Received: 18 May 2019 | Revised: 30 June 2019 | Accepted: 8 July 2019

DOI: 10.1002/sta4.246

ORIGINAL ARTICLE

WILEY

Logistic regression error-in-covariate models for longitudinal high-dimensional covariates

Hyung Park[1] | Seonjoo Lee[2]

[1] Division of Biostatistics, Department of Population Health, New York University, New York, NY 10016

[2] Research Foundation for Mental Hygiene, Inc., New York, NY 10032

Correspondence
Seonjoo Lee, Research Foundation for Mental Hygiene, Inc., New York, NY 10032.
Email: sl3670@cumc.columbia.edu

Funding information
National Institutes of Health, Grant/Award Number: K01AG051348

We consider a logistic regression model for a binary response where part of its covariates are subject-specific random intercepts and slopes from a large number of longitudinal covariates. These random effect covariates must be estimated from the observed data, and therefore, the model essentially involves errors in covariates. Because of high dimension and high correlation of the random effects, we employ longitudinal principal component analysis to reduce the total number of random effects to some manageable number of random effects. To deal with errors in covariates, we extend the conditional-score equation approach to this moderate dimensional logistic regression model with random effect covariates. To reliably solve the conditional-score equations in moderate/high dimension, we apply a majorization on the first derivative of the conditional-score functions and a penalized estimation by the smoothly clipped absolute deviation. The method was evaluated through a set of simulation studies and applied to a data set with longitudinal cortical thickness of 68 regions of interest to identify biomarkers that are related to dementia transition.

KEYWORDS

conditional-score equations, errors in covariates, longitudinal functional principal component analysis

1 | INTRODUCTION

In this paper, we estimate a logistic regression model for a binary response where part of its covariates are subject-specific random intercepts and random slopes from a large number of longitudinal covariates. As the random intercepts and slopes are not observable, these random effect covariates must be estimated from the observed longitudinal high-dimensional data. Therefore, the model essentially involves errors in covariates, and we aim to improve parameter estimation accuracy by taking into account the errors. This study was motivated by Alzheimer's Disease Neuroimaging Initiative (ADNI) whose primary aim is to identify biomarkers that are related to dementia transition in elderly participants. Particularly, we analyse longitudinal trajectories of cortical thickness measures to identify biomarkers of dementia transition among the participants with mild cognitive impairment (MCI).

In the neuroimaging literature, statistical models are typically based on measurement error-free assumptions. However, the accuracy of parameter estimates deteriorates due to scan-to-scan variation (e.g., Iscan et al., 2015) present in imaging variables. In particular, if the covariates are subject-specific random effects obtained from a longitudinal high-dimensional (imaging) covariate, the effect of the measurement error can be amplified due to the uncertainty associated with the random effect estimation. To address this, we develop a method accounting for errors in covariates that is applicable to a longitudinal high-dimensional covariate setting.

For generalized linear models (GLMs) with errors in covariates, various statistical procedures have been developed. For earlier development, see Bickel and Ritov (1987) and Carroll, Knickerbocker, and Wang (1995) among many others. Further systematic review can be found in Carroll, Ruppert, Stefanski, and Crainiceanu (2006) and Yi (2016) and many others cited therein. Most related to our approach is the conditional-score

method of Stefanski and Carroll (1987) for an unbiased estimation of the parameter in the presence of errors in covariates. Li, Zhang, and Davidian (2004) adopted their conditional-score method to estimate a GLM with errors in covariates, where its covariate is a subject-specific univariate trajectory modelled by longitudinal random effects. Our approach is a high-dimensional counterpart of the work by Li et al. (2004). However, the extension to a high-dimensional longitudinal covariate is not trivial, as the number of random effects needed for characterizing the high-dimensional longitudinal trajectory increases dramatically and the associated random effects can be highly correlated.

In spite of the rich literature in the subject, there are only few studies on estimation of a logistic regression measurement error model where its covariates are moderate/high dimensional. Generally, estimation of a high-dimensional non-linear measurement error model is challenging. Ma and Li (2010) developed a class of covariate selection procedures for partially linear logistic regression measurement error models. However, their procedures and algorithms may not scale well to a high-dimensional covariate setting. Recently, Datta and Zou (2017) further studied theoretical properties and cross-validation for measurement errors. Besides classical high-dimensional regression, Stefanski, Wu, and White (2014) proposed a variable selection method for nonparametric classification. Cai (2015) considered the simulation extrapolation (SIMEX) method and applied it to a functional covariate under the assumption of uncorrelated measurement errors in a scalar-on-function regression.

For a regression with a longitudinal high-dimensional covariate, the covariance of the covariate measurement error is not easily characterizable due to the high correlation/high dimensionality of the covariate, which makes applications of SIMEX-like algorithms to handle the covariate measurement errors generally difficult. In the presence of high correlation among the covariates, employing a regularization estimation such as the elastic-net (Zou & Hastie, 2005) for the outcome regression model is also prone to severe inaccuracy (as illustrated in our simulation examples), as most regularization-based methods assume weakly correlated or independent covariates.

In this paper, we employ the longitudinal functional principal component analysis (LFPCA; Greven, Crainiceanu, Caffo, & Reich, 2010; Zipunnikov et al., 2014) to account for the correlation structure in the longitudinal covariates. To account for errors in covariates that occur from the estimation of subject-specific longitudinal random effects, we extend the unbiased conditional-score equations of Stefanski and Carroll (1987) to this moderate/high-dimensional logistic regression model with random effect covariates. To reliably solve the conditional-score equations in moderate/high dimension, we apply a majorization on the first derivative of the conditional-score functions and a penalized estimation by the smoothly clipped absolute deviation (SCAD; Fan & Li, 2001).

Although not directly related to this paper, several other works on measurement error models include Li, Shao, and Palta (2005) for a longitudinal outcome, Li, Tang, and Lin (2009) and Huque, Bondell, Carroll, and Ryan (2016) for a spatial regression setting, Midthune, Carroll, Freedman, and Kipnis (2016) for including interaction terms, and Zhang, Wang, Ma, and Carroll (2017) for model selection for prediction.

The paper is organized as follows. In Section 2.1, we introduce an LFPC representation of the covariate model. In Section 2.2, we construct a set of unbiased conditional-score equations in the longitudinal high-dimensional covariate setting. In Section 2.3, we develop an algorithm to obtain a solution to the penalized conditional-score equations. We present simulation studies for assessing the performance of the proposed method in Section 3, and an application of the method to ADNI data is illustrated in Section 4. The paper concludes with discussion in Section 5.

