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

DOI: xxx/xxxx

ORIGINAL ARTICLE

Multivariate single index modeling of longitudinal data with multiple 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 medical studies, composite indices and/or scores are routinely used for predicting medical conditions of patients. These indices are usually developed from observed data of certain disease risk factors, and it has been demonstrated in the literature that single index models can provide a powerful tool for this purpose. In practice, the observed data of disease risk factors are often longitudinal in the sense that they are collected at multiple time points for individual patients, and there are often multiple aspects of a patient's medical condition that are of our concern. However, most existing single index models are developed for cases with independent data and a single response variable, which are inappropriate for the problem just described in which within-subject observations are usually correlated and there are multiple mutually correlated response variables involved. This paper aims to fill this methodological gap by developing a single index model for analyzing longitudinal data with multiple responses. Both theoretical and numerical justifications show that the proposed new method provides an effective solution to the related research problem. It is also demonstrated using a dataset from the English Longitudinal Study of Aging.

KEYWORDS:
Asymptotic normality, EM algorithm, local linear kernel smoothing, mixed-effects modeling, multiple responses, single index model

1 | INTRODUCTION

Combining information from different aspects can reduce complexity in both feature selection and interpretation. While it may help gain feasibility in analysis, a composite score or index is argued to be more accessible for advocacy and political intervention decisions[1]. In health and clinical research, well-designed indices can be used to monitor and predict health outcomes of interest. For instance, World Health Organization developed the Urban Health Index that measured disparities in health determinants

and outcomes in the urban area and can utilize it to monitor the status of urban areas and determine the effects of program interventions[2]. In this paper, we focus on constructing such indices when multiple outcomes of interest need to be studied simultaneously and longitudinally.

To construct a composite index, it has been demonstrated in the literature that single index models can provide a powerful tool (e.g., Wu and Tu[3]). A single-index model is a semi-parametric model that can reduce the dimensionality of predictors and build a flexible relationship between the outcomes and predictors. It links the mean of the response to a linear combination of the predictors through an unknown nonparametric link function, where the linear combination of predictors is used to suppress the multidimensional predictors into a single index and avoid the so-called "curse of dimensionality". Moreover, the nonparametric link function is used to accommodate a potential nonlinear relationship between the outcome and the index. In the literature, there is an extensive discussion on the estimation of the index coefficients and the link function as well as the statistical properties of their estimators. See, for instance, Hardle and Stoker[4], Hardle et al.[5], Ichimura[6], Xia et al.[7], Xia[8], and Yu and Ruppert[9]. However, all these works were for cases with independent data and a single response variable.

In practice, especially in clinical research, repeated measurements are routinely taken for individual subjects. With the goal of incorporating within-subject serial correlation, investigators have extended conventional model-fitting approaches for independent data to the longitudinal setting. For instance, in linear regression modeling, mixed-effects models have been a popular tool for analyzing longitudinal data. By introducing subject-specific random effects, the resulting mixed-effects model allows subjects to have their own subject-specific mean trajectories over time[10]. Proper specification of the random effects can accommodate the covariance structure among the outcome variables as well. To take advantage of random-effects modeling, Pang and Xue[11] proposed a single-index model with random-effects terms included to accommodate within-subject data correlation, and used the generalized estimating equations (GEE) to estimate model coefficients. Since a working covariance structure should be specified in the GEE approach, their estimation of the random components was limited to the variance of a random intercept and the variance of the pure measurement error. In addition, Wu and Tu[3] extended the penalized spline estimation method originally proposed by Yu and Ruppert[9] to estimate their single-index model with random effects. Whilst penalized splines had a mixed-effects model representation with unpenalized (fixed) and penalized (random) components[12], all of the index-coefficient parameters, variance components, and the smoothing parameter were estimated directly by the restricted maximum likelihood method (REML).

In practice, there are often multiple response variables of interest. In some medical studies, for instance, they can measure different aspects of the medical condition of a patient. To accommodate their mutual correlation, Wu and Tu[13] further extended their penalized spline method to a multivariate setting. Given the same set of predictors for each response variable, their approach allows for different link functions used for modeling different response variables. However, the index coefficients for different response variables are assumed to be the same in their approach, which means that a common index is used for predicting different

response variables. While some well-defined medical indices like Body Mass Index (BMI) are routinely used to predict different response variables in practice for convenience, the assumption that different response variables depend on a set of disease risk factors through a common index could be questionable in some scenarios. As an example, assume that we want to predict the incidence of stroke and liver cancer (two response variables) from a set of disease risk factors like alcohol consumption, cholesterol level, systolic blood pressure, and more. In this example, the relative importance of individual disease risk factors should be quite different in predicting the two response variables. For instance, alcohol consumption might be more relevant to liver cancer although it is also an important risk factor of stroke. Similarly, cholesterol level and systolic blood pressure should be more relevant to stroke. In such cases, it might be more appropriate to use two different indices to predict the two response variables. In addition, the penalized spline method mentioned above needs a pre-specified set of knots for defining the basis functions used. Even if the penalty term can protect the procedure from over-fitting, the choice of the number and positions of the knots, which may be highly dependent on the true mean curve and variance structure, can be tricky and may lead to inefficient estimation of the index parameters.

In this paper, we propose a flexible single-index model to describe longitudinal data with multiple response variables. In our proposed model, the index coefficient parameters for different response variables could be different, and the within-subject correlation could also be accommodated by including random-effects terms in the model. Then, the model is estimated in the context of a multivariate single-index model with random effects. Namely, the index coefficient parameters are estimated by combining the ideas of the refined conditional minimum average variance estimation (rMAVE) method proposed by Xia et al.[7] and the expectation-maximization (EM) algorithm introduced by Dempster et al.[14], and the link functions for individual response variables are estimated locally by the local linear kernel smoothing method. To our knowledge, this is the first multivariate single-index model that can allow different sets of index coefficient parameters for different response variables and accommodate within-subject data correlation as well. Both theoretical and numerical justifications show that it is an effective single-index model for describing longitudinal data with multiple response variables.

The rest of the paper is organized as follows. Section 2 describes specification and estimation of the proposed multivariate single-index model with random effects. Some asymptotic results about the estimated model are given in Section 3. Simulation studies evaluating its numerical performance are presented in Section 4. In Section 5, we apply the proposed method to a dataset from the English Longitudinal Study of Aging (ELSA). Section 6 concludes the article with some remarks.

2 | METHODOLOGY

2.1 | Model specification

In cases with a univariate response variable Y and a p-dimensional vector of predictors $\mathbf{X} \in \mathbb{R}^p$, a single-index model takes the form of $\mathbb{E}(Y|\mathbf{X}) = \psi(\boldsymbol{\beta}^T\mathbf{X})$, where $\boldsymbol{\beta}$ is the p-dimensional vector of coefficients and $\psi(\cdot)$ is an unknown smooth link function. For model identifiability reasons, one often assumes that $\boldsymbol{\beta}^T\boldsymbol{\beta} = 1$ and the first element of $\boldsymbol{\beta}$ is nonnegative. While $\boldsymbol{\beta}^T\mathbf{X}$ suppresses the multidimensional vector $\mathbf{X}$ into a single index of dimension one, the link function $\psi(\cdot)$ preserves some flexibility in the relationship between the response variable and the constructed index.

Now, we generalize the univariate single-index model to a multivariate one with random-effects to describe longitudinal data with multiple response variables. Assume that there are a total of M subjects included in the longitudinal data. For the i-th subject, m_i repeated measurements are taken on both the q response variables and p predictors. We want to use the data of the p predictors to construct q single indices for the q response variables. Let Y_{ijk} be the observed k-th response variable of the i-th subject at the j-th observation time $t_{ij} \in [T_0, T_1]$, for $i = 1, \ldots, M$, $j = 1, \ldots, m_i$, and $k = 1, \ldots, q$. It is assumed that the observation times are independent not only with each other but also with both the observed response variables and predictors. The observed vector of predictors for the i-th subject at the j-th observation time is denoted as $\mathbf{X}_{ij} = (X_{ij1}, \ldots, X_{ijp})^T \in \mathbb{R}^p$. Then, the proposed multivariate single-index model with random effects is defined to be

$$Y_{ijk} = \psi_k(\boldsymbol{\beta}_k^T\mathbf{X}_{ij}) + \mathbf{b}_{ik}^T\mathbf{g}(t_{ij}) + \varepsilon_{ijk}, \tag{1}$$

for $i = 1, \ldots, M$, $j = 1, \ldots, m_i$, and $k = 1, \ldots, q$. In Model (1), $\psi_k(\cdot)$ is the link function for the k-th response variable, $\boldsymbol{\beta}_k$ is a p-dimensional vector of index coefficients that satisfies the identifiability conditions mentioned above, $\mathbf{b}_{ik} = (b_{ik1}, \ldots, b_{iks})^T$ is the vector of random-effects corresponding to the k-th response variable of the i-th subject, $\mathbf{g}(t_{ij}) = [g_1(t_{ij}), \ldots, g_s(t_{ij})]^T$ is a vector of pre-specified functions of t_{ij} that potentially accounts for the complexity in the subject-specific trajectory of the mean response over time, and ε_{ijk} is the pure measurement error. For simplicity and adequacy, we consider $s = 2$ and $\mathbf{g}(t_{ij}) = (1, t_{ij})^T$, which results in the commonly used random intercept and random slope random-effects model. Among the limited literature on the single-index models with random effects, no existing work considered random effects more than the random intercept. Let $\mathbf{b}_i = (\mathbf{b}_{i1}^T, \ldots, \mathbf{b}_{iq}^T)^T$ be the vector of all random effects for the i-th subject. Then, it is assumed that $\mathbf{b}_i$ follows a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}_b$. For the i-th subject at the j-th observation time, the correlation between the k-th and k'-th response variables is induced by the correlation between $\mathbf{b}_{ik}$ and $\mathbf{b}_{ik'}$. For different subjects, the vectors of random effects are assumed to be independent. For all i, j and k, the pure measurement errors $\{\varepsilon_{ijk}\}$ are assumed to be independent with each other, and each of them follow a normal distribution with mean zero and an outcome-specific variance σ_k^2. The independence between random effects and pure measurement errors is also assumed.

2.2 | Model estimation

In Model (1), the index coefficients $\boldsymbol{\beta}_k$ and the link functions $\psi_k(\cdot)$ are the two main parts to be estimated. In this section, we describe how to estimate $\boldsymbol{\beta}_k$ and $\psi_k(\cdot)$ by combining the ideas of the EM algorithm and an extended version of the rMAVE method. The EM algorithm is a commonly used tool to numerically fit mixed-effects models[15]. More specifically, we can regard the random effects as unobserved data and write out the log-likelihood for the complete data $l_c(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{b})$, where $\mathbf{Y}$ is the vector of the observed data, $\mathbf{b}$ is the vector of the random effects, and $\boldsymbol{\theta}$ is the vector of all unknown parameters. Then, we iteratively update the parameter estimates 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 convergence, where $\boldsymbol{\theta}^*$ denotes the parameter estimates obtained in the previous iteration.

