiovascular diseases would have a great impact on the well-being of the patients,[33] the correlation between the quality of life score and the risk of cardiovascular diseases thus should not be neglected. Besides the two response variables described above, the social isolation score, loneliness score, and age are used as predictors in our study to construct single indices to predict the response variables. The social isolation score is adapted from the index of social isolation developed by Shankar et al.[34] The index assigns one point for each of the following five items: living alone; less than monthly face-to-face, telephone, or written/e-mail contact with children outside the household; less than monthly contact with other relatives outside the household; less than monthly contact with friends; and not participating in any organizations, religious groups, or committees. Therefore, it ranges from 0 to 5 with a higher score indicating a greater social isolation. The loneliness score is constructed based on the University of California, Los Angeles (UCLA) three-item loneliness scale, which covers the frequency and intensity of loneliness feelings.[35] Particularly, the three items are: "How often do you feel you lack companionship?", "How often do you feel left out?", and "How often do you feel isolated from others?". For each question, participants can answer "hardly ever or never" (score of 1), "some of the time" (score of 2), or "often" (score of 3). Thus, a total score ranges from 3 to 9, with a higher value indicating a greater level of loneliness.

Because the measurements of interest in Wave 2-7 of ELSA are quite complete and the ones in other waves have many missing values, we consider a dataset that contains $M = 746$ participants with $m = 6$ observation times each. For the ith participant, Y_{ij1} and Y_{ij2} denote the binary observations of the cardiovascular disease status and the indicator whether the participant has a higher-than-median quality of life score, respectively, at the jth observation time, for each i and j. Similarly, X_{ij1}, X_{ij2} and X_{ij3} denote the age, loneliness score, and isolation score, respectively, of the ith participant at the jth observation time. The three predictors are then standardized to have mean 0 and variance 1 so that a more intuitive interpretation of the index coefficients can be made.

Then, Model (1) is fitted. From the fitted model, we find that there are 5 subjects whose fitted values of the single-indices are far away from the fitted single-indices of other subjects. Thus, the observed data of these 5 subjects are deleted since they are potentially the "influential points" (cf., Cook and Weisberg[36]) for estimating the link functions. Model (1) is then re-fitted, and the estimated index coefficients and their standard errors (SEs) that are computed based on a bootstrap procedure with 200 bootstrap samples are presented in Table 4. The two left panels of Figure 4 show the estimated link functions, and the two right panels show the estimated subject-specific probabilities of developing cardiovascular diseases and having a larger-than-median quality of life scores, respectively, as a function of the estimated single indices. In the two right panels, the observed response values are also presented by little circles. From the two top panels of the figure, it can be seen that $\widehat{\psi}_1(\cdot)$ is almost a linear function of the first single-index with a positive slope when the first single-index takes values in $[-1.5, 3]$. Namely, when the first single-index value ranges from -1.5 to 3, we can see that a larger index value is associated with a higher risk of cardiovascular diseases and the probability increases faster as the single-index value increases. From Table 4, only age has a significant relationship with the

first single-index, since the mean estimated coefficient of age is more than 2 times of the SE. Accordingly, we conclude that when the first single-index value belongs to $[-1.5, 3]$, Age is positively associated with the likelihood of cardiovascular diseases and the impact of the loneliness score and the isolation score is small. When the first single-index takes the values between -3 and -1.5, however, we observe a non-linear relationship between $\widehat{\psi}_1(\cdot)$ and the single-index. While additional analyses and careful interpretations should be made in such cases, the subjects in this scenario have a very low risk of developing cardiovascular diseases. From the two bottom panels of Figure 4, we can see that $\widehat{\psi}_2(\cdot)$ is almost linear with a negative slope when the second single-index takes its value in $[-1, 4]$. This implies that a participant with a smaller value of the second single index would have a larger chance to have a greater-than-median quality of life score. The fluctuation of $\widehat{\psi}_2(\cdot)$ when the second single-index values are from 4 to 6 may be due to the limited observations with such single-index values. From the results in Table 4, both the loneliness score and the isolation score have positive and significant index coefficients while age does not have a significant index coefficient. Therefore, based on the overall trend of $\widehat{\psi}_2(\cdot)$, we can conclude that both the loneliness score and the isolation score have negative association with the likelihood of greater-than-median quality of life score, the loneliness score has a greater impact on the likelihood between the two score types, and age does not have a significant impact on the likelihood.