2 | METHOD

We consider a random intercepts and slopes model for characterizing a longitudinal covariate vector $W_{ij} \in \mathbb{R}^p$

$$W_{ij} = X_i^{(0)} + X_i^{(1)} t_{ij} + U_{ij} \quad (i = 1, \ ..., n; \quad j = 1, \ ..., J_i), \tag{1}$$

for the visit times $t_{i1}, \ ..., t_{iJ_i}$. In Equation (1), subject-specific vectors $X_i^{(0)} \in \mathbb{R}^p$ and $X_i^{(1)} \in \mathbb{R}^p$correspond to the random intercepts and random slopes, respectively. These random effects $X_i := \left(X_i^{(0)'}, X_i^{(1)'}\right)' \in \mathbb{R}^{2p}$ (i=1,...,n) are assumed to be distributed with zero mean and a covariance $\Sigma_X = \begin{bmatrix} \Sigma_X^{(0,0)} & \Sigma_X^{(0,1)} \\ \Sigma_X^{(1,0)} & \Sigma_X^{(1,1)} \end{bmatrix}$, where $\Sigma_X^{(a,b)} = \mathbb{E}\left(X_i^{(a)} X_i^{(b)'}\right)$, $a,b \in \{0,1\}$, uncorrelated with the p×1 noise vectors $U_{ij} \sim \mathcal{N}(\mathbf{0}, \Sigma_U)$.

For a binary outcome $Y_i \in \{0,1\}$, we consider a logistic regression model

$$\mathbb{E}(Y_i | Z_i, X_i) = F(b_0 + \beta_0' Z_i + \beta_1' X_i) \quad (i = 1, \ ..., n), \tag{2}$$

in which $F(u) := 1/(1+e^{-u})$ (the inverse-logit function). In Equation (2), $b_0 \in \mathbb{R}$ is an intercept, $Z_i \in \mathbb{R}^q$ corresponds to a vector of "error-free" covariates (such as age and gender), and $X_i \in \mathbb{R}^{2p}$ corresponds to the vector of the subject-specific "error-prone" random effects, which must be estimated from the observed longitudinal covariates $W_{ij} \in \mathbb{R}^p$, in the presence of the noise $U_{ij} \in \mathbb{R}^p$ in Equation (1). Our focus is on the estimation of the coefficient vector $\beta_1 \in \mathbb{R}^{2p}$ associated with X_i, accounting for the errors in covariates that occur from U_{ij}.

In Equation (1), without loss of generality, W_{ij} is assumed to have mean zero, as it can be priorly shifted by a fixed effect, which can be consistently estimated, for example, by variable-wise means (Greven et al., 2010).

2.1 | Longitudinal principal component analysis representation

The high dimensionality/high correlation among W_{ij} makes the covariance $\mathbf{\Sigma}_U$ of the noise U_{ij} not easily characterizable. To handle this, we represent model (1) using the LFPC (Greven et al., 2010; Zipunnikov et al., 2014) mixed effects model framework. Let us represent $X_i \approx \Phi_X \tilde{X}_i$, where $\Phi_X = \left(\Phi_X^{(0)'}, \Phi_X^{(1)'}\right)'$ is a $2p \times N_X$ matrix of the leading N_X eigenvectors of $\mathbf{\Sigma}_X$, and $\tilde{X}_i = (\tilde{x}_{i1}, \ldots, \tilde{x}_{iN_X})' \in \mathbb{R}^{N_X}$ is the associated subject i-specific eigenscores. Let us represent $U_{ij} \approx \Phi_U \tilde{U}_{ij}$, where $\mathbf{\Phi}_U$ is a $p \times N_U$ matrix of the leading N_U eigenvectors of $\mathbf{\Sigma}_U$, and $\tilde{U}_{ij} = (\tilde{u}_{ij1}, \ldots, \tilde{u}_{ijN_U})' \in \mathbb{R}^{N_U}$ is the associated subject(i)- and visit(j)-specific eigenscores. Under the LFPC framework, model (1) can be rewritten as

$$W_{ij} = \Phi_X^{(0)} \tilde{X}_i + t_{ij} \Phi_X^{(1)} \tilde{X}_i + \Phi_U \tilde{U}_{ij} \quad (i = 1, \ldots, n;\ j = 1, \ldots, J_i), \tag{3}$$

where $(\tilde{x}_{ia}, \tilde{x}_{ib}) \sim (0, 0; \lambda_X^{(a)}, \lambda_X^{(b)}; 0)$and $(\tilde{u}_{ija}, \tilde{u}_{ijb}) \sim \mathcal{N}(0, 0; \lambda_U^{(a)}, \lambda_U^{(b)}; 0)$, in which "$\cdot \sim (\mu_1, \mu_2; \sigma_1^2, \sigma_2^2; \rho)$" represents a pair of variables that has a distribution with mean (μ_1,μ_2), variance (σ_1^2, σ_2^2), and correlation ρ; $\mathcal{N}$ denotes the Gaussian distribution. The eigenvectors $(\mathbf{\Phi}_X,\mathbf{\Phi}_U)$ and the eigenscores $(\tilde{X}_i, \tilde{U}_{ij})$ of the LFPC model (3) can be obtained by the least squares estimation of the covariance matrices $\mathbf{\Sigma}_X$ and $\mathbf{\Sigma}_U$, and by the best unbiased linear predictors (Greven et al., 2010; Zipunnikov et al., 2014).

Representation (3) implies $\sum_X = \Phi_X \Lambda_X \Phi_X'$and $\sum_U = \Phi_U \Lambda_U \Phi_U'$, where the diagonal matrices $\mathbf{\Lambda}_X$ and $\mathbf{\Lambda}_U$ are the matrices with the diagonal elements $(\lambda_X^{(1)}, \ldots, \lambda_X^{(N_X)})$ and $(\lambda_U^{(1)}, \ldots, \lambda_U^{(N_U)})$, respectively. Several ways for choosing N_X and N_U were discussed in Greven et al. (2010).

The random effect dimension ($2p$) often far exceeds the number of observations (n). However, the dimension $2p$ can be considerably reduced by exploiting the intrinsic correlation structure in X_i, represented by the covariance $\mathbf{\Sigma}_X$. By representing the coefficients $\beta_1 \in \mathbb{R}^{2p}$ on the basis of the eigenvectors $\mathbf{\Phi}_X$ of $\mathbf{\Sigma}_X$,

$$\beta_1 = \Phi_X \tilde{\beta}_1, \tag{4}$$

where $\tilde{\beta}_1 := \Phi_X' \beta_1 \in \mathbb{R}^{N_X}$. Under representations (3) and (4), the primary outcome model (2) is rewritten as

$$\mathbb{E}\left(Y_i | Z_i, \tilde{X}_i\right) = F\left(b_0 + \beta_0' Z_i + \tilde{\beta}_1' \tilde{X}_i\right) \quad (i = 1, \ldots, n). \tag{5}$$