In our setting, let $\mathbf{Y}_{i\bullet k} = (Y_{i1k}, \ldots, Y_{im_ik})^T$ be the observations of subject i on the k-th response variable, $\mathbf{Y}_{i\bullet\bullet} = (\mathbf{Y}_{i\bullet 1}^T, \ldots, \mathbf{Y}_{i\bullet q}^T)^T$ be the vector of all observations of the i-th subject, $\mathbf{Y}_{\bullet\bullet k} = (\mathbf{Y}_{1\bullet k}^T, \ldots, \mathbf{Y}_{M\bullet k}^T)^T$ be the vector of all observations on the k-th response variable, 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. Denote $\boldsymbol{\theta}$ as a collection of all unknown parameters $(\{\boldsymbol{\beta}_k\}, \{\sigma_k^2\}, \boldsymbol{\Sigma}_b)$ and the unknown link functions $\{\psi_k(\cdot)\}$. Then, the log-likelihood of the complete data has the following expression:

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

where C denotes the terms omitted that have nothing to do with $\boldsymbol{\theta}$, and $|\boldsymbol{\Sigma}_b|$ denotes the determinant of $\boldsymbol{\Sigma}_b$.

If the link functions are known, then we can use the conventional EM algorithm directly by updating the parameter estimates iteratively using the following formulas: for $k = 1, \ldots, q$,

$$\widehat{\boldsymbol{\beta}}_k = \underset{\boldsymbol{\beta}_k}{\operatorname{argmax}}\, \mathbb{E}_{\mathbf{b}|\mathbf{Y},\widehat{\boldsymbol{\Sigma}}_b,\{\widehat{\sigma}_k^2\},\{\widetilde{\boldsymbol{\beta}}_k\}}\left\{\log f(\mathbf{Y}|\mathbf{b}, \boldsymbol{\beta}_k, \hat{\sigma}_k^2)\middle|\mathbf{Y}, \widehat{\boldsymbol{\Sigma}}_b, \{\widehat{\sigma}_k^2\}, \{\widetilde{\boldsymbol{\beta}}_k\}\right\}, \tag{3}$$

$$\widehat{\sigma}_k^2 = \underset{\sigma_k^2}{\operatorname{argmax}}\, \mathbb{E}_{\mathbf{b}|\mathbf{Y},\widehat{\boldsymbol{\Sigma}}_b,\{\widetilde{\sigma}_k^2\},\{\widehat{\boldsymbol{\beta}}_k\}}\left\{\log f(\mathbf{Y}|\mathbf{b}, \widehat{\boldsymbol{\beta}}_k, \sigma_k^2)\middle|\mathbf{Y}, \widehat{\boldsymbol{\Sigma}}_b, \{\widetilde{\sigma}_k^2\}, \{\widehat{\boldsymbol{\beta}}_k\}\right\}, \tag{4}$$

$$\widehat{\boldsymbol{\Sigma}}_b = \underset{\boldsymbol{\Sigma}_b}{\operatorname{argmax}}\, \mathbb{E}_{\mathbf{b}|\mathbf{Y},\widetilde{\boldsymbol{\Sigma}}_b,\{\widehat{\sigma}_k^2\},\{\widehat{\boldsymbol{\beta}}_k\}}\left\{\log f(\mathbf{b}|\boldsymbol{\Sigma}_b)\middle|\mathbf{Y}, \widetilde{\boldsymbol{\Sigma}}_b, \{\widehat{\sigma}_k^2\}, \{\widehat{\boldsymbol{\beta}}_k\}\right\}, \tag{5}$$

where

$$\log f(\mathbf{Y}|\mathbf{b},\boldsymbol{\beta}_k,\sigma_k^2) = \sum_{i=1}^{M}\sum_{j=1}^{m_i}\left[-\frac{1}{2}\log(\sigma_k^2) - \frac{1}{2\sigma_k^2}\{Y_{ijk} - \psi_k(\boldsymbol{\beta}_k^T\mathbf{X}_{ij}) - \mathbf{g}(t_{ij})^T\mathbf{b}_{ik}\}^2\right],$$

$$\log f(\mathbf{b}|\boldsymbol{\Sigma}_b) = \sum_{i=1}^{M}\left\{-\frac{1}{2}\log|\boldsymbol{\Sigma}_b| - \frac{1}{2}\mathbf{b}_i^T\boldsymbol{\Sigma}_b^{-1}\mathbf{b}_i\right\},$$

$\widehat{\boldsymbol{\beta}}_k$, $\widehat{\sigma}_k^2$ and $\widehat{\boldsymbol{\Sigma}}_b$ are the parameter estimates in the current iteration, $\widetilde{\boldsymbol{\beta}}_k$, $\widetilde{\sigma}_k^2$ and $\widetilde{\boldsymbol{\Sigma}}_b$ are the parameter estimates obtained in the previous iteration.

To accommodate the estimation of the link functions, Equation (3) is modified into a two-step procedure that updates the estimates of $\{\boldsymbol{\beta}_k\}$ and $\{\psi_k(\cdot)\}$ separately. Following the idea of rMAVE, for any given k and $\mathbf{X}_{i'j'}$, $\mathbb{E}(Y_{ijk}|\mathbf{X}_{ij})$ can be approximated by a linear expansion at $\boldsymbol{\beta}_k^T\mathbf{X}_{i'j'}$, i.e., $\mathbb{E}(Y_{ijk}|\mathbf{X}_{ij}) \approx a_{i'j'k} + c_{i'j'k}\boldsymbol{\beta}_k^T(\mathbf{X}_{ij} - \mathbf{X}_{i'j'})$. Then, $\mathbb{E}\{Y_{ijk} - \mathbf{g}(t_{ij})^T\mathbf{b}_{ik} - \mathbb{E}(Y_{ijk}|\boldsymbol{\beta}_k^T\mathbf{X}_{ij})\}^2$ can be approximated by

$$\sum_{i,i'=1}^{M}\sum_{j=1}^{m_i}\sum_{j'=1}^{m_{i'}}\left[Y_{ijk} - \mathbf{g}(t_{ij})^T\mathbf{b}_{ik} - \{a_{i'j'k} + c_{i'j'k}\boldsymbol{\beta}_k^T(\mathbf{X}_{ij} - \mathbf{X}_{i'j'})\}\right]^2 w_{iji'j'},$$

where

$$w_{iji'j'} = \frac{K_{h_{1k}}\{\boldsymbol{\beta}_k^T(\mathbf{X}_{ij} - \mathbf{X}_{i'j'})\}}{\sum_{i=1}^{M}\sum_{j=1}^{m_i}K_{h_{1k}}\{\boldsymbol{\beta}_k^T(\mathbf{X}_{ij} - \mathbf{X}_{i'j'})\}},$$

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

Notably, minimizing $\mathbb{E}\{Y_{ijk} - \mathbf{g}(t_{ij})^T\mathbf{b}_{ik} - \mathbb{E}(Y_{ijk}|\boldsymbol{\beta}_k^T\mathbf{X}_{ij})\}^2$ with respect to each $\boldsymbol{\beta}_k$ is equivalent to optimizing the complete data log-likelihood. Therefore, under the EM framework, for the k-th response variable, we update the estimates of $\{\boldsymbol{\beta}_k\}$ by minimizing the following loss function with respect to $\boldsymbol{\beta}_k$, $\mathbf{a}_k$, and $\mathbf{c}_k$:

$$\sum_{i,i'=1}^{M}\sum_{j=1}^{m_i}\sum_{j'=1}^{m_{i'}}\mathrm{E}_{\mathbf{b}|\mathbf{Y}}\left([Y_{ijk} - g(t_{ij})^T\mathbf{b}_{ik} - \{a_{i'j'k} + c_{i'j'k}\boldsymbol{\beta}_k^T(\mathbf{X}_{ij} - \mathbf{X}_{i'j'})\}]^2 w_{iji'j'}|\mathbf{Y},\widehat{\boldsymbol{\theta}}\right), \tag{6}$$

where $\mathbf{a}_k$ and $\mathbf{c}_k$ are the vectors of $a_{i'j'k}$ and $c_{i'j'k}$, respectively, that permute i' and j' in the same way as $\mathbf{Y}_k$, and the expectation is over the posterior distribution of $\mathbf{b}$ that depends on the collection of current estimates of all parameters and the estimates of the link functions to be discussed soon. For convenience, we use $\widehat{\boldsymbol{\theta}}$ to denote the collection of all current parameter estimates.

After obtaining the estimates of $\{\boldsymbol{\beta}_k\}$, denoted as $\{\widehat{\boldsymbol{\beta}}_k\}$, a vector of single indices corresponding to the $\sum_{i=1}^{M} m_i$ observations can be computed for each response variable. Similar to the definition of the vectors of the observed response variables, let $\boldsymbol{\Psi}_{i\bullet k} = [\psi_k(\widehat{\boldsymbol{\beta}}_k^T\mathbf{X}_{i1}),\ldots,\psi_k(\widehat{\boldsymbol{\beta}}_k^T\mathbf{X}_{im_i})]^T$ be the vector of mean response for the k-th response variable of the i-th subject conditional on the estimates of the single indices, $\boldsymbol{\Psi}_{i\bullet\bullet} = (\boldsymbol{\Psi}_{i\bullet 1}^T,\ldots,\boldsymbol{\Psi}_{i\bullet q}^T)^T$ be the vector of all conditional mean responses for the i-th subject, and $\boldsymbol{\Psi}_{\bullet\bullet k} = (\boldsymbol{\Psi}_{1\bullet k}^T,\ldots,\boldsymbol{\Psi}_{M\bullet k}^T)^T$ be the vector of all conditional mean responses for the k-th response variable. For $k = 1,\ldots,q$, we estimate all individual elements of $\boldsymbol{\Psi}_{\bullet\bullet k}$ by using the local linear kernel smoothing procedure. More specifically, let $\mathbf{X}$ be the $(\sum_{i=1}^{M} m_i)\times p$ matrix of all covariate data whose $[\sum_{l=1}^{i-1} m_l + j]$-th row is $\mathbf{X}_{ij}^T$, for $1 \le j \le m_i$ and $1 \le i \le M$. Then, for a given u

in the range of all elements of $\mathbf{X}\widehat{\boldsymbol{\beta}}_k$, the local linear kernel estimator of $\psi_k(u)$ is given by

$$\widehat{\psi}_k(u) = \mathbf{e}_1^T \left\{ \sum_{i=1}^{M} \sum_{j=1}^{m_i} K_{h_{2k}}(\widehat{\boldsymbol{\beta}}_k^T \mathbf{X}_{ij} - u) \begin{pmatrix} 1 \\ \widehat{\boldsymbol{\beta}}_k^T \mathbf{X}_{ij} - u \end{pmatrix} \begin{pmatrix} 1 \\ \widehat{\boldsymbol{\beta}}_k^T \mathbf{X}_{ij} - u \end{pmatrix}^T \right\}^{-1} \times \sum_{i=1}^{M} \sum_{j=1}^{m_i} K_{h_{2k}}(\widehat{\boldsymbol{\beta}}_k^T \mathbf{X}_{ij} - u) \begin{pmatrix} 1 \\ \widehat{\boldsymbol{\beta}}_k^T \mathbf{X}_{ij} - u \end{pmatrix} Y_{ijk}, \tag{7}$$

where $\mathbf{e}_1 = (1, 0)^T$, and h_{2k} is a bandwidth chosen properly to accommodate the correlated data $\mathbf{Y}_{\bullet\bullet k}$, for $k = 1, \ldots, q$. Details about bandwidth selection will be discussed in the next subsection.