[Table 4 about here]

[Figure 4 about here]

5 | CONCLUDING REMARKS

In the previous sections, we have introduced a multivariate single-index longitudinal logistic regression model for analyzing longitudinal data with multiple binary response variables. It has been confirmed by theoretical justifications and numerical studies that the proposed method can provide an effective analytic tool for solving the related problem.

It should be pointed out that the proposed method can be generalized in several directions. For instance, more complicated forms of random effects (e.g., random intercepts plus random slopes over time) can be used in Model (1). In this paper, all response variables are assumed to be binary. Actually, cases with a mixture of different types of response variables (e.g., some response variables are continuous while some others take binary or count values) are possible in practice, and they can be handled in a similar way to that discussed in the paper by using different linkage functions for the means of different response variables. In addition, the proposed method still has some issues to address. First, while the estimation of the link functions is separated from the procedure for generating the posterior samples of the random effects in our EM algorithm (cf., the related discussion in Section 2), other more efficient sampling procedure used in the estimation of GLMM might be possible to refine the proposed method. Second, since the random effects in Model (1) introduce the within-subject correlation for the logit transformation of

the mean response variables, we can only make subject-specific interpretation of the single-indices. In addition, there could be a large number predictors in some applications. Thus, variable selection and other issues related to high-dimensional data should be addressed. All these research problems require much future research.

ACKNOWLEDGMENTS

The authors thank the editor, the associate editor, and two anonymous referees for many constructive comments and suggestions, which improved the quality of the paper greatly.

SUPPORTING INFORMATION

Some supplementary materials on the proposed estimation procedure in Section 2 and some additional numerical results related to Section 3 can be found in the online supplementary file available at the Wiley Library Online.

CONFLICT OF INTEREST

The authors have no conflict of interest.

DATA AVAILABILITY STATEMENT

The ELSA datasets analyzed during this study are available in the UK Data Service, https://beta.ukdataservice.ac.uk/datacatalogue/series/series?id=200011.

References

1. Alkhalaqi A, Al-Naimi F, Qassmi R, et al. Visceral adiposity index is a better predictor of type 2 diabetes than body mass index in Qatari population. *Med*. 2020; **99**(35): e21327.

2. Härdle W, Stoker TM. Investigating smooth multiple regression by the method of average derivatives. *J Am Stat Assoc*. 1989; **84**(408): 986–995.

3. Ichimura H. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J Econom*. 1993; **58**(1-2): 71–120.

4. Härdle W, Hall P, Ichimura H. Optimal smoothing in single-index models. *Ann Stat*. 1993; **21**(1): 157–178.

5. Xia Y, Tong H, Li W, Zhu LX. An adaptive estimation of dimension reduction space. *J R Stat Soc Series B Stat Methodol*. 2002; **64**(3): 363–410.

6. Xia Y. Asymptotic distributions for two estimators of the single-index model. *Econ Theory*. 2006; **22**(6): 1112–1137.

7. Yu Y, Ruppert D. Penalized spline estimation for partially linear single-index models. *J Am Stat Assoc*. 2002; **97**(460): 1042–1054.

8. Carroll RJ, Fan J, Gijbels I, Wand MP. Generalized partially linear single-index models. *J Am Stat Assoc*. 1997; **92**(438): 477–489.

9. Cui X, Härdle WK, Zhu L. The EFM approach for single-index models. *Ann Stat*. 2011; **39**(3): 1658—1688.

10. Diggle P. *Analysis of longitudinal data*. Oxford: Oxford University Press; 2002.

11. Chen J, Li D, Liang H, Wang S. Semiparametric GEE analysis in partially linear single-index models for longitudinal data. *Ann Stat*. 2015; **43**(4): 1682–1715.

12. Wu J, Tu W. A multivariate single-index model for longitudinal data. *Stat Modelling*. 2016; **16**(5): 392–408.

13. Tian Z, Qiu P. Multivariate single index modeling of longitudinal data with multiple responses. *Stat in Med*. 2023; **42**(17): 2982-2998

14. Qiu P. *Image processing and jump regression analysis*. New York: John Wiley & Sons; 2005.

15. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Series B Stat Methodol*. 1977; **39**(1): 1–22.

16. Chowdhury SK, Sinha SK. Semiparametric marginal models for binary longitudinal data. *Int J Stat Probab*. 2015; **4**(3): 107.