2.2 | Conditional-score equations

We can rearrange model (3) as

$$W_i = \mathbf{B}_i^{(X)} \tilde{X}_i + \mathbf{B}_i^{(U)} \tilde{U}_i \quad (i = 1, \ldots, n), \tag{6}$$

where $W_i := (W_{i1}', \ldots, W_{iJ_i}')' \in \mathbb{R}^{pJ_i}$ is the stack-up vector containing $W_{ij} \in \mathbb{R}^p$ for j=1,...,J_i; the $pJ_i \times N_X$ matrix $\mathbf{B}_i^{(X)} := \mathbf{1}_{J_i} \otimes \Phi_X^{(0)} + \mathbf{t}_i \otimes \Phi_X^{(1)}$, in which $\otimes$ is the Kronecker product, $\mathbf{1}_{J_i} := (1, 1, \ldots, 1, 1)' \in \mathbb{R}^{J_i}$and $\mathbf{t}_i := (t_{i1}, \ldots, t_{iJ_i})' \in \mathbb{R}^{J_i}$; the $pJ_i \times N_U J_i$ matrix $\mathbf{B}_i^{(U)} := \mathbf{I}_{J_i} \otimes \Phi_U$, in which $\mathbf{I}_{J_i}$ is the $J_i \times J_i$ identity matrix; $\tilde{U}_i := (\tilde{U}_{i1}', \ldots, \tilde{U}_{iJ_i}')' \in \mathbb{R}^{N_U J_i}$ is the stack-up vector containing $\tilde{U}_{ij} \in \mathbb{R}^{N_U}$ for j=1,...,J_i. Note that $\tilde{U}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{J_i} \otimes \Lambda_U)$. Multiplying both sides of model (6) by $\left(\mathbf{B}_i^{(X)'} \mathbf{B}_i^{(X)}\right)^{-1} \mathbf{B}_i^{(X)'}$ gives

$$\left(\mathbf{B}_i^{(X)'} \mathbf{B}_i^{(X)}\right)^{-1} \mathbf{B}_i^{(X)'} W_i = \tilde{X}_i + \left(\mathbf{B}_i^{(X)'} \mathbf{B}_i^{(X)}\right)^{-1} \mathbf{B}_i^{(X)'} \mathbf{B}_i^{(U)} \tilde{U}_i,$$

which is equivalently

$$W_i^* = \tilde{X}_i + \tilde{U}_i^* \quad (i = 1, \ldots, n), \tag{7}$$

where $W_i^* := \left(\mathbf{B}_i^{(X)'} \mathbf{B}_i^{(X)}\right)^{-1} \mathbf{B}_i^{(X)'} W_i \in \mathbb{R}^{N_X}$ and $\tilde{U}_i^* := \left(\mathbf{B}_i^{(X)'} \mathbf{B}_i^{(X)}\right)^{-1} \mathbf{B}_i^{(X)'} \mathbf{B}_i^{(U)} \tilde{U}_i = \mathbf{H}_i \tilde{U}_i \in \mathbb{R}^{N_X}$, in which $\mathbf{H}_i := \left(\mathbf{B}_i^{(X)'} \mathbf{B}_i^{(X)}\right)^{-1} \mathbf{B}_i^{(X)'} \mathbf{B}_i^{(U)}$ is an $N_X \times N_U J_i$ matrix. As $\tilde{U}_i^* = \mathbf{H}_i \tilde{U}_i$, we have $\tilde{U}_i^* \sim \mathcal{N}(\mathbf{0}, \Omega_i)$ with $\Omega_i := \mathbf{H}_i (\mathbf{I}_{J_i} \otimes \Lambda_U) \mathbf{H}_i'$. In Equation (7), the "noise" component $\tilde{U}_i^* \in \mathbb{R}^{N_X}$ associated with the "signal" component $\tilde{X}_i$is additive, mean zero, and normally distributed. Thus, representation (7) is under the setting of the *conditional-score function* method of Stefanski and Carroll (1987). For the joint model of the covariates (7) and the outcomes (5), one useful sufficient statistic for the latent component $\tilde{X}_i$ is given by Stefanski and Carroll (1985, 1987),

$$\Delta_i := W_i^* + Y_i\Omega_i\tilde{\beta}_1 \quad (i = 1, \ ..., \ n). \tag{8}$$

For joint models (7) and (5), the distribution of Y_i given $(Z_i, \tilde{X}_i, \Delta_i)$ has the following conditional expectation:

$$\mathbb{E}\left(Y_i|Z_i, \tilde{X}_i, \Delta_i\right) = \mathbb{E}(Y_i|Z_i, \Delta_i) = F\left(b_0 + \beta_0'Z_i + \tilde{\beta}_1'\Delta_i - \tilde{\beta}_1'\Omega_i\tilde{\beta}_1/2\right),$$

free of the "error-prone" variable $\tilde{X}_i$, which leads to the *unbiased score equations* of Stefanski and Carroll (1985, 1987):

$$n^{-1}\sum_{i=1}^{n}\{Y_i - F(b_0 + \beta_0'Z_i + \tilde{\beta}_1'\boldsymbol{s}_i)\}(1, Z_i', \boldsymbol{s}_i')' = \boldsymbol{0}, \tag{9}$$

where $\boldsymbol{s}_i := \Delta_i - \Omega_i\tilde{\beta}_1/2$ (i=1,...,n), in which Δ_i is defined in Equation (8). Let us write

$$G_i(\beta) := \left(1, Z_i', \boldsymbol{s}_i'\right)' \in \mathbb{R}^{1+d} \quad (i = 1, \ ..., \ n), \tag{10}$$

and then Equation (9) can be rewritten as

$$n^{-1}\sum_{i=1}^{n}\Psi(Y_i, G_i(\beta), \beta) := n^{-1}\sum_{i=1}^{n}\{Y_i - F(\beta'G_i)\}G_i = \boldsymbol{0}, \tag{11}$$

where $\beta := (b_0, \beta_0', \tilde{\beta}_1')' \in \mathbb{R}^{1+d}$; the number $d(=q+N_X)$ denotes the dimensionality of $(\beta_0', \tilde{\beta}_1')' \in \mathbb{R}^d$. In Equation (11), for the notational simplicity, the functional argument ($\boldsymbol{\beta}$) of $G_i(\boldsymbol{\beta})$ was omitted. Equation (11) is in the form of standard score equations in logistic regression, except for the dependence of the "covariate" G_i on the parameter $\boldsymbol{\beta}$ (as $\boldsymbol{s}_i$ of Equation 10 depends on $\tilde{\beta}_1$), and thus the "linear predictor" $\boldsymbol{\beta}'G_i$ in Equation (11) is non-linear with respect to $\boldsymbol{\beta}$.

Remark 1. If the LFPC dimension reduction (3) is not performed, then in representation (6), $\boldsymbol{B}_i^{(X)} = (\boldsymbol{1}_{J_i}\otimes\boldsymbol{I}_p; \ \boldsymbol{t}_i\otimes\boldsymbol{I}_p)$ is a $pJ_i\times 2p$ matrix; $\boldsymbol{B}_i^{(U)} = \boldsymbol{I}_{J_i}\otimes\boldsymbol{I}_p$ is a $pJ_i\times pJ_i$ matrix; $\tilde{X}_i$ (=X_i) is a $2p\times 1$ vector; $\tilde{U}_i$ ($= U_i = (U_{i1}', \ ..., U_{iJ_i}')'$) represents a $pJ_i\times 1$ vector that follows $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{J_i}\otimes\sum_{\boldsymbol{U}})$; the noise covariance $\boldsymbol{\Sigma_U}$ can be estimated via (restricted) maximum likelihood estimation (MLE) on the basis of model (1), given observed covariates. The sufficient "statistic" (8) of the random effects X_i is given by Δ_i=W_i+$Y_i\boldsymbol{\Omega}_i\boldsymbol{\beta}_1$, in which $\Omega_i = \boldsymbol{H}_i(\boldsymbol{I}_{J_i}\otimes\sum_{\boldsymbol{U}})\boldsymbol{H}_i'$, and the parameter $\boldsymbol{\beta}$ in Equation (11) is $\beta = (b_0, \beta_0', \beta_1')'$; that is, $\boldsymbol{\beta}_1$ takes the part of $\tilde{\beta}_1$.