After $\{\widehat{\boldsymbol{\beta}}_k\}$ and $\{\widehat{\boldsymbol{\Psi}}_{\bullet\bullet k}\}$ are computed, the posterior distribution of $\mathbf{b}$ and the estimates of σ_k^2 and $\boldsymbol{\Sigma}_b$ can be updated using Equations (4) and (5). Practically, most optimization procedures mentioned above can be simplified using closed-form formulae. For clarity, our proposed model estimation method is summarized in the following algorithm.

1. Obtain estimates of the index coefficient parameters and variance components for each response variable, say $\{\widehat{\boldsymbol{\alpha}}_k\}$, $\{\widehat{\sigma}_k^2\}$ and $\{\widehat{\boldsymbol{\Sigma}}_{bk}\}$, by independently fitting q linear mixed effects models with random intercept and random slope across time. Then, define $\widehat{\boldsymbol{\beta}}_k^{(0)} = \widehat{\boldsymbol{\alpha}}_k / ||\widehat{\boldsymbol{\alpha}}_k||$ and $\widehat{\sigma}_k^{2(0)} = \widehat{\sigma}_k^2$ as the initial value of $\boldsymbol{\beta}_k$ and σ_k^2, respectively, for $k = 1, \ldots, q$, where $||\cdot||$ denotes the L_2-norm of the underlying vector. Set the initial values of $\boldsymbol{\Sigma}_b$ as $\widehat{\boldsymbol{\Sigma}}_b^{(0)} = \text{diag}\{\widehat{\boldsymbol{\Sigma}}_{b1}, \ldots, \widehat{\boldsymbol{\Sigma}}_{bq}\}$ which is a $2q \times 2q$ block diagonal matrix of $\{\widehat{\boldsymbol{\Sigma}}_{bk}\}$.

2. Use the estimates obtained from the $(r-1)$-th iteration to update estimates of $\{\boldsymbol{\beta}_k\}$ and $\{\boldsymbol{\Psi}_{\bullet\bullet k}\}$ evaluated at the current indices vectors $\{\mathbf{X}\widehat{\boldsymbol{\beta}}_k^{(r)}\}$, as described below.

   2.1 Use $\{\widehat{\boldsymbol{\beta}}_k^{(r-1)}\}$ as the initial value to iteratively solve the following equation system: for $k = 1, \ldots, q$,

$$\begin{aligned} \begin{pmatrix} \widehat{a}_{i'j'k} \\ \widehat{c}_{i'j'k} \end{pmatrix} = & \left[\sum_{i=1}^{M} \sum_{j=1}^{m_i} K_{h_{1k}}\{\widehat{\boldsymbol{\beta}}_k^T(\mathbf{X}_{ij} - \mathbf{X}_{i'j'})\} \begin{pmatrix} 1 \\ \widehat{\boldsymbol{\beta}}_k^T(\mathbf{X}_{ij} - \mathbf{X}_{i'j'}) \end{pmatrix} \begin{pmatrix} 1 \\ \widehat{\boldsymbol{\beta}}_k^T(\mathbf{X}_{ij} - \mathbf{X}_{i'j'}) \end{pmatrix}^T \right]^{-1} \\ & \times \sum_{i=1}^{M} \sum_{j=1}^{m_i} K_{h_{1k}}\{\widehat{\boldsymbol{\beta}}_k^T(\mathbf{X}_{ij} - \mathbf{X}_{i'j'})\} \begin{pmatrix} 1 \\ \widehat{\boldsymbol{\beta}}_k^T(\mathbf{X}_{ij} - \mathbf{X}_{i'j'}) \end{pmatrix} (Y_{ijk} - \widehat{z}_{ijk}^{(r-1)}), \\ & \text{for } i' = 1, \ldots, M, j' = 1, \ldots, m_i, \end{aligned} \tag{8}$$

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_k = & \left[\sum_{i,i'=1}^{M} \sum_{j=1}^{m_i} \sum_{j'=1}^{m_{i'}} K_{h_{1k}}\{\widehat{\boldsymbol{\beta}}_k^T(\mathbf{X}_{ij} - \mathbf{X}_{i'j'})\} \widehat{c}_{i'j'k}^2 (\mathbf{X}_{ij} - \mathbf{X}_{i'j'})(\mathbf{X}_{ij} - \mathbf{X}_{i'j'})^T / \hat{f}_{\beta_k,\mathbf{x}}(\widehat{\boldsymbol{\beta}}_k^T \mathbf{X}_{i'j'}) \right]^{-1} \\ & \times \sum_{i,i'=1}^{M} \sum_{j=1}^{m_i} \sum_{j'=1}^{m_{i'}} K_{h_{1k}}\{\widehat{\boldsymbol{\beta}}_k^T(\mathbf{X}_{ij} - \mathbf{X}_{i'j'})\} \widehat{c}_{i'j'k} (\mathbf{X}_{ij} - \mathbf{X}_{i'j'})(Y_{ijk} - \widehat{a}_{i'j'k} - \widehat{z}_{ijk}^{(r-1)}) / \hat{f}_{\beta_k,\mathbf{x}}(\widehat{\boldsymbol{\beta}}_k^T \mathbf{X}_{i'j'}), \\ & \text{with } \hat{f}_{\beta_k,\mathbf{x}}(\widehat{\boldsymbol{\beta}}_k^T \mathbf{X}_{i'j'}) = \frac{1}{\sum_{i=1}^{M} m_i} \sum_{i=1}^{M} \sum_{j=1}^{m_i} K_{h_{1k}}\{\widehat{\boldsymbol{\beta}}_k^T(\mathbf{X}_{ij} - \mathbf{X}_{i'j'})\}. \end{aligned} \tag{9}$$

Here, $\{h_{1k}\}$ are the bandwidths that need to be selected properly once $\{\widehat{\boldsymbol{\beta}}_k\}$ are updated, $\widehat{z}_{ijk}^{(r-1)} = \mathbf{g}(t_{ij})^T \mathbb{E}(\mathbf{b}_{ik}|\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(r-1)})$, $\widehat{\boldsymbol{\theta}}^{(r-1)}$ is the collection of all parameter estimates and the link function estimates obtained from the $(r-1)$-th iteration, and $\mathrm{E}(\mathbf{b}_{ik}|\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(r-1)})$ is a vector consisting of the $(2k-1)$-th to $2k$-th elements of the vector

$$\mathbb{E}(\mathbf{b}_i|\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(r-1)}) = \widehat{\boldsymbol{\Sigma}}_b^{(r-1)} \mathbf{Z}_i^T (\mathbf{Z}_i \widehat{\boldsymbol{\Sigma}}_b^{(r-1)} \mathbf{Z}_i^T + \widehat{\boldsymbol{\Lambda}}_i^{(r-1)})^{-1} (\mathbf{Y}_{i\bullet\bullet} - \widehat{\boldsymbol{\Psi}}_{i\bullet\bullet}^{(r-1)}),$$

where $\mathbf{Z}_i$ is the design matrix of $\mathbf{b}_i$, which is a block diagonal matrix with q blocks, i.e., $\mathbf{Z}_i = \mathrm{diag}\{\mathbf{G}_i, \ldots, \mathbf{G}_i\}$ where $\mathbf{G}_i = [\mathbf{g}(t_{i1}), \ldots, \mathbf{g}(t_{im_i})]^T$, and $\widehat{\boldsymbol{\Lambda}}_i^{(r-1)}$ is a diagonal matrix with diagonal elements $(\widehat{\sigma}_1^{2(r-1)} \mathbf{1}_{m_i}^T, \ldots, \widehat{\sigma}_q^{2(r-1)} \mathbf{1}_{m_i}^T)$ in which $\mathbf{1}_{m_i}$ is a vector of ones with length m_i. Then, we define $\{\widehat{\boldsymbol{\beta}}_k^{(r)}\}$ to be the convergent values of $\widehat{\boldsymbol{\beta}}_k$ in (9), for $k = 1, \ldots, q$. Note that $\{\widehat{\boldsymbol{\beta}}_k\}$ should be standardized in each iteration when solving the equation system (8) and (9).

2.2 Update the indices $\{\mathbf{X}\widehat{\boldsymbol{\beta}}_k^{(r)}\}$ and apply the local linear kernel smoothing procedure described in (7) to obtain $\widehat{\boldsymbol{\Psi}}_{\bullet\bullet k}^{(r)}$, for $k = 1, \ldots, q$.