17. Prentice RL. Correlated binary regression with covariates specific to each binary observation. *Biometrics*. 1988; **44**(4):: 1033–1048.

18. Yi GY, He W, Liang H. Analysis of correlated binary data under partially linear single-index logistic models. *J Multivar Anal*. 2009; **100**(2): 278–290.

19. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics*. 1982; **38**(4): 963–974.

20. McCulloch CE. Maximum likelihood algorithms for generalized linear mixed models. *J Am Stat Assoc*. 1997; **92**(437): 162–170.

21. Booth JG, Hobert JP. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J R Stat Soc Series B Stat Methodol*. 1999; 61(1): 265–285.

22. Gilks WR, Richardson S, Spiegelhalter D. *Markov chain Monte Carlo in practice*. Boca Raton, FL: CRC press; 1995.

23. Fan J, Heckman NE, Wand MP. Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J Am Stat Assoc*. 1995; **90**(429): 141–150.

24. Epanechnikov VA. Non-parametric estimation of a multivariate probability density. *Theory Probab App*. 1969; **14**(1): 153–158.

25. Schrempft S, Jackowska M, Hamer M, Steptoe A. Associations between social isolation, loneliness, and objective physical activity in older men and women. *BMC Public Health*. 2019; **19**(1): 1–10.

26. Steptoe A, Shankar A, Demakakos P, Wardle J. Social isolation, loneliness, and all-cause mortality in older men and women. *Proc Natl Acad Sci*. 2013; **110**(15): 5797–5801.

27. Holt-Lunstad J, Smith TB, Baker M, Harris T, Stephenson D. Loneliness and social isolation as risk factors for mortality: a meta-analytic review. *Perspect Psychol Sci*. 2015; **10**(2): 227–237.

28. Hajek A, Kretzler B, König HH. Multimorbidity, loneliness, and social isolation. A systematic review. *Int J Environ Res Public Health*. 2020; **17**(22): 8688.

29. Sharma A, Pillai DR, Lu M, et al. Impact of isolation precautions on quality of life: a meta-analysis. *J Hosp Infect*. 2020; **105**(1): 35–42.

30. Cadar D, Abell J, Matthews FE, et al. Cohort profile update: the harmonised cognitive assessment protocol sub-study of the English longitudinal study of ageing (ELSA-HCAP). *Int J Epidemiol*. 2021; **50**(3): 725–726i.

31. Steptoe A, Breeze E, Banks J, Nazroo J. Cohort profile: the English longitudinal study of ageing. *Int J Epidemiol*. 2013; **42**(6): 1640–1648.

32. Hyde M, Wiggins RD, Higgs P, Blane DB. A measure of quality of life in early old age: the theory, development and properties of a needs satisfaction model (CASP-19). *Aging Ment Health*. 2003; **7**(3): 186–194.

33. Fletcher AE, Hunt BM, Bulpitt CJ. Evaluation of quality of life in clinical trials of cardiovascular disease. *J Chronic Dis*. 1987; **40**(6): 557–566.

34. Shankar A, McMunn A, Banks J, Steptoe A. Loneliness, social isolation, and behavioral and biological health indicators in older adults.. *Health Psychol*. 2011; **30**(4): 377.

35. Kotwal AA, Holt-Lunstad J, Newmark RL, et al. Social isolation and loneliness among San Francisco Bay Area older adults during the COVID-19 shelter-in-place orders. *J Am Geriatr Soc*. 2021; **69**(1): 20–29.

36. Cook RD, Weisberg S. *Applied regression including computing and graphics*. New York: John Wiley & Sons; 2009.

Table 1 Bias, variance (VAR) and MSE values of the parameter estimates when $m = 5$ and $M = 50, 100$, or 200. The true parameter values are $\beta_{11} = 0.7071$, $\beta_{12} = -0.7071$, $\beta_{13} = 0$, $\beta_{21} = 0.5345$, $\beta_{22} = 0.2673$, $\beta_{23} = 0.8018$, $\sigma_{11} = 1$, $\sigma_{12} = 0.5$, and $\sigma_{22} = 1$.