2.3 | Penalized estimation via coordinate descent

Carroll et al. (2006) noted that a Newton-Raphson-type algorithm can be utilized to solve the unbiased score equations of Stefanski and Carroll (1985, 1987). If $\hat{\beta}$ represents a solution to the estimating equation (11), then $\sqrt{n}(\hat{\beta} - \beta)$ is asymptotically $\mathcal{N}(\boldsymbol{0}, \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{-1})')$, where matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ represent the matrices of the first derivative (with respect to $\boldsymbol{\beta}$) and the covariance, respectively, of the conditional-score function $\Psi(Y, G(\beta), \beta)$ in Equation (11). Matrix $\boldsymbol{A}$ can be consistently estimated by

$$\hat{\boldsymbol{A}} = n^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial\beta'}\Psi(Y_i, G_i(\beta), \beta)|_{\beta=\hat{\beta}} \tag{12}$$

and the solution $\hat{\beta}$ can be obtained via Newton–Raphson iterations:

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + \left[\hat{\boldsymbol{A}}^{(k)}\right]^{-1}n^{-1}\sum_{i=1}^{n}\Psi\left(Y_i, G_i(\hat{\beta}^{(k)}), \hat{\beta}^{(k)}\right), \tag{13}$$

where $\hat{\boldsymbol{A}}^{(k)}$ is $\hat{\boldsymbol{A}}$ in Equation (12), evaluated at its kth iteration, $\beta = \hat{\beta}^{(k)}$. In this paper, to handle the moderate/high-dimensional parameter $\boldsymbol{\beta}$, we consider a penalized estimating equation (Ma & Li, 2010) of the form

$$n^{-1}\sum_{i=1}^{n}\Psi(Y_i, G_i(\beta), \beta) - \boldsymbol{p}_\lambda'(|\beta|)\text{sign}(\beta) = \boldsymbol{0}, \tag{14}$$

and use a coordinate descent (CD) algorithm to iteratively solve Equation (14) for $\boldsymbol{\beta}$. In Equation (14), the function $\boldsymbol{p}_\lambda'(|\beta|) := (0, p_\lambda'(|\beta_1|), \ ..., p_\lambda'(|\beta_d|)) \in \mathbb{R}^{d+1}$ represents the vector of the first derivatives of a penalty function $p_\lambda(\cdot)$ evaluated at $|\boldsymbol{\beta}|$ (notice that

intercept b_0 is not penalized), $\text{sign}(\beta) = (\text{sgn}(b_0),\ \text{sgn}(\beta_1),\ \ldots, \text{sgn}(\beta_d))'$, where $\text{sgn}(t) := I(t>0) - I(t<0)$, and the notation $\boldsymbol{p}'_\lambda(|\beta|)\text{sign}(\beta)$ denotes the component-wise product between the two terms. In Equation (14), if the SCAD (Fan & Li, 2001) penalty is used, we can write

$$p'_\lambda(|\beta|) = \lambda I(|\beta| \le \lambda) + \frac{(a\lambda - |\beta|)_+}{a-1} I(|\beta| > \lambda), \tag{15}$$

for $|\beta|>0$, and $p'_\lambda(0) = 0$, in which $\lambda>0$ needs to be selected to optimize the estimation/covariate selection performance, and $a=3.7$ is often used in many applications.

Paralleling iteration (13), given the (kth) iteration $\hat{\beta}^{(k)}$, the subsequent Newton–Raphson iteration associated with Equation (14) is given by

$$\hat{\beta}^{(k+1)} = \underset{\beta}{\text{solve}} \left\{ n^{-1} \sum_{i=1}^{n} w_i^{(k)} G_i^{(k)} \left(r_i^{(k)} - G_i^{(k)'} \beta \right) - \boldsymbol{p}'_\lambda(|\beta|)\text{sign}(\beta) = \boldsymbol{0} \right\}, \tag{16}$$

in which the score function $\Psi(Y_i, G_i(\beta)|\beta)$ in Equation (14) is linearly approximated at $\beta = \hat{\beta}^{(k)}$, with

$$\begin{aligned} w_i^{(k)} &= \dot{F}(G_i^{(k)'} \hat{\beta}^{(k)}) \\ r_i^{(k)} &= G_i^{(k)'} \hat{\beta}^{(k)} + [\dot{F}(G_i^{(k)'} \hat{\beta}^{(k)})]^{-1} [Y_i - F(G_i^{(k)'} \hat{\beta}^{(k)})], \end{aligned} \tag{17}$$

where $G_i^{(k)}$ denotes the vector $G_i(\boldsymbol{\beta})$ in Equation (10) evaluated at the (kth) iteration $\hat{\beta}^{(k)}$, and $\dot{F}(\cdot)$ is the first derivative of the (inverse-logit) function $F(\cdot)$ with respect to $\cdot$. The Newton–Raphson step (16) constitutes the (kth) "outer" loop of solving Equation (14). We will use a CD algorithm to obtain the subsequent iteration $\hat{\beta}^{(k+1)}$ of the "outer" step (16); this constitutes the "inner" CD loop, defined within each (the kth) "outer" loop step (16). We use the notation $\hat{\beta}^{(k,m)}$ to keep track of the (mth) "inner" CD iteration, defined within each (kth) step of Equation (16). Jiang and Huang (2014) developed an efficient implementation of CD, termed the majorization minimization by CD (MMCD), which we employ in this paper to obtain Equation (16), for each (kth) "outer" loop step. CD cyclically updates each (jth) coordinate while holding the other coordinates fixed, until convergence of its iteration:

$$\hat{\beta}_j^{(k,m)} = (\hat{\beta}_0^{(k,m+1)},\ \ldots, \hat{\beta}_j^{(k,m+1)}, \hat{\beta}_{j+1}^{(k,m)},\ \ldots, \hat{\beta}_d^{(k,m)})' \in \mathbb{R}^{1+d} \quad (j = 0,\ 1,\ \ldots,\ d), \tag{18}$$

which represents the value of the mth "inner" iteration $\hat{\beta}^{(k,m)}$, at the time of the jth coordinate's update. Given the equations in Equation (16), a typical CD updates the iteration $\hat{\beta}_{j-1}^{(k,m)}$ to $\hat{\beta}_j^{(k,m)}$ by solving the following equation:

$$n^{-1} \left\{ \sum_{i=1}^{n} w_i^{(k)} (G_{ij}^{(k)})^2 \right\} \beta_j + p'_\lambda(|\beta_j|)\text{sign}(\beta_j) - n^{-1} \sum_{i=1}^{n} w_i^{(k)} G_{ij}^{(k)} \left(r_i^{(k)} - G_i^{(k)'} \hat{\beta}_{j-1}^{(k,m)} \right) - n^{-1} \left\{ \sum_{i=1}^{n} w_i^{(k)} (G_{ij}^{(k)})^2 \right\} \hat{\beta}_j^{(k,m)} = 0, \tag{19}$$