3. Use the estimates from step 2 to compute

$$\begin{aligned}\widehat{\sigma}_k^{2(r)} = \frac{1}{\sum_{i=1}^M m_i} \Bigg[&(\mathbf{Y}_{\bullet\bullet k} - \widehat{\boldsymbol{\Psi}}_{\bullet\bullet k}^{(r)})^T (\mathbf{Y}_{\bullet\bullet k} - \widehat{\boldsymbol{\Psi}}_{\bullet\bullet k}^{(r)}) + \sum_{i=1}^M \mathrm{tr}\{\mathbf{G}_i^T \mathbf{G}_i \, \mathbb{E}(\mathbf{b}_{ik}\mathbf{b}_{ik}^T|\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(r-1)*})\} \\ &-2 \sum_{i=1}^M (\mathbf{Y}_{i\bullet k} - \widehat{\boldsymbol{\Psi}}_{i\bullet k}^{(r)})^T \mathbf{G}_i \, \mathbb{E}(\mathbf{b}_{ik}|\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(r-1)*}) \Bigg], \text{ for } k = 1, \ldots, q, \end{aligned} \tag{10}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix,

$$\mathbb{E}(\mathbf{b}_{ik}\mathbf{b}_{ik}^T|\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(r-1)*}) = \mathrm{Var}(\mathbf{b}_{ik}|\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(r-1)*}) + \mathbb{E}(\mathbf{b}_{ik}|\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(r-1)*}) \, \mathbb{E}(\mathbf{b}_{ik}|\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(r-1)*})^T$$

with $\mathbb{E}(\mathbf{b}_{ik}|\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(r-1)*})$ and $\mathrm{var}(\mathbf{b}_{ik}|\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(r-1)*})$ obtained respectively from the proper entries of

$$\mathbb{E}(\mathbf{b}_i|\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(r-1)*}) = \widehat{\boldsymbol{\Sigma}}_b^{(r-1)} \mathbf{Z}_i^T (\mathbf{Z}_i \widehat{\boldsymbol{\Sigma}}_b^{(r-1)} \mathbf{Z}_i^T + \widehat{\boldsymbol{\Lambda}}_i^{(r-1)})^{-1} (\mathbf{Y}_i - \widehat{\boldsymbol{\Psi}}_i^{(r)}), \text{ and}$$

$$\mathrm{Var}(\mathbf{b}_i|\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(r-1)*}) = \{(\widehat{\boldsymbol{\Sigma}}_b^{(r-1)})^{-1} + \mathbf{Z}_i^T (\widehat{\boldsymbol{\Lambda}}_i^{(r-1)})^{-1} \mathbf{Z}_i\}^{-1}.$$

4. Update the estimates of the covariance matrix of the random effects via

$$\widehat{\boldsymbol{\Sigma}}_b^{(r)} = \frac{1}{M} \sum_{i=1}^M \left[\mathrm{var}(\mathbf{b}_i|\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(r-1)**}) + \mathbb{E}(\mathbf{b}_i|\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(r-1)**}) \, \mathbb{E}(\mathbf{b}_i|\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(r-1)**})^T \right], \tag{11}$$

where

$$\mathbb{E}(\mathbf{b}_i|\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(r-1)**}) = \widehat{\boldsymbol{\Sigma}}_b^{(r-1)} \mathbf{Z}_i^T (\mathbf{Z}_i \widehat{\boldsymbol{\Sigma}}_b^{(r-1)} \mathbf{Z}_i^T + \widehat{\boldsymbol{\Lambda}}_i^{(r)})^{-1} (\mathbf{Y}_{i\bullet\bullet} - \widehat{\boldsymbol{\Psi}}_{i\bullet\bullet}^{(r)}), \text{ and}$$

$$\mathrm{Var}(\mathbf{b}_i|\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(r-1)**}) = \{(\widehat{\boldsymbol{\Sigma}}_b^{(r-1)})^{-1} + \mathbf{Z}_i^T (\widehat{\boldsymbol{\Lambda}}_i^{(r)})^{-1} \mathbf{Z}_i\}^{-1}.$$

5. Repeat steps 2, 3, and 4 until convergence of all parameter estimates. The final estimates of the link functions are based on indices computed from the final estimates of $\{\boldsymbol{\beta}_k\}$.

Derivation of Equations (8)-(11) is provided in the supplementary materials.

Since the above model estimation method estimates the index coefficients, variance components and link functions simultaneously, its computational cost would be higher than the previous methods like rMAVE that assume data independence. More specifically, in each iteration of the proposed iterative model estimation algorithm for updating all parameter estimates in concern, updating the q sets of index coefficients can be regarded as q implementations of the conventional rMAVE approach. Note that the closed-form formulae have been obtained for updating the variance components. So, updating $\widehat{\sigma}_k^2$'s and $\widehat{\boldsymbol{\Sigma}}_b$ in each iteration is actually quite straightforward. But, calculation of the inverse of some $2q \times 2q$ matrices could be computationally intensive when q is large. Therefore, the current version of the proposed method is mainly for cases when q is small (e.g., $q \leq 10$). In cases when q is large, it might be helpful to incorporate a variable selection procedure in the proposed method to reduce the dimensionality of the response variables, which is left to our future research.

2.3 | Selection of the kernel function and the bandwidths

Since the local linear kernel smoothing procedure is used when updating the coefficients $\{\boldsymbol{\beta}_k\}$ and the vectors of the conditional mean responses $\{\boldsymbol{\Psi}_{\cdot\cdot k}\}$, the kernel functions $K(\cdot)$ should be specified in advance, and the bandwidths $\{h_{1k}\}$ and $\{h_{2k}\}$ should be selected properly based on the updated indices $\{\mathbf{X}\widehat{\boldsymbol{\beta}}_k\}$.

Given the good theoretical properties of the Epanechnikov kernel function illustrated in the kernel smoothing literature[16], $K(\cdot)$ is chosen to be that kernel function, which takes the form of $K(x) = 0.75(1-x^2)\mathbf{I}(|x| \leq 1)$. For the k-th response variable, the bandwidth h_{1k} is used when solving the minimization problem (6), which is a modified version of the loss function used in rMAVE. Empirically, the difference between the random effects $\mathbf{b}_i$ and $\mathbb{E}(\mathbf{b}_i|\mathbf{Y}, \widehat{\boldsymbol{\theta}})$ becomes negligible as the algorithm runs. Therefore, for each k, we view $Y_{ijk} - \mathbf{g}(t_{ij})^T\, \mathbb{E}(\mathbf{b}_{ik}|\mathbf{Y}, \widehat{\boldsymbol{\theta}})$ as independent observations, for $i = 1, \ldots, M$ and $j = 1, \ldots, m_i$, and choose the bandwidth h_{1k} simply by the rule of thumb provided in Mack and Silverman[17]. This was also suggested by Xia[8] when they implemented the rMAVE method for independent data. In practice, it is not feasible to choose $\{h_{1k}\}$ by using a cross-validation (CV) procedure because the bandwidths $\{h_{1k}\}$ need to be updated once the indices are changed due to the new estimates of $\{\boldsymbol{\beta}_k\}$ in each sub-iteration described in step 2.1 of the proposed algorithm discussed in the previous subsection.

The bandwidth h_{2k} is used in the estimation of the conditional mean response $\{\boldsymbol{\Psi}_{\cdot\cdot k}\}$. We choose $\{h_{2k}\}$ by applying a modified CV (MCV) procedure suggested by De Brabanter et al.[18], which can accommodate potential correlation among the observed data by using a bimodal kernel function. Namely, for the k-th response variable, h_{2k} is chosen by minimizing the MCV score

below.

$$\mathrm{MCV}(h_{2k}) = \frac{1}{M}\sum_{i=1}^{M}\left[\frac{1}{m_i}\sum_{j=1}^{m_i}\left\{\widehat{\psi}_{k,-(ij)}(\widehat{\boldsymbol{\beta}}_k^T\mathbf{X}_{ij}) - Y_{ijk}\right\}^2\right], \tag{12}$$

where $\widehat{\psi}_{k,-(ij)}(\widehat{\boldsymbol{\beta}}_k^T\mathbf{X}_{ij})$ is the leave-one-out estimate of $\psi_k(\widehat{\boldsymbol{\beta}}_k^T\mathbf{X}_{ij})$ obtained by a modified version of (7) in which the observation $(Y_{ijk}, \widehat{\boldsymbol{\beta}}_k^T\mathbf{X}_{ij})$ is omitted in the computation and $K(\cdot)$ is changed to

$$K_\epsilon(u) = \frac{4}{4-3\epsilon-\epsilon^3}\begin{cases}\frac{3}{4}(1-u^2)\mathbf{I}(|u|\le 1), & \text{if } |u|\ge\epsilon,\\ \frac{3(1-\epsilon^2)}{4\epsilon}|u|, & \text{otherwise,}\end{cases} \tag{13}$$

where $\epsilon \in (0,1)$ is a pre-specified constant. We use $\epsilon = 0.1$ by adopting the suggestion in De Brabanter et al.[18] which was justified there by numerical studies.

3 | ASYMPTOTIC RESULTS

In this section, we establish the asymptotic normality of the estimated index coefficient parameters for different response variables. Following Jiang and Wang[19], for $k = 1,\ldots,q$, the initial value $\boldsymbol{\beta}_k^{(0)}$ is assumed to be in the $\sqrt{n}-$neighborhood of the true value $\boldsymbol{\beta}_k$. The kernel function $K(\cdot)$ is assumed to be a symmetric density function. For simplicity, it is also assumed that $\int u^2K(u)du = 1$. This can be achieved by any symmetric density kernel function after proper normalization. For the k-th response variable, let $\boldsymbol{\mu}_{\beta_k,\mathbf{x}}(\mathbf{x}) = \mathbb{E}(\mathbf{X}_{ij}|\boldsymbol{\beta}_k^T\mathbf{X}_{ij} = \boldsymbol{\beta}_k^T\mathbf{x})$, $\boldsymbol{\nu}_{\beta_k,\mathbf{x}}(\mathbf{x}) = \boldsymbol{\mu}_{\beta_k,\mathbf{x}}(\mathbf{x}) - \mathbf{x}$, $\mathbf{w}_{\beta_k,\mathbf{x}}(\mathbf{x}) = \mathbb{E}(\mathbf{X}_{ij}\mathbf{X}_{ij}^T|\boldsymbol{\beta}_k^T\mathbf{X}_{ij} = \boldsymbol{\beta}_k^T\mathbf{x})$, and $\mathbf{W}(\mathbf{x}) = \mathbf{w}_{\beta_k,\mathbf{x}}(\mathbf{x}) - \boldsymbol{\mu}_{\beta_k,\mathbf{x}}(\mathbf{x})\boldsymbol{\mu}_{\beta_k,\mathbf{x}}(\mathbf{x})^T$, for $i = 1,\ldots,M$ and $j = 1,\ldots,m_i$. Then, we have the following asymptotic property.

Theorem 1. Under the regularity conditions given in Appendix B.1, we have

$$\sqrt{M}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k) \xrightarrow{d} N(0, \mathbf{W}_{\psi_k}^{+}\boldsymbol{\Sigma}_k\mathbf{W}_{\psi_k}^{+}), \text{ for } k = 1,\ldots,q,$$

where $\mathbf{W}_{\psi_k} = \mathbb{E}\{\psi_k'(\boldsymbol{\beta}_k^T\mathbf{X})^2\mathbf{W}(\mathbf{X})\}/2$, the superscript "+" denotes the Moore-Penrose generalized inverse, and

$$\boldsymbol{\Sigma}_k = \lim_{M\to\infty}\frac{1}{M}\sum_{i=1}^{M}\mathrm{Var}\left\{\frac{1}{\overline{m}}\sum_{j=1}^{m_i}\psi_k'(\boldsymbol{\beta}_k^T\mathbf{X}_{ij})\boldsymbol{\nu}_{\beta_k,\mathbf{x}}(\mathbf{X}_{ij})\left(\varepsilon_{ijk} + z_{ijk} - \widehat{z}_{ijk}\right)\right\},$$

$\overline{m} = M^{-1}\sum_{i=1}^{M} m_i$, $z_{ijk} = \mathbf{g}(t_{ij})^T\mathbf{b}_{ik}$, $\widehat{z}_{ijk} = \mathbf{g}(t_{ij})^T\,\mathbb{E}(\mathbf{b}_{ik}|\mathbf{Y},\widehat{\boldsymbol{\theta}})$, and $\widehat{\boldsymbol{\theta}}$ denotes the collection of the estimates of $\boldsymbol{\Sigma}_b$, $\{\sigma_k^2\}$ and $\{\boldsymbol{\Psi}_{\bullet\bullet k}\}$.