	$M = 50$			$M = 100$			$M = 200$		
Parameter	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{11}	-0.0210	0.0264	0.0268	-0.0147	0.0159	0.0162	-0.0023	0.0044	0.0044
β_{12}	0.0308	0.0183	0.0192	0.0162	0.0122	0.0125	0.0090	0.0042	0.0043
β_{13}	0.0021	0.0276	0.0276	0.0022	0.0153	0.0153	0.0017	0.0073	0.0073
β_{21}	-0.0693	0.0607	0.0655	-0.0351	0.0287	0.0299	-0.0150	0.0102	0.0104
β_{22}	0.0003	0.0327	0.0327	-0.0024	0.0117	0.0117	-0.0047	0.0050	0.0050
β_{23}	-0.0313	0.0255	0.0265	-0.0104	0.0140	0.0141	-0.0013	0.0053	0.0053
σ_{11}	0.1094	0.3631	0.3750	0.0222	0.1050	0.1055	-0.0197	0.0554	0.0558
σ_{12}	-0.0247	0.1183	0.1189	-0.0522	0.0437	0.0464	-0.0275	0.0272	0.0279
σ_{22}	-0.0641	0.2158	0.2199	-0.1299	0.1077	0.1245	-0.0712	0.0637	0.0688

Table 2 Bias, variance (VAR) and MSE values of the parameter estimates when $m = 10$ and $M = 50, 100$, or 200. The true parameter values are $\beta_{11} = 0.7071$, $\beta_{12} = -0.7071$, $\beta_{13} = 0$, $\beta_{21} = 0.5345$, $\beta_{22} = 0.2673$, $\beta_{23} = 0.8018$, $\sigma_{11} = 1$, $\sigma_{12} = 0.5$, and $\sigma_{22} = 1$.

	$M = 50$			$M = 100$			$M = 200$		
Parameter	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_{11}	-0.0124	0.0148	0.0150	0.0078	0.0045	0.0046	-0.0073	0.0029	0.0029
β_{12}	0.0158	0.0108	0.0111	0.0200	0.0042	0.0046	-0.0002	0.0027	0.0027
β_{13}	-0.0113	0.0140	0.0141	-0.0037	0.0082	0.0083	0.0031	0.0045	0.0045
β_{21}	-0.0379	0.0379	0.0393	-0.0246	0.0143	0.0149	-0.0063	0.0063	0.0064
β_{22}	-0.0045	0.0124	0.0125	-0.0062	0.0053	0.0053	-0.0009	0.0027	0.0027
β_{23}	-0.0136	0.0130	0.0132	0.0026	0.0053	0.0053	-0.0029	0.0029	0.0030
σ_{11}	0.0364	0.1516	0.1529	-0.0259	0.0560	0.0566	0.0020	0.0420	0.0420
σ_{12}	-0.0346	0.0667	0.0679	-0.0293	0.0320	0.0328	-0.0204	0.0157	0.0161
σ_{22}	-0.0343	0.1354	0.1366	-0.0430	0.0678	0.0696	-0.0498	0.0335	0.0360

Figure 1 Histograms for the estimates of β_{11} when M and m change. The true value of β_{11} is labeled by the dashed vertical line.