for $\beta_j \in \mathbb{R}$ and plugging the solution into the jth "coordinate," $\hat{\beta}_j^{(k,m+1)}$, of iteration (18). (In Equation 19, $G_{ij}^{(k)} \in \mathbb{R}$ represents the jth element of $G_i^{(k)} \in \mathbb{R}^{d+1}$.) For Equation (19), the MMCD approach of Jiang and Huang (2014) assumes a majorization of the "scaling" factor $\{\sum_{i=1}^n w_i^{(k)} (G_{ij}^{(k)})^2\}$ attached to β_j, by some constant $M>0$, that is, $\sum_{i=1}^n w_i^{(k)} (G_{ij}^{(k)})^2 \le M$, for all $j=0,\ldots,d$ (for each k). (That this majorization is useful for solving Equation 14 will be explained at the end of this section.) Given such M, the update rule associated with Equation (18), using the SCAD penalty (15), is given by

$$\hat{\beta}_j^{(k,m+1)} = \frac{\text{sgn}(\tau_j)(|\tau_j| - \lambda)_+}{M} \boldsymbol{I}_{|\tau_j| \le (1+M)\lambda} + \frac{\text{sgn}(\tau_j)(|\tau_j| - a\lambda/(a-1))_+}{M - 1/(a-1)} \boldsymbol{I}_{(1+M)\lambda < |\tau_j| \le a\lambda M} + \frac{\tau_j}{M} \boldsymbol{I}_{|\tau_j| > a\lambda M}, \tag{20}$$

where $\tau_j = M\hat{\beta}_j^{(k,m)} + n^{-1} \sum_{i=1}^n G_{ij}^{(k)} (Y_i - F(G_i^{(k)'} \hat{\beta}_{j-1}^{(k,m)}))$, for $j=1,\ldots,d$. For the intercept, that is, for $j=0$,

$$\hat{\beta}_0^{(k,m+1)} = \tau_0 / M, \tag{21}$$

where $\tau_0 = M\hat{\beta}_0^{(k,m)} + n^{-1} \sum_{i=1}^n G_{i1}^{(k)} (Y_i - F(G_i^{(k)'} \hat{\beta}^{(k,m)}))$, in which $\hat{\beta}^{(k,m)} = (\hat{\beta}_0^{(k,m)}, \hat{\beta}_1^{(k,m)},\ \ldots, \hat{\beta}_d^{(k,m)})'$. The convergence of iteration (18) (over the "inner" loop m) defined by the rules (20) and (21) to a solution of the (kth) equations of Equation (16) (for each k) is given by the Theorem 1 of Jiang and Huang (2014). (In Equation 20, $(\cdot)_+$ denotes the thresholding operator.)

For each $j \in \{1,\ldots,d\}$, if we standardize $\{(G_{ij}^{(k)})_{i=1}^n\}$ to have (mean 0 and) unit variance, then the process of seeking a majorization constant M, which satisfies $\sum_{i=1}^n w_i^{(k)} (G_{ij}^{(k)})^2 \le M$, for all j and k, is reduced to finding an upper bound for the weights $w_i^{(k)}$, uniformly over all i and k. In logistic regression, the weight $w_i^{(k)}$ in Equation (17) can be written as $w_i^{(k)} = \hat{\mu}_i^{(k)}(1 - \hat{\mu}_i^{(k)})$, for some probability $\hat{\mu}_i^{(k)}$ (i.e., $0 \le \hat{\mu}_i^{(k)} \le 1$), which implies

$w_i^{(k)} \leq 1/4$, for all i and k. Thus, the uniform upper bound M can be set at 1/4 for the case of a logistic regression. We summarize below the proposed algorithm for estimating model (2) via solving Equation (14).

Algorithm 1 Estimation procedure

LFPCA: Estimate LFPC model (3) and obtain $\boldsymbol{\Phi}_X^{(0)}$, $\boldsymbol{\Phi}_X^{(1)}$, $\boldsymbol{\Phi}_U$, and $\tilde{X}_i$.
Initialization: Obtain a naive estimate, $\hat{\boldsymbol{\beta}}^{(0)}$, by estimating GLM (5), treating $\tilde{X}_i$ as error free.
Given a set of the tuning parameters (λ, a):
for the "outer" loop $k = 0, 1, 2, \ldots$, until convergence of $\hat{\boldsymbol{\beta}}^{(k)}$, **do**
  Given $\hat{\boldsymbol{\beta}}^{(k)}$, update $G_i^{(k)}$ by Equation (10), and centre and scale $G_i^{(k)}$.
  Given $G_i^{(k)}$ (and with $M = 1/4$), obtain $\hat{\boldsymbol{\beta}}^{(k+1)}$ by:
  for the "inner" loop $m = 1, 2, \ldots$, until convergence of the $\hat{\boldsymbol{\beta}}_j^{(k,m)}$'s in Equation (18), **do**
    update $\hat{\boldsymbol{\beta}}_j^{(k,m)}$ by Equation (21), for $j = 0$;
    update $\hat{\boldsymbol{\beta}}_j^{(k,m)}$ by Equation (20), for $j = 1, \ldots, d$.
  end for
  Scale the iteration $\hat{\boldsymbol{\beta}}^{(k+1)}$ on the original scale of $G_i^{(k)}$.
end for

Throughout the paper, λ>0 is selected by maximizing a five-fold cross-validated averaged predictive area under the receiver operating characteristic curve, and a=3.7. Given an estimate of $\tilde{\beta}_1$ of Equation (5), the estimate of $\boldsymbol{\beta}_1$ of Equation (2) can be obtained via relationship (4).

As a consequence of the majorization with M, updating rules (20) and (21) does not depend on the quantities $w_i^{(k)}$ and $r_i^{(k)}$ in Equation (17). Therefore, we do not need to update $w_i^{(k)}$ and $r_i^{(k)}$ but only update $G_i^{(k)}$ for each "outer" loop step k, which simplifies Algorithm 1. Moreover, the majorization on the "scaling" factor $\{\sum_{i=1}^n w_i^{(k)}(G_{ij}^{(k)})^2\}$ of Equation (19) removes the potential numerical instability that might exist in solving Equation (14). As $w_i^{(k)}$ and $G_{ij}^{(k)}$ depend on $\hat{\beta}^{(k)}$, the "scaling" factor $\{\sum_{i=1}^n w_i^{(k)}(G_{ij}^{(k)})^2\}$ attached to β_j in Equation (19) is a highly non-linear function of $\hat{\beta}^{(k)}$. This implies that the scale of the solution β_j of Equation (19) can be highly sensitive to a small change in the values of the (kth) "outer" iteration, $\hat{\beta}^{(k)}$, which might cause instability over the "outer" iteration (16). (This instability is also expected for the Newton–Raphson iteration 13, due to the matrix of the first derivative A in Equation 12, which involves the derivative of the non-linear "linear predictor" $\boldsymbol{\beta}'G_i$ in Equation 11 with respect to $\boldsymbol{\beta}$, resulting in a high degree of non-linearity in the update rule of Equation 13.) However, by using the majorization and the associated MMCD in performing the "inner" loop step, Algorithm 1 becomes independent of this highly non-linear (k-dependent) scaling factor.