If we check the regularity conditions given in Web Appendix B.1, it can be found that the asymptotic normality of $\{\widehat{\boldsymbol{\beta}}_k\}$ given in Theorem 1 actually does not require the pure measurement errors to be normally distributed. The proof of Theorem 1 is provided in Web Appendix B.2 in the supplementary materials. While $\widehat{z}_{ijk}$ is denoted like an estimate, it is actually a function of the response vector $\mathbf{Y}$ and hence a random variable.

4 | SIMULATION STUDY

We conduct two sets of simulations to assess the performance of our proposed method in estimating the multivariate single-index model (1) with random effects. The first set of simulations aims to evaluate its finite-sample performance when the number of subjects and the number of observation times increase. The second set of simulations compares our method with the rMAVE method proposed by Xia[8] and the p-spline method proposed by Wu and Tu[13] under different scenarios. In all simulation studies, m_i's are the same to be m, and four sample sizes with $M = 50$ or 100 and $m = 5$ or 10 are considered in the main article.

4.1 | Finite-sample performance of the proposed method

Assume that there are two correlated outcome variables and their observations are generated from the following model:

$$\begin{cases} Y_{ij1} = \exp(\beta_{11}X_{ij1} + \beta_{12}X_{ij2} + \beta_{13}X_{ij3}) + b_{i11} + b_{i12}t_{ij} + \epsilon_{ij1} \\ Y_{ij2} = (\beta_{21}X_{ij1} + \beta_{22}X_{ij2} + \beta_{23}X_{ij3})^2 + b_{i21} + b_{i22}t_{ij} + \epsilon_{ij2}, \end{cases}$$

where the single-indices are linked to the conditional mean responses via the link functions $\psi_1(u) = \exp(u)$ and $\psi_2(u) = u^2$. The true values of the index coefficients are $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \beta_{13})^T = (1, 2, 1.5)^T/\sqrt{7.25}$ and $\boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22}, \beta_{23})^T = (2, -1, 3)^T/\sqrt{14}$. For $i = 1, \ldots, M$ and $j = 1, \ldots, m$, we assume $t_{ij} \in [0, 1]$, and the j-th observation time for the i-th subject is generated from the uniform distribution $\mathrm{U}((j-1)/m, j/m)$. The predictors at each observation time are generated as follows: X_{ij1} is generated from $\mathrm{U}(0, 1)$, X_{ij2} is generated from $\mathrm{U}(-1, 1)$, and X_{ij3} is generated by a random number from $\mathrm{U}(1, 2)$ multiplying by the corresponding observation times t_{ij}. Thus, the first two predictors are time-independent and the third predictor is time-dependent. For each subject, the random-effects $(b_{i11}, b_{i12}, b_{i21}, b_{i22})^T$ are generated from a multivariate normal distribution $\mathrm{N}_4(\mathbf{0}, \boldsymbol{\Sigma}_b)$ with

$$\begin{aligned} \boldsymbol{\Sigma}_b &= \begin{pmatrix} 0.4 & 0 & 0 & 0 \\ 0 & 0.4 & 0 & 0 \\ 0 & 0 & 0.6 & 0 \\ 0 & 0 & 0 & 0.6 \end{pmatrix} \begin{pmatrix} 1 & 0.5 & 0.25 & 0.25 \\ 0.5 & 1 & 0.25 & 0.25 \\ 0.25 & 0.25 & 1 & 0.5 \\ 0.25 & 0.25 & 0.5 & 1 \end{pmatrix} \begin{pmatrix} 0.4 & 0 & 0 & 0 \\ 0 & 0.4 & 0 & 0 \\ 0 & 0 & 0.6 & 0 \\ 0 & 0 & 0 & 0.6 \end{pmatrix} \\ &= \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_{44} \end{pmatrix} = \begin{pmatrix} 0.16 & 0.08 & 0.06 & 0.06 \\ 0.08 & 0.16 & 0.06 & 0.06 \\ 0.06 & 0.06 & 0.36 & 0.18 \\ 0.06 & 0.06 & 0.18 & 0.36 \end{pmatrix}. \end{aligned} \tag{14}$$

The pure measurement errors ϵ_{ij1} and ϵ_{ij2} are generated independently from $N(0, 0.01)$ and $N(0, 0.04)$, respectively.

To assess the finite-sample performance of our proposed method, we first compare the parameter estimates with their true values under different sample sizes by reporting some summary statistics based on 250 repeated simulation runs. Tables 1 and 2 present the mean values of the parameter estimates, their mean biases, their variances (Var), and the mean squared errors (MSE) under $m = 5$ and 10, respectively. From the tables, it can be seen that i) the index coefficient estimates are overall estimated more accurately than the variance components with respect to bias and variance, and ii) the MSE values of all parameter estimates decrease as either the number of subjects or the number of repeated measurements of each subject increases.

[Table 1 about here]

[Table 2 about here]

We also studied the estimation of the link functions under different sample sizes. Figures 1 and 2 compare the true link function with their estimates in different sample size cases. From the figures, it can be seen that the averaged pointwise estimates of both link functions are almost identical to the true functions, illustrating the good performance of our proposed model estimation method. The mean integrated squared error (MISE) values are also reported in different cases. As expected, they decrease as the number of subjects and/or the number of repeated measurements increase.

[Figure 1 about here]

[Figure 2 about here]

In Appendix C of the supplementary materials, we present another set of simulation results under a more complex scenario with $q = 5$ response variables, $p = 10$ predictors, and a larger sample size. These results also show that our proposed method can estimate the index coefficients and link functions accurately.

4.2 | Method comparison

In this part, we compare the numerical performance of our proposed method with the competing rMAVE method and the multivariate p-spline method discussed in Section 1. The comparison is conducted in four scenarios corresponding to four different assumptions on the correlation structure of the response variables. In Section 1, we have discussed the main features of the rMAVE and multivariate p-spline methods. The rMAVE method was proposed to handle cases with a univariate response variable and independent data, and it cannot accommodate the within-subject correlation and the correlation among different response variables. The p-spline method describes the within-subject correlation by including random intercepts in its model. However, while the link functions can be different for different response variables, this method assumes the index coefficients

are the same for different response variables. When implementing the p-spline method, we follow the suggestions by Wu and Tu[13] to choose the cubic splines with 20 knots.

Similar to the setups of the simulation study in Section 4.1, assume that there are two continuous outcome variables whose observations are generated from the following model:

$$\begin{cases} Y_{ij1} = \exp(\beta_{11}X_{ij1} + \beta_{12}X_{ij2} + \beta_{13}X_{ij3}) + e_{ij1} \\ Y_{ij2} = (\beta_{21}X_{ij1} + \beta_{22}X_{ij2} + \beta_{23}X_{ij3})^2 + e_{ij2}, \end{cases}$$

where e_{ij1} and e_{ij2} are the error terms. Then, the following four scenarios are considered.

- Scenario 1: Let $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \beta_{13})^T = \boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22}, \beta_{23})^T = (2, -1, 3)^T/\sqrt{14}$. The error terms only includethe pure measurement error, i.e., $e_{ij1} = \varepsilon_{ij1}$ and $e_{ij2} = \varepsilon_{ij2}$. The pure measurement errors ϵ_{ij1} and ϵ_{ij2} are generated independently from $N(0, 0.01)$ and $N(0, 0.04)$, respectively.

- Scenario 2: Let $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \beta_{13})^T = \boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22}, \beta_{23})^T = (2, -1, 3)^T/\sqrt{14}$. The error terms consist of a random intercept and a pure measurement error, i.e., $e_{ij1} = b_{i1} + \varepsilon_{ij1}$ and $e_{ij2} = b_{i2} + \varepsilon_{ij2}$. The random-effects terms $(b_{i1}, b_{i2})^T$ are generated from $\mathrm{N}_2(\mathbf{0}, \boldsymbol{\Sigma}_b)$ with
$$\boldsymbol{\Sigma}_b = \begin{pmatrix} 0.4 & 0 \\ 0 & 0.6 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0.4 & 0 \\ 0 & 0.6 \end{pmatrix} = \begin{pmatrix} 0.16 & 0 \\ 0 & 0.36 \end{pmatrix}.$$
The pure measurement errors ϵ_{ij1} and ϵ_{ij2} are generated in the same way as that in Scenario 1.

- Scenario 3: Let $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \beta_{13})^T = \boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22}, \beta_{23})^T = (2, -1, 3)^T/\sqrt{14}$. The error terms consist of a random intercept, a random slope, and a pure measurement error, i.e., $e_{ij1} = b_{i11} + b_{i12}t_{ij} + \varepsilon_{ij1}$ and $e_{ij2} = b_{i21} + b_{i22}t_{ij} + \varepsilon_{ij2}$. The random-effects terms $(b_{i11}, b_{i12}, b_{i21}, b_{i22})^T$ are generated from $\mathrm{N}_4(\mathbf{0}, \boldsymbol{\Sigma}_b)$ with $\boldsymbol{\Sigma}_b$ be the 4×4 matrix given in (14). The pure measurement errors ϵ_{ij1} and ϵ_{ij2} are generated in the same way as that in Scenario 1.

- Scenario 4: Same as the one considered in Section 4.1. This scenario can be viewed as a modified version of Scenario 3 above with different index coefficients for the two response variables. Namely, $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \beta_{13})^T = (1, 2, 1.5)^T/\sqrt{7.25}$ and $\boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22}, \beta_{23})^T = (2, -1, 3)^T/\sqrt{14}$.

For $k = 1, 2$, the sum of squares of the estimation errors, $\sum_{p=1}^{3}(\hat{\beta}_{kp} - \beta_{kp})^2$, is used to measure the estimation errors in each case considered. Table 3 presents the results under different scenarios described above, where "MK" denotes our proposed method that is based on the **m**ultivariate **k**ernel smoothing procedure. From the table, it can be seen that the rMAVE method has the best performance in Scenario 1 when the observed data are independent over time and among different response variables. This is reasonable because both the p-spline and MK methods need to estimate the variance components in addition to the index coefficient parameters, which would result in some extra variability in the estimates of the index coefficients. However,

in Scenarios 2-4 with some random-effects terms included in the model, the MK method overperforms the other two methods unanimously, with the margins getting larger from Scenario 2 to Scenario 4. Even in Scenario 2 when the true model only contains random intercepts, our proposed method gives the smallest estimation error. It is intuitively reasonable for MK to outperform rMAVE in this case because the latter assumes independent data which is invalid in this scenario. A possible reason for the p-spline method underperforming ours is as follows. In the original work of Wu and Tu[13], they only considered cases with two covariates. When they implemented the proposed estimation procedure, the Nelder-Mead algorithm was used to optimize the index coefficients given the estimates of spline parameters in each iterative. This could be inefficient even when the dimension of the index coefficient vector increase from 2 to 3. Besides, in Scenario 4 when the true indices have different coefficient parameters, it can be seen that the p-spline method is not capable of handling this scenario with different index coefficient vectors.