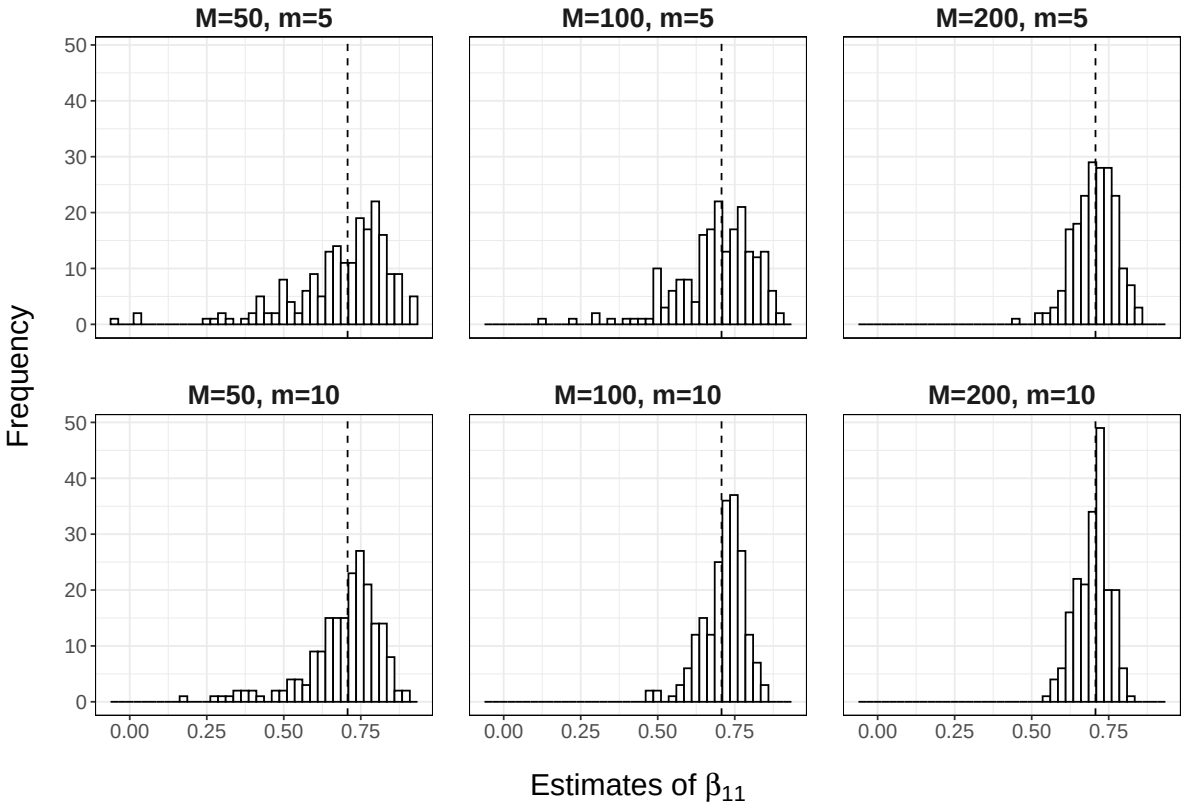

Figure 2 In each plot, the black solid line denotes the true link function ψ_1, the gray dashed line denotes its pointwise estimate, and the black dashed lines denote the 95% pointwise confidence interval. The results are based on 200 repeated simulations.