2.4 | Theoretical result

Concerns about the estimation bias due to the error in covariates in a moderate/high-dimensional random effect covariate setting prompted us to consider the penalized conditional-score equation of form (14). In this subsection, we provide an asymptotic property of the penalized estimating equation estimator $\hat{\beta}$ that solves Equation (14), which follows from the asymptotic property of the penalized estimating equation of Ma and Li (2010), when the dimension, $1+d_n$, of the parameter $\boldsymbol{\beta}$ increases along with the sample size n. Here, we write d as d_n to emphasize its dependence on n.

Let us denote $\beta^* = (b_0^*, \beta_1^*, \ldots, \beta_{d_n}^*)'$ as the true value of $\boldsymbol{\beta}$. Let $\alpha_n = \max\{|p'_{\lambda_n}(|\beta_j^*|)| : \beta_j^* \neq 0\}$, where we write λ as λ_n to emphasize its dependence on n. We state below regularity conditions on the conditional-score function $\Psi(Y, G(\boldsymbol{\beta}), \boldsymbol{\beta})$ and the penalty function $\boldsymbol{p}'_\lambda(|\beta|)$in Equation (14).

Assumption 1. The expectation of the first derivative of $\Psi(Y, G(\boldsymbol{\beta}), \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, that is, $\partial\Psi(Y, G(\boldsymbol{\beta}), \boldsymbol{\beta})/\partial\boldsymbol{\beta}'$, exists at $\boldsymbol{\beta}=\boldsymbol{\beta}^*$, and its eigenvalues are bounded below and above by positive constants. For any entry S_{jk} in $\partial\Psi(Y, G(\boldsymbol{\beta}^*), \boldsymbol{\beta}^*)/\partial\boldsymbol{\beta}'$, $\mathbb{E}(S_{jk}^2)$ is bounded above by a constant.

Assumption 2. The second derivatives of $\Psi(Y, G(\boldsymbol{\beta}), \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ exist, and the entries are uniformly bounded by some constant in a large enough neighbourhood of $\boldsymbol{\beta}^*$.

Assumption 3. For the penalty function, there exist constants C and D such that if $\gamma_1, \gamma_2 > C\lambda$, then $|p''_\lambda(\gamma_1) - p''_\lambda(\gamma_2)| \leq D|\gamma_1 - \gamma_2|$.

The SCAD penalty (15) satisfies Assumption 3, and Assumptions 1-2 are mild regularity conditions on $\Psi(Y, G(\boldsymbol{\beta}), \boldsymbol{\beta})$.

Theorem 1. *Under Assumptions 1-3, and if* $d_n^4 n^{-1} \to 0$, $\lambda_n \to 0$ *as* $n \to \infty$, *then, with probability tending to 1, there is an estimator* $\hat{\beta}$ *that solves Equation (14) such that* $\|\hat{\beta} - \beta^*\| = O_p\{\sqrt{d_n}(n^{-1/2} + \alpha_n)\}$.

The proof of Theorem 1 is provided in the Supporting Information. Theorem 1 indicates that the convergence rate depends on λ_n and the penalty function through α_n. To achieve root(n/d_n) convergence rate of $\hat{\beta}$, $\alpha_n = O(n^{-1/2})$. Note that, for the SCAD penalty (15), $\alpha_n = 0$ as $\lambda_n \to 0$ (for the L^1 penalty, $\alpha_n = \lambda_n$), and therefore the resulting penalized conditional-score estimate is root(n/d_n) consistent.

3 | SIMULATION STUDY

In Section 3.1, we present simulation studies for assessing the estimation performance of the proposed method, in settings where the outcome is regressed on a longitudinal functional covariate. In Section 3.2, we present the variable selection/estimation performance in moderate dimension.

3.1 | Simulation 1: Models with a longitudinal functional covariate

Following Greven et al. (2010), we consider the longitudinal functional covariates $W_{ij}(v)$ from

$$W_{ij}(v) = \sum_{k=1}^{N_X} \tilde{x}_{i,k} \Phi_{X,k}^{(0)}(v) + t_{ij} \sum_{k=1}^{N_X} \tilde{x}_{i,k} \Phi_{X,k}^{(1)}(v) + \sum_{k=1}^{N_U} \tilde{u}_{ij,k} \Phi_{U,k}(v) + \epsilon_{ij}(v), \tag{22}$$

where the LFPC scores $\tilde{x}_{i,k} \sim \mathcal{N}(0, \lambda_X^{(k)})$and $\tilde{u}_{ij,k} \sim \mathcal{N}(0, \sigma^2 \lambda_U^{(k)})$, in which $\lambda_X^{(k)} = \lambda_U^{(k)} = \frac{1}{2}^{k-1}$, k=1,2,..., and the scale parameter $\sigma^2 \in \{0.25, 0.5, 1, 1.5\}$ associated with the LFPC noise scores, $\tilde{u}_{ij,k}$, controls the contribution of the LFPC noise component to the variance of $W_{ij}(v)$. The term $\epsilon_{ij}(v)$ in Equation (22) accounts for the random homoscedastic white noise and is assumed to be i.i.d. $\mathcal{N}(0, s^2)$ for some $s^2 > 0$ (which will be specified later) independently of all other LFPC processes. We set N_X=8 and N_U=4. In the Supporting Information, we provide the sets of basis $\{\Phi_{X,k}^{(0)}(v),\ k = 1,\ \ldots,\ N_X\}$, $\{\Phi_{X,k}^{(1)}(v),\ k = 1,\ \ldots,\ N_X\}$, and $\{\Phi_{U,k}(v), k=1,\ldots,N_U\}$ of Equation (22), used to specify a nonsparse (Set "A") or sparse (Set "B") LFPC model.

To obtain a (length p) vector $W_{ij} (\in \mathbb{R}^p)$ in the framework of model (1), we evaluate model (22) on the grid of $p \in \{50, 200\}$ equidistant points in [0,1]. Each of the evaluated LFPC basis vectors is normalized to have a unit L^2 norm, and the corresponding (length p) vectors $W_{ij} (\in \mathbb{R}^p)$ are obtained from Equation (22). The visit time points, t_{ij}, are generated from Unif(0,J_i), sorted in an increasing order, and standardized to have mean 0 and unit variance, for each subject i. We set the number of visits J_i=4 for all i=1,...,n. With the normalized LFPC basis and the scaled and centred visit times, the signal-to-noise ratio (SNR) of the LFPC component of covariate model (22) is $\left(\sum_{k=1}^{8} \lambda_X^{(k)} + \sum_{k=1}^{4} \sigma^2 \lambda_U^{(k)}\right) / (p s^2)$. We set s^2=0.05 for p=50 (and s^2=0.0125 for p=200), setting SNR of the LFPC component of Equation (22) approximately at 1, for the case of σ^2=0.25. This setting of (LFPC) SNR=1 with σ^2=0.25 is similar to that of the data set analysed in Section 4.