[Table 3 about here]

5 | CASE STUDY

Maintaining cognitive functioning is a key component of healthy aging[20]. Previous studies have shown the association between cognitive decline and the increased risk of mortality, disability, and poor quality of life[21,22]. While cognitive decline is a long-term process, Zheng et al.[23] found through their prospective studies that a greater number of cognitive assessments were preferable when studying the impact of risk factors on the subsequent trajectory of cognitive function. Over the last decade, there was sizable literature on studying the predictors of the progression of cognitive decline. By applying different analytic methods, including the linear mixed-effects model and the linear growth curve model, some existing studies have figured out a series of key predictors, including age variation, positive well-being, physical activity level, and more[24,25,26,27,23].

With the goal of a better understanding of the association between cognitive function and its key predictors under the longitudinal setting and in a more flexible manner (e.g., taking into account the potential heterogeneity in subject-specific trends over time), we apply the proposed methodology to the analysis of a dataset from the English Longitudinal Study of Aging (ELSA). The ELSA project is an ongoing panel study of adults aged 50 and over. The study commenced in 2002, and the participants have been followed up roughly every 2 years. Most of the raw data were collected through face-to-face interviews and self-completed questionnaires. See Cadar et al.[28] and Steptoe et al.[29] for more detailed information about the ELSA project. In this paper, we choose two major domains of cognitive function, i.e., memory and executive function, as the main endpoints to investigate. To quantify the two endpoints of interest, we follow the data preparation procedure suggested in Zaninotto et al.[27] to define the memory score to be the sum of scores on the immediate and delayed recall tests (ranging from 0 to 20), and the executive function score to be the score on the animal naming task (ranging from 0 to 50). Since the cognitive function measures were fully assessed only at each of the first 5 waves of data collection, we consider the observation times ranged from 0 to 10 years. Then,

the age, assessments of positive well-being, and frequency of engaging in physical activities are chosen to be the predictors for constructing the single indices. The well-being that measures the 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[30]. Higher scores on CAPS-19 index represent higher levels of positive well-being. In ELSA, the frequency of doing physical activities was originally asked in three questions about frequencies in engaging in mildly energetic, moderately energetic, and vigorous activities, respectively. For simplicity, we proceed with the responses to those questions as a vector of ordinal variables with the same levels of 0, 1, 2, and 3, and then compute a physical activity index as their summation. Higher scores on the activity index represent more frequent participation in physical activities.

In our analysis, we randomly select $M = 200$ subjects with complete data on all variables of interest. Then, the number of repeated measurements on each subject is $m = 5$. To get more intuitive interpretations of the index coefficients, we standardize each predictor to have zero mean and unity variance. For $i = 1, \dots, M$ and $j = 1, \dots, m$, Y_{ij1} and Y_{ij2} denote the executive function score and memory score, respectively. For the i-th subject, let X_{ij1}, X_{ij2} and X_{ij3} be the well-being index, physical activity index, and age in years of the subjects measured at the j-th observation time respectively. We assume that the conditional mean of the executive function scores and the memory scores are linked to their corresponding single indices $(\beta_{11}X_{ij1}+\beta_{12}X_{ij2}+\beta_{13}X_{ij3})$ and $(\beta_{21}X_{ij1} + \beta_{22}X_{ij2} + \beta_{23}X_{ij3})$ through the link functions $\psi_1(\cdot)$ and $\psi_2(\cdot)$, respectively. Then, the index parameters $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \beta_{13})^T$ and $\boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22}, \beta_{23})^T$, and the unknown link functions $\psi_1(\cdot)$ and $\psi_2(\cdot)$ need to be estimated using the following model:

$$\begin{cases} Y_{ij1} = \psi_1(\beta_{11}X_{ij1} + \beta_{12}X_{ij2} + \beta_{13}X_{ij3}) + b_{i11} + b_{i12}t_{ij} + \epsilon_{ij1} \\ Y_{ij2} = \psi_2(\beta_{21}X_{ij1} + \beta_{22}X_{ij2} + \beta_{23}X_{ij3}) + b_{i21} + b_{i22}t_{ij} + \epsilon_{ij2}. \end{cases}$$

Table 4 presents the estimated index coefficient parameters for the two response variables, together with their standard errors (SE) estimated by a bootstrap procedure with 250 bootstrap samples. Figure 3 shows the estimated link functions. From the figure, it can be observed that when the single index value does not take an extreme value (e.g., between -2 to 4), the derived single index 1 has a roughly negative association with the corresponding executive function score. When the index is less than -2 or greater than 4, we find the underlying association becomes positive, and the variability of the curve becomes larger. The relationship between the derived single index 2 and the memory score has a clearer pattern depending on the value of the single index. When the index ranges between -2 to 3, the memory score is negatively associated with the underlying index, and it is positively associated with the index when the index value is greater than 3. The interpretation of the index coefficient parameters can be more straightforward when the estimated link function is monotonic. For instance, for the executive function scores, for moderate values of the single index, the positive estimate of the coefficient of the well- being index indicates that positive well-being or better quality of life is positively associated with the decline in executive function. Similarly, frequent engagement in

physical activities is negatively associated with a decline in executive function. The considerable standard error corresponding to the coefficients of age makes it hard to conclude anything about its role in the decline of executive function without any further explorations. Since we have standardized the predictors used to construct the single indices, we can have further conclusions about the relative importance of different factors in affecting the response variables. When the estimated link function has a more complicated shape than a monotonic curve, the interpretation of the index coefficients should be more careful. For instance, if the estimated link function shows a clear pattern of piecewise monotonicity, then it might be reasonable to make separate conclusions in different intervals with monotonic patterns of the estimated link function.

[Table 4 about here]

[Figure 3 about here]

By using the estimates of the variance components of the two response variables, we can compute the correlation between them at each measurement time as follows: for each i and j,

$$\widehat{\text{Corr}}(Y_{ij1}, Y_{ij2}) = \frac{\widehat{\sigma}_{13} + t_{ij}(\widehat{\sigma}_{14} + \widehat{\sigma}_{23}) + t_{ij}^2\widehat{\sigma}_{24}}{\sqrt{(\widehat{\sigma}_{11} + 2t_{ij}\widehat{\sigma}_{12} + t_{ij}^2\widehat{\sigma}_{22} + \widehat{\sigma}_1^2)(\widehat{\sigma}_{33} + 2t_{ij}\widehat{\sigma}_{34} + t_{ij}^2\widehat{\sigma}_{44} + \widehat{\sigma}_2^2)}},$$

where $\widehat{\sigma}_{l_1 l_2}$, for $l_1, l_2 = 1, 2, 3, 4$, denote the estimates of the covariance between two random effects as in Equation (14), and $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$ are the variance estimates of the pure measurement error for the two response variables. Figure 4 shows the estimate of the correlation between the two response variables over time. From the figure, the two response variables are positively correlated, although the correlation changes slightly over time, which is consistent with our intuition that our cognitive function and executive function would roughly increase or decrease together.

[Figure 4 about here]

6 | CONCLUDING REMARKS

In previous sections, we have described a new methodology for analyzing multivariate longitudinal data with multiple response variables using a multivariate single-index model with random-effects. The proposed model allows for different link functions for different response variables and can accommodate within-subject correlation and mutual correlation among different response variables. Both theoretical justifications and numerical studies confirm that it can work well in practice. However, there are still some issues related to the proposed method. First, while the consistency of the index coefficients has been established in the paper, the statistical properties of the estimated variance components still need to be explored. More future research is needed to establish the consistency of the estimated variance components and explore whether there are appropriate methods

(e.g., the Restricted Maximum Likelihood approach) to reduce the bias in the estimated variance components. With a better understanding of the impact of the local linear kernel smoothing technique used in estimating the link functions in the proposed method on the estimation of the variance components, it might be possible to develop some information criteria like AIC and BIC, and/or statistical tests like the likelihood ratio test and tests based on the restricted likelihood estimation to assess the goodness-of-fit of the estimated model. Second, in the current model formulation, the single-indices used in Model (1) are defined to be linear combinations of the predictors. Although the nonparametric link functions used in the model allow for a quite flexible relationship between the response variables and their single indices, it might be more reasonable to define single indices to be certain parametric functions of the predictors (e.g., linear functions of parametrically transformed predictors) in some applications. In addition, the current method can only handle cases when all response variables are continuous. In many applications, however, there could be binary or categorical response variables (e.g., whether a patient is recovered from a disease or not). Generalization of the current method to such cases may not be straightforward. All these issues require much future research to address.

SUPPORTING INFORMATION

Some supplementary materials on the proposed estimation procedure in Section 2, the asymptotic normality theorem in Section 3, and an additional set of simulation study mentioned in Section 4 can be found in the online supplementary file available at the Wiley Library Online.

ACKNOWLEDGEMENTS

We thank the editor, the associate editor and anonymous reviewers for providing many constructive comments and suggestions that have improved the quality of the paper greatly. This research is supported in part by the NSF grant DMS-1914639.

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. Köhler L. Monitoring children's health and well-being by indicators and index: apples and oranges or fruit salad? *Child Care Health Dev* 2016; 42(6): 798–808.
2. Weaver S, Dai D, Stauber CE, Luo R. *The Urban Health Index: a handbook for its calculation and use*. World Health Organization. 2014.
3. Wu J, Tu W. Development of a pediatric body mass index using longitudinal single-index models. *Stat Methods Med Res* 2013; 25(2): 872–884.
4. Hardle W, Stoker TM. Investigating smooth multiple regression by the method of average derivatives. *J Am Stat Assoc* 1989; 84(408):986–995.
5. Hardle W, Hall P, Ichimura H. Optimal smoothing in single-index models. *Ann Stat* 1993; 21(1): 157–178.
6. Ichimura H. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J Econom* 1993; 58(1-2): 71–120.
7. 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: 363–410.
8. Xia Y. Asymptotic distributions for two estimators of the single-index model. *Econ Theory* 2006; 22(6): 1112–1137.
9. Yu Y, Ruppert D. Penalized spline estimation for partially linear single-index models. *J Am Stat Assoc* 2002; 97(460): 1042–1054.
10. Fitzmaurice GM, Laird NM, Ware JH. *Applied longitudinal analysis*. 998. John Wiley & Sons. 2012.
11. Pang Z, Xue L. Estimation for the single-index models with random effects. *Comput Stat Data Anal* 2012; 56(6): 1837–1853.
12. Ruppert D, Wand MP, Carroll RJ. *Semiparametric regression*. No. 12. Cambridge university press. 2003.
13. Wu J, Tu W. A multivariate single-index model for longitudinal data. *Stat Modelling* 2016; 16(5): 392–408.
14. 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.
15. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982: 963–974.
16. Epanechnikov VA. Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications* 1969; 14(1): 153–158.