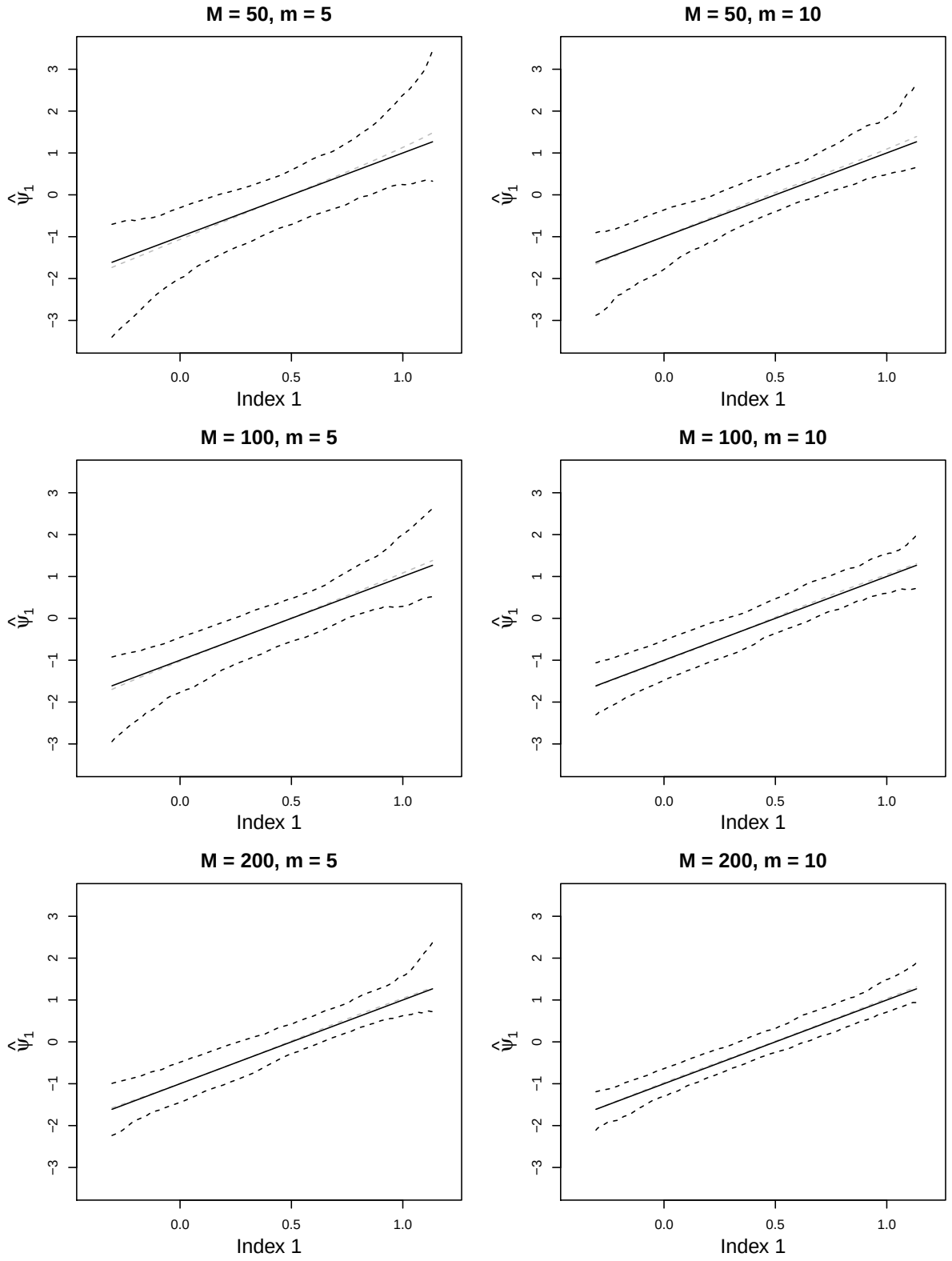

Figure 3 In each plot, the black solid line denotes the true link function ψ_2, the gray dashed line denotes its pointwise estimate, and the black dashed lines denote the 95% pointwise confidence interval. The results are based on 200 repeated simulations.