We generate the "error-free" covariates $Z_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$. Then we generate the outcome $Y_i \in \{0,1\}$ using model (5), on the basis of the longitudinal scores $\tilde{X}_i = (\tilde{x}_{i1},\ \ldots, \tilde{x}_{iN_X})'$ obtained from covariate model (22). In Equation (5), we set $\boldsymbol{\beta}_0 = (1, 0.5)'$, $\tilde{\beta}_1 = \left(\tilde{\beta}_{11},\ \ldots, \tilde{\beta}_{18}\right)' = (4, -2, 1, -0.5, 0, 0, 0, 0)' \in \mathbb{R}^{N_X}$, and b_0=0. In the framework of outcome model (2), this sets the "true" $\beta_1 (\in \mathbb{R}^{2p})$ (associated with the $2p$ number of random intercepts/slopes) as $\beta_1 = \Phi_X \tilde{\beta}_1$, in which $\Phi_X = \left(\left(\tilde{\Phi}_{X,1}^{(0)};\ \ldots; \tilde{\Phi}_{X,N_X}^{(0)}\right)', \left(\tilde{\Phi}_{X,1}^{(1)};\ \ldots; \tilde{\Phi}_{X,N_X}^{(1)}\right)'\right)'$ (a $2p \times N_x$ matrix), where the $p \times 1$ vectors $\tilde{\Phi}_{X,k}^{(0)}$ and $\tilde{\Phi}_{X,k}^{(1)}$ correspond to the functions $\Phi_{X,k}^{(0)}(v)$ and $\Phi_{X,k}^{(1)}(v)$in Equation (22), respectively, evaluated (and normalized to have unit norm) at the p grid points.

To compare the performance of different estimation methods, we report the estimation error, defined as $\|\hat{\beta}_1 - \beta_1\| / \sqrt{2p}$, in which $\hat{\beta}_1$ denotes an estimate of the "true" parameter $\boldsymbol{\beta}_1$ in model (2). In the Supporting Information, we also provide the estimation error, $\|\hat{\beta}_0 - \beta_0\|$, of the coefficient vector $\boldsymbol{\beta}_0$ associated with the error-free covariates Z_i. We consider the cases with the number of subjects $n \in \{100, 200, 400\}$, the number of evaluation points $p \in \{50, 200\}$, and the basis type {"A","B"} which specifies nonsparse ("A") or sparse ("B") basis for covariate model (22). We compare the following three methods of estimating the coefficients.

- **Cond.(LFPCA)**: The proposed conditional method that solves Equation (14) by Algorithm 1; the approach utilizes the LFPCA dimension reduction described in Section 2.1 and the SCAD penalization described in Section 2.3.
- **Naive(LFPCA)**: A naive method ignoring the noise in $\tilde{X}_i$, where reduced model (5) is estimated by maximizing the SCAD-penalized likelihood, given a set of $\tilde{X}_i$ from LFPCA reduction (3). For implementation, Algorithm 1 with the single "outer" loop (i.e., k=0 only) is utilized, in which the vector $(1, Z_i', \tilde{X}_i')'$ takes the part of $G_i^{(k)}$ throughout the whole algorithm (i.e., the update on $G_i^{(k)}$ is not performed).
- **Naive(Variable-wise)**: A naive approach ignoring the noise in X_i, in which model (2) is estimated by maximizing the elastic-net (Zou & Hastie, 2005) penalized likelihood, given a set of X_i obtained from p number of separate variable-wise regressions (i.e., LFPCA reduction 3 is not

performed). The mixing parameter of the elastic-net penalty is set to be 0.5, and its sparsity parameter is selected by maximizing a five-fold cross-validated averaged predictive area under the receiver operating characteristic curve.

For the LFPCA-based methods Cond.(LFPCA) and Naive(LFPCA), to determine appropriate values of N_X and N_U for representation (3), we employ the approach suggested by Greven et al. (2010) that uses the proportion of variance explained as a "cut-off" criterion. Throughout this example, we use a cut-off value that explains 85% of the observed longitudinal data variations. The elastic-net regularization is often considered as a more appealing method of regularization for highly correlated covariatesthan the SCAD regularization, which tends to select only one covariate from a group of highly correlated covariates and ignore others, and this is why the elastic-net regularization is employed for the method Naive(Variable-wise). We conduct simulations 100 times for each scenario.

The boxplots of the estimation errors for the methods Cond.(LFPCA), Naive(LFPCA), and Naive(Variable-wise) are reported in Figures 1 and 2, for the settings "A" and "B," respectively. In all scenarios, the proposed method, Cond.(LFPCA), of solving Equation (14) outperforms the naive approaches ignoring the noise in covariates and results in a smaller average estimation error. In both settings "A" and "B," when we compare the two LFPCA-based approaches, Cond.(LFPCA) and Naive(LFPCA), the benefit of employing the conditional method (that solves Equation 14) becomes more evident, as the sample size n and the noise variance σ^2 increase. Naive(Variable-wise), which uses a set of estimates of the intercepts and slopes obtained from p number of variable-wise regressions as covariates, performs very poorly due to the high correlations among the random intercepts and slopes. In the Supporting Information, we additionally provide an illustration for computational efficiency of Algorithm 1 used in implementing Cond.(LFPCA) for this simulation example, in comparison with Naive(LFPCA) and Naive(Variable-wise).

3.2 | Simulation 2: Variable selection and estimation performance

We investigate the performance of variable selection for significant random intercepts/slopes and that of estimation accuracy, with a varying intensity of noise level in covariates in a moderate dimensional setting. The LFPCA dimension reduction for X_i is not performed to distinguish the effect of LFPCA from that of using the (SCAD-penalized) measurement error models on the performance (see Remark 1 for the notation in