17. Mack YP, Silverman BW. Weak and strong uniform consistency of kernel regression estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 1982; 61(3): 405–415.

18. De Brabanter K, De Brabanter J, Suykens JA, De Moor B. Kernel regression in the presence of correlated errors. *J Mach Learn Res* 2011; 12(6): 1955–1976.

19. Jiang CR, Wang JL. Functional single index models for longitudinal data. *Ann Stat* 2011: 362–388.

20. Jeste DV, Depp CA, Vahia IV. Successful cognitive and emotional aging. *World psychiatry* 2010; 9(2): 78.

21. Batty GD, Deary IJ, Zaninotto P. Association of cognitive function with cause-specific mortality in middle and older age: follow-up of participants in the English Longitudinal Study of Ageing. *Am J Epidemiol* 2016; 183(3): 183–190.

22. Plassman BL, Williams Jr JW, Burke JR, Holsinger T, Benjamin S. Systematic review: factors associated with risk for and possible prevention of cognitive decline in later life. *Ann Intern Med* 2010; 153(3):182–193.

23. Zheng F, Yan L, Yang Z, Zhong B, Xie W. HbA1c, diabetes and cognitive decline: the English Longitudinal Study of Ageing. *Diabetologia* 2018; 61(4): 839–848.

24. Allerhand M, Gale CR, Deary IJ. The dynamic relationship between cognitive function and positive well-being in older people: a prospective study using the English Longitudinal Study of Aging. *Psychol Aging* 2014; 29(2): 306.

25. Hamer M, Terrera GM, Demakakos P. Physical activity and trajectories in cognitive function: English longitudinal study of ageing. *J Epidemiol Community Health* 2018; 72(6): 477–483.

26. Shankar A, Rafnsson SB, Steptoe A. Longitudinal associations between social connections and subjective wellbeing in the English Longitudinal Study of Ageing. *Psychology & health* 2015; 30(6): 686–698.

27. Zaninotto P, Batty GD, Allerhand M, Deary IJ. Cognitive function trajectories and their determinants in older people: 8 years of follow-up in the English Longitudinal Study of Ageing. *J Epidemiol Community Health* 2018; 72(8): 685–694.

28. 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.

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

30. 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.

Table 1 Summary of parameter estimates under $M = 50$ or 100 and $m = 5$. The true parameter values are $\beta_{11} = 0.3714$, $\beta_{12} = 0.7428$, $\beta_{13} = 0.5571$, $\beta_{21} = 0.5345$, $\beta_{22} = -0.2673$, $\beta_{23} = 0.8018$, $\sigma_1^2 = 0.01$, $\sigma_2^2 = 0.04$, $\sigma_{11} = 0.16$, $\sigma_{12} = 0.08$, $\sigma_{13} = 0.06$, $\sigma_{14} = 0.06$, $\sigma_{22} = 0.16$, $\sigma_{23} = 0.06$, $\sigma_{24} = 0.06$, $\sigma_{33} = 0.36$, $\sigma_{34} = 0.18$, and $\sigma_{44} = 0.36$. The Var and MSE values are both in the unit of 10^{-3}.

	$M = 50$				$M = 100$			
Parameter	Mean	Bias	Var	MSE	Mean	Bias	Var	MSE
β_{11}	0.3717	0.0003	0.1561	0.1562	0.3711	-0.0003	0.0687	0.0688
β_{12}	0.7356	-0.0072	0.0798	0.1310	0.7417	-0.0011	0.0260	0.0272
β_{13}	0.5660	0.0089	0.1189	0.1985	0.5586	0.0015	0.0495	0.0519
β_{21}	0.5261	-0.0084	0.5935	0.6647	0.5352	0.0006	0.2190	0.2194
β_{22}	-0.2643	0.0030	0.2595	0.2682	-0.2679	-0.0006	0.1196	0.1199
β_{23}	0.8076	0.0059	0.2276	0.2620	0.8009	-0.0009	0.0968	0.0976
σ_1^2	0.0140	0.0040	0.0127	0.0290	0.0137	0.0037	0.0046	0.0186
σ_2^2	0.0482	0.0082	0.0730	0.1411	0.0475	0.0075	0.0311	0.0866
σ_{11}	0.1519	-0.0081	0.4070	0.4721	0.1546	-0.0054	0.1279	0.1574
σ_{12}	0.0719	-0.0081	0.3035	0.3696	0.0739	-0.0061	0.1391	0.1759
σ_{13}	0.0574	-0.0026	0.4489	0.4558	0.0573	-0.0027	0.2159	0.2234
σ_{14}	0.0569	-0.0031	0.8106	0.8201	0.0575	-0.0025	0.3867	0.3932
σ_{22}	0.1608	0.0008	0.9369	0.9377	0.1587	-0.0013	0.4253	0.4270
σ_{23}	0.0548	-0.0052	0.7363	0.7637	0.0558	-0.0042	0.3448	0.3628
σ_{24}	0.0565	-0.0035	1.2275	1.2394	0.0554	-0.0046	0.5234	0.5444
σ_{33}	0.3486	-0.0114	2.3939	2.5235	0.3502	-0.0098	1.0481	1.1445
σ_{34}	0.1670	-0.0130	1.9169	2.0851	0.1645	-0.0155	0.9415	1.1826
σ_{44}	0.3512	-0.0088	9.3728	9.4501	0.3541	-0.0059	3.2946	3.3299

Table 2 Summary of parameter estimates under $M = 50$ or 100 and $m = 10$. The true values of the parameters are $\beta_{11} = 0.3714$, $\beta_{12} = 0.7428$, $\beta_{13} = 0.5571$, $\beta_{21} = 0.5345$, $\beta_{22} = -0.2673$, $\beta_{23} = 0.8018$, $\sigma_1^2 = 0.01$, $\sigma_2^2 = 0.04$, $\sigma_{11} = 0.16$, $\sigma_{12} = 0.08$, $\sigma_{13} = 0.06$, $\sigma_{14} = 0.06$, $\sigma_{22} = 0.16$, $\sigma_{23} = 0.06$, $\sigma_{24} = 0.06$, $\sigma_{33} = 0.36$, $\sigma_{34} = 0.18$, and $\sigma_{44} = 0.36$. The Var and MSE values are both in the unit of 10^{-3}.

	$M = 50$				$M = 100$			
Parameter	Mean	Bias	Var	MSE	Mean	Bias	Var	MSE
β_{11}	0.3713	-0.0001	0.0587	0.0587	0.3711	-0.0003	0.0297	0.0298
β_{12}	0.7409	-0.0019	0.0401	0.0435	0.7417	-0.0011	0.0234	0.0246
β_{13}	0.5595	0.0024	0.0932	0.0989	0.5586	0.0016	0.0492	0.0516
β_{21}	0.5347	0.0002	0.2170	0.2170	0.5339	-0.0006	0.1340	0.1344
β_{22}	-0.2681	-0.0009	0.1002	0.1010	-0.2672	0.0001	0.0474	0.0474
β_{23}	0.8011	-0.0007	0.1039	0.1043	0.8021	0.0003	0.0582	0.0582
σ_1^2	0.0141	0.0041	0.0053	0.0223	0.0128	0.0028	0.0025	0.0105
σ_2^2	0.0471	0.0071	0.0265	0.0766	0.0448	0.0048	0.0118	0.0345
σ_{11}	0.1520	-0.0080	0.3339	0.3979	0.1565	-0.0035	0.1122	0.1244
σ_{12}	0.0714	-0.0086	0.2273	0.3006	0.0759	-0.0041	0.0904	0.1075
σ_{13}	0.0578	-0.0022	0.3375	0.3421	0.0592	-0.0008	0.1683	0.1689
σ_{14}	0.0555	-0.0045	0.5467	0.5665	0.0578	-0.0022	0.2255	0.2304
σ_{22}	0.1565	-0.0035	0.5669	0.5791	0.1580	-0.0020	0.2691	0.2730
σ_{23}	0.0531	-0.0069	0.4174	0.4652	0.0564	-0.0036	0.1960	0.2091
σ_{24}	0.0557	-0.0043	0.6434	0.6617	0.0563	-0.0037	0.3023	0.3159
σ_{33}	0.3518	-0.0082	1.6286	1.6963	0.3527	-0.0073	0.6559	0.7094
σ_{34}	0.1641	-0.0159	1.2756	1.5289	0.1719	-0.0081	0.5500	0.6157
σ_{44}	0.3448	-0.0152	6.7558	6.9857	0.3489	-0.0111	1.9506	2.0741

Table 3 Estimation errors $\sum_{p=1}^{3}(\hat{\beta}_{1p} - \beta_{1p})^2$ and $\sum_{p=1}^{3}(\hat{\beta}_{2p} - \beta_{2p})^2$ computed based on 250 repeated simulations under different scenarios. All values are in the unit of 10^{-3}, and the smallest value in each case among the three methods is in bold.

			rMAVE		P-spline		MK	
Scenario	M	m	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$
1	50	5	**0.0714**	**0.6153**	1.2538	1.2538	0.0777	0.6264
		10	**0.0334**	**0.2641**	0.9071	0.9071	0.0353	0.2691
	100	5	**0.0344**	**0.2622**	0.9659	0.9659	0.0351	0.2676
		10	**0.0188**	**0.1510**	0.9429	0.9429	0.0192	0.1561
2	50	5	1.0687	4.9204	1.3866	1.3866	**0.1051**	**0.7887**
		10	0.5301	3.0934	1.2542	1.2542	**0.0418**	**0.3049**
	100	5	0.5422	2.7060	1.2460	1.2460	**0.0498**	**0.3446**
		10	0.2769	1.3737	1.0708	1.0708	**0.0217**	**0.1730**
3	50	5	2.0256	11.5785	2.0781	2.0781	**0.1952**	**1.2008**
		10	1.0855	5.5032	1.5277	1.5277	**0.0819**	**0.4236**
	100	5	1.0902	6.5289	1.5758	1.5758	**0.0884**	**0.4318**
		10	0.5847	2.8336	1.2806	1.2806	**0.0423**	**0.2401**
4	50	5	3.3489	11.5785	184.7875	97.2147	**0.4857**	**1.1949**
		10	1.9145	5.5032	1.2944	1082.6353	**0.2012**	**0.4224**
	100	5	1.5131	6.5289	1.3015	1082.1406	**0.1479**	**0.4369**
		10	0.9531	2.8336	0.6901	1087.0904	**0.1059**	**0.2401**

Figure 1 In each plot, the black solid line denotes the true link function ψ_1, the gray dashed line denotes its averaged pointwise estimate, and the black dashed lines denote the pointwise 95% confidence interval. The results are based on 250 repeated simulations.