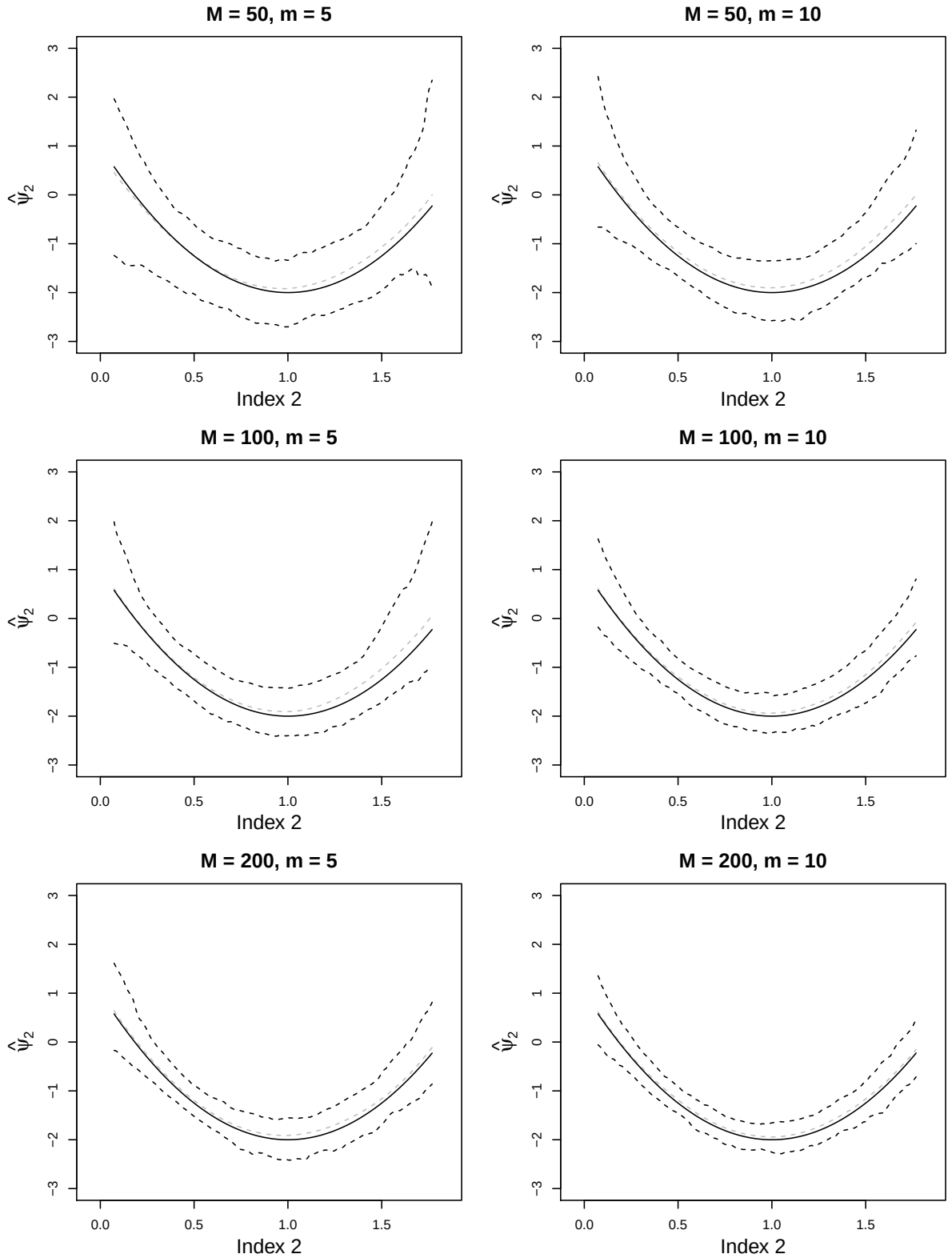

Table 3 Numbers presented in the table are $\sum_{p=1}^{3}(\hat{\beta}_{kp} - \beta_{kp})^2$, for $k = 1$ and 2, which are computed based on 200 repeated simulations in various cases considered. The smallest value obtained by the three methods in each case is presented in bold.

			rMAVE		EFM		PROPOSED	
Scenario	M	m	β_1	β_2	β_1	β_2	β_1	β_2
1	50	5	0.1556	0.1975	**0.0828**	**0.1892**	0.1575	0.1984
		10	0.0572	0.0891	**0.0455**	**0.0843**	0.0594	0.0942
	100	5	0.0486	0.0471	**0.0361**	**0.0463**	0.0490	0.0536
		10	**0.0169**	0.0192	0.0224	**0.0181**	0.0189	0.0202
	200	5	**0.0189**	0.0317	0.0205	**0.0300**	0.0196	0.0311
		10	**0.0095**	0.0101	0.0096	**0.0098**	0.0097	0.0121
2	50	5	0.1875	0.3856	0.1689	0.3660	**0.1663**	**0.3168**
		10	0.0599	0.1133	0.0604	0.1031	**0.0593**	**0.0940**
	100	5	0.1033	0.1022	0.0947	0.0972	**0.0866**	**0.0928**
		10	0.0263	0.0235	0.0324	0.0215	**0.0236**	**0.0203**
	200	5	0.0550	0.0294	0.0524	0.0271	**0.0476**	**0.0261**
		10	0.0130	0.0124	0.0214	0.0114	**0.0125**	**0.0138**
3	50	5	0.2258	0.3652	0.2228	**0.3352**	**0.2141**	0.3404
		10	0.0501	0.1105	0.0586	0.1091	**0.0442**	**0.0986**
	100	5	0.1013	0.0740	0.0926	0.0734	**0.0898**	**0.0670**
		10	0.0337	0.0507	0.0292	0.0458	**0.0278**	**0.0435**
	200	5	0.0280	0.0293	0.0256	0.0276	**0.0240**	**0.0256**
		10	0.0160	0.0145	0.0245	0.0136	**0.0137**	**0.0116**

Table 4 Estimated index coefficients and their standard errors (in parentheses) computed by a bootstrap procedure with 200 bootstrap samples. The two response variables are the status of cardiovascular diseases (CVD) and an indicator whether a participant has a larger-than-median quality of life score (HQOL).

Responses	Age	Loneliness score	Isolation score
Status of CVD	0.9946 (0.1382)	-0.0908 (0.1446)	-0.0494 (0.0290)
Indicator of HQOL	0.0952 (0.0930)	0.9278 (0.1042)	0.3608 (0.0512)

Figure 4 Estimated link functions $\widehat{\psi}_1$ and $\widehat{\psi}_2$ (left panels), and estimated probabilities of having cardiovascular diseases or a higher-than-median quality of life score (right panels). The observed binary response values are shown by the small circles in the two right panels.