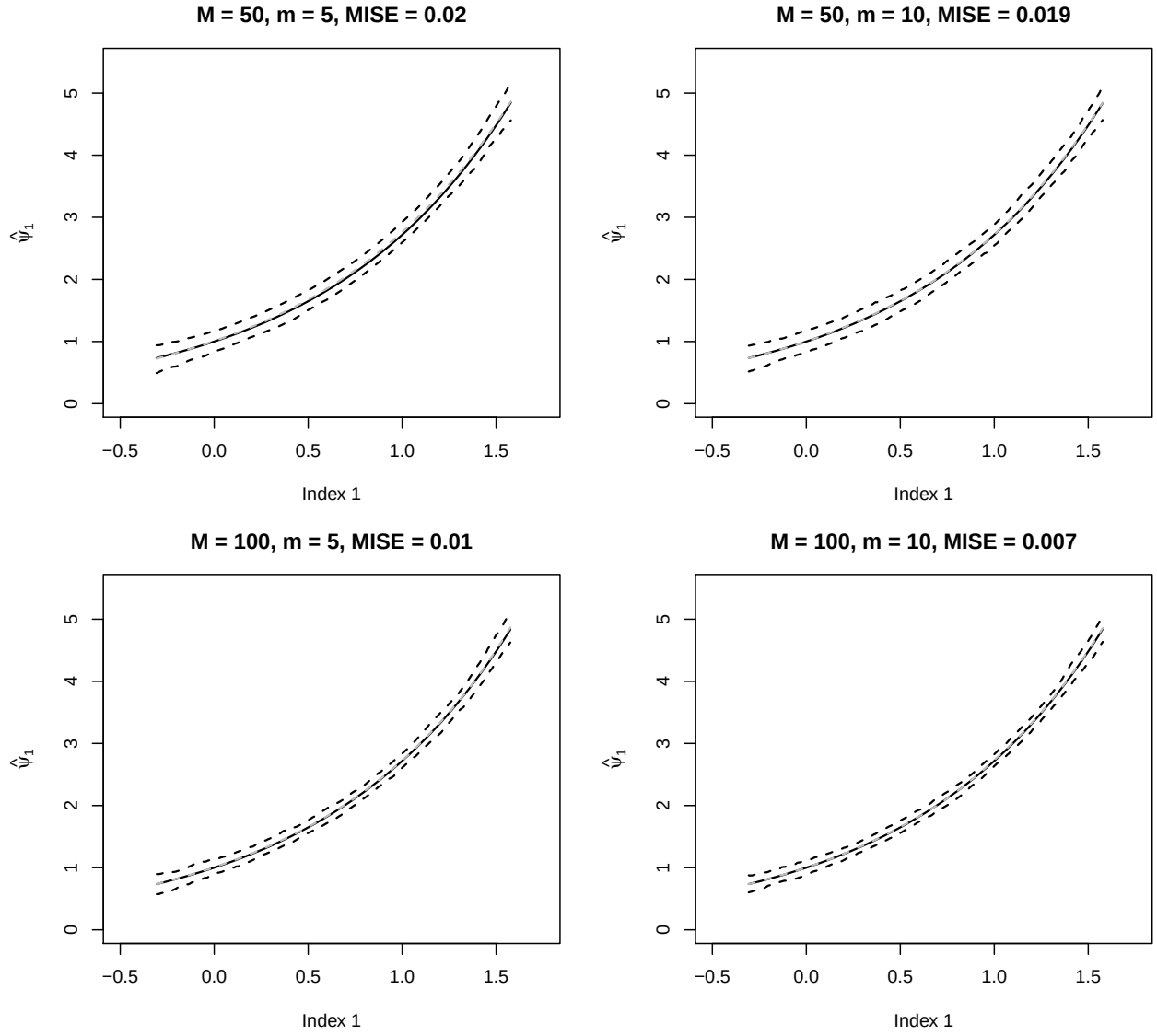

Table 4 Estimated index coefficient parameters with their standard errors (in parentheses) computed by a bootstrap procedure with 250 bootstrap samples.

Outcomes	Well-being	Physical activity	Age
Executive function score	0.7614 (0.2541)	-0.5206 (0.2052)	0.3863 (0.3705)
Memory score	0.5129 (0.2148)	-0.3410 (0.3290)	0.7878 (0.7317)

Figure 2 In each plot, the black solid line denotes the true link function ψ_2, the gray dashed line denotes its averaged pointwise estimate, and the black dashed lines denote the pointwise 95% confidence interval. The results are based on 250 repeated simulations.

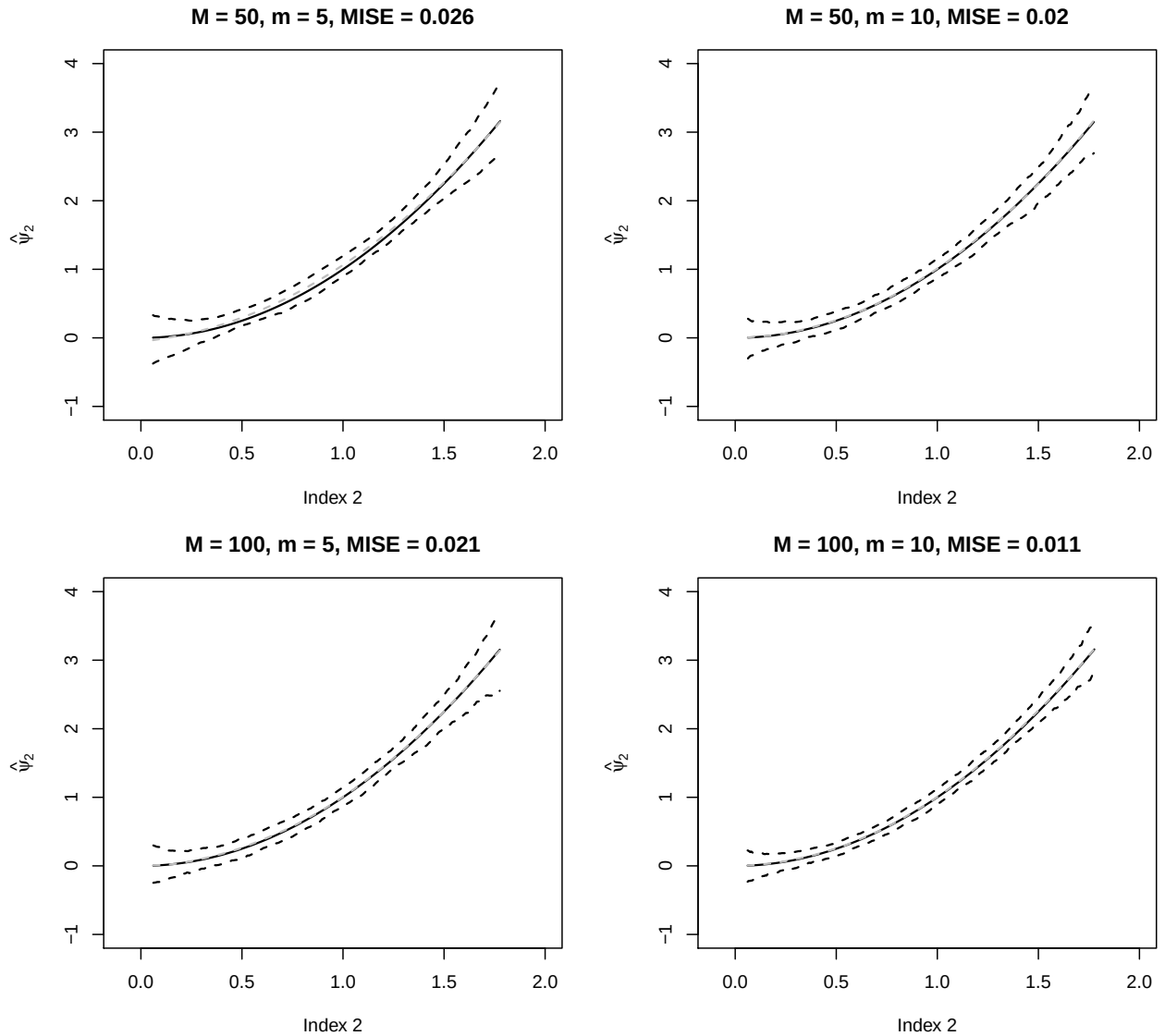

Figure 3 Estimates of the link functions ψ_1 and ψ_2 (solid lines) and their 95% pointwise confidence intervals (dashed lines) computed by a bootstrap procedure with 250 bootstrap samples.

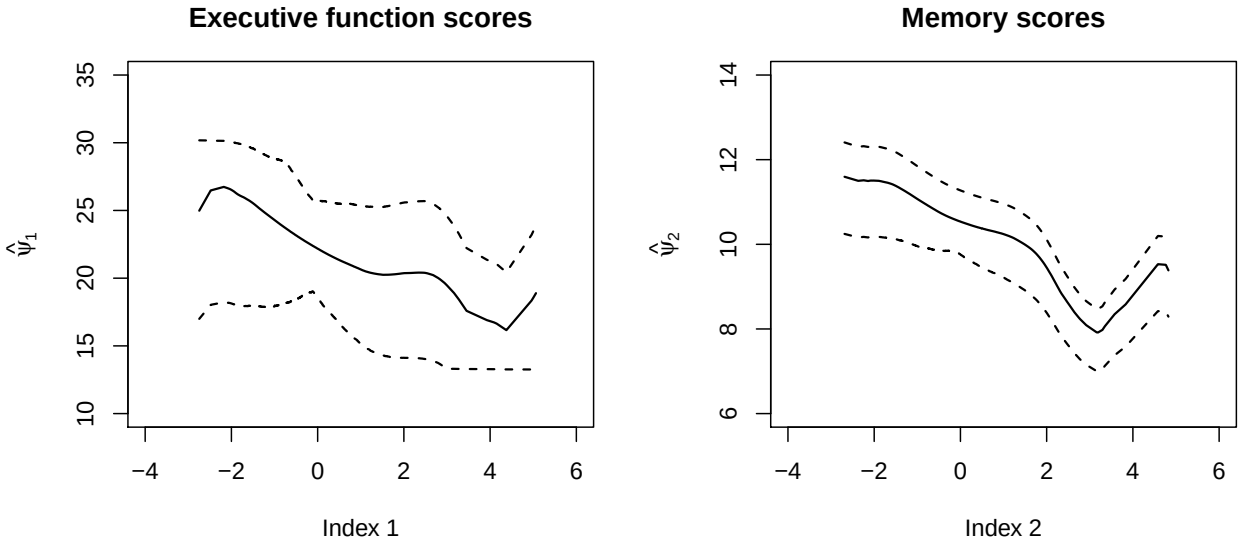

Figure 4 Estimated correlation between executive function score and memory score of a specific subject over measurement time in years.