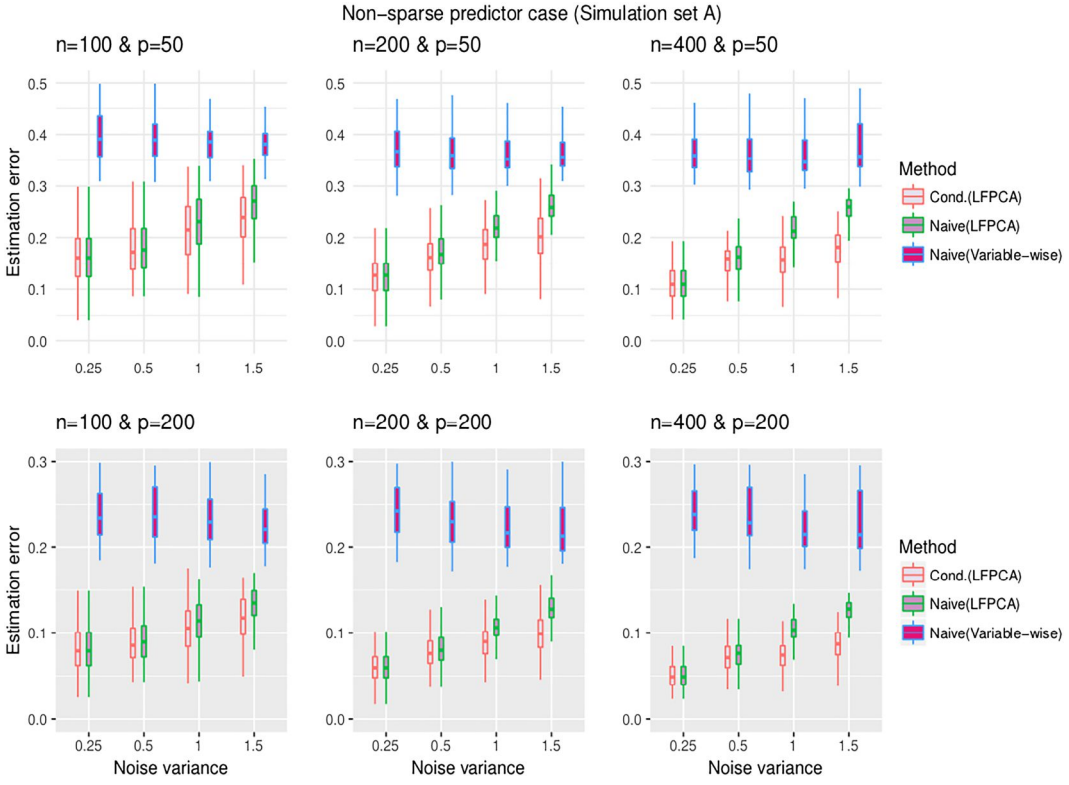

FIGURE 1 Boxplots of the estimation errors, obtained from 100 simulation runs, comparing (1) Cond.(LFPCA): the LFPCA-based conditional method accounting for the noise in covariates; (2) Naive(LFPCA): the LFPCA-based naive method ignoring the noise in covariates; and(3) Naive(Variable-wise): the variable-wise naive method ignoring the noise in covariates estimated via the elastic-net penalization. Each panel represents one of the combinations of $n \in \{100,200,400\}$ and $p \in \{50,200\}$, as a function of a varying noise (in covariates) level $\sigma^2 \in \{0.25,0.5,1.0,1.5\}$

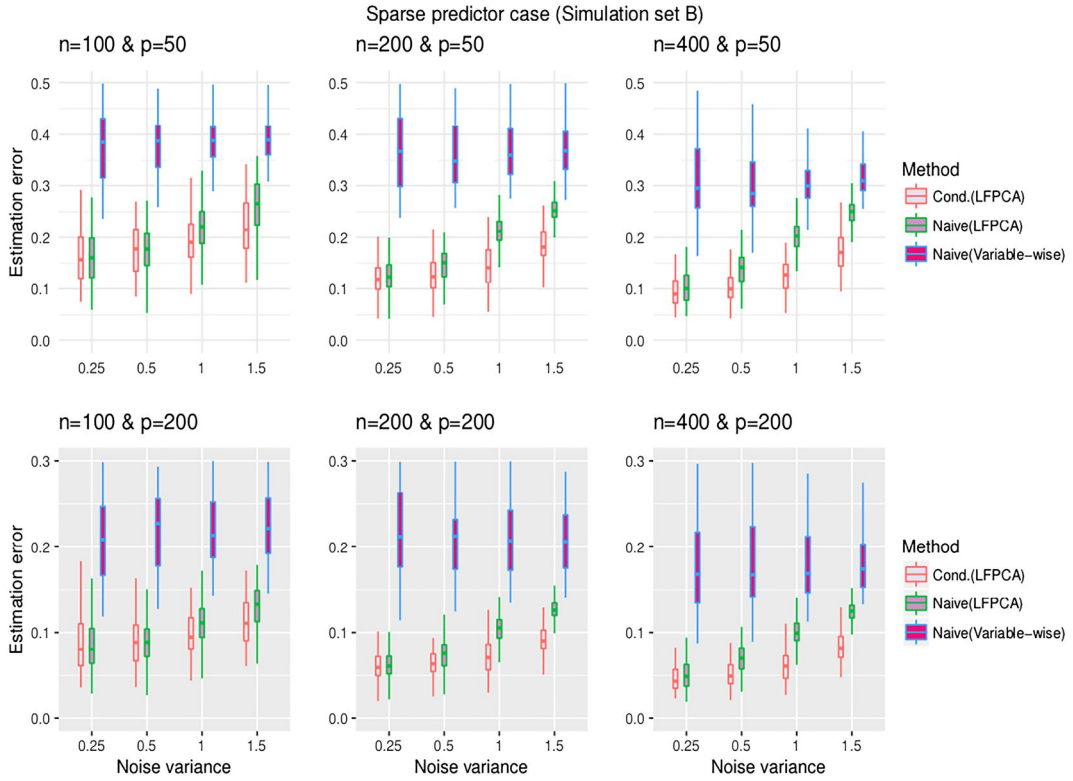

FIGURE 2 Boxplots of the estimation errors, obtained from 100 simulation runs, comparing (1) Cond.(LFPCA): the LFPCA-based conditional method accounting for the noise in covariates; (2) Naive(LFPCA): the LFPCA-based naive method ignoring the noise in covariates; and (3) Naive(Variable-wise): the variable-wise naive method ignoring the noise in covariates estimated via the elastic-net penalization. Each panel represents one of the combinations of $n \in \{100,200,400\}$ and $p \in \{50,200\}$, as a function of a varying noise (in covariates) level $\sigma^2 \in \{0.25,0.5,1.0,1.5\}$

this setting). We take the sample size to be $n \in \{250,500\}$, and the number of subject-specific random intercepts and slopes from p independent regions, $2p \in \{10,20,30,40\}$. For each scenario, we conduct simulations 200 times.

For each simulation run, we generate $Y_i \in \{0,1\}$ on the basis of model (2), with $\beta_0=0$ and $\beta_1 = (\underbrace{1.5,1,0.5}_{3 \text{ active intercepts}}, \underbrace{0,\ldots,0}_{p-3}, \underbrace{0,\ldots,0}_{p-3}, \underbrace{0.5,1,1.5}_{3 \text{ active slopes}})' \in \mathbb{R}^{2p}$, that is, there are six active random intercepts/slopes associated with Y_i. The error-free covariates Z_i are not considered in this example. The $2p \times 1$ vector of random effect vectors $X_i = (X_i^{(0)'}, X_i^{(1)'})'$, with $X_i^{(0)} \overset{D}{=} X_i^{(1)} \in \mathbb{R}^p$ are generated from the multivariate normal distribution with the identity covariance $\mathbf{\Sigma}_X = \mathbf{I}_{2p}$. The measurement error $U_{ij} \in \mathbb{R}^p$ is generated from $U_{ij} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \sum_U)$, with a $p \times p$ auto-regressive correlation matrix $\mathbf{\Sigma}_U$ where its (a,b)th entry is given as $0.5^{|a-b|}$ for any $a,b \in \{1,\ldots,p\}$. The intensity of the covariate noise U_{ij} is controlled by $\sigma^2 \in \{0.25,0.5,1,1.5\}$. The longitudinal vectors $W_{ij} \in \mathbb{R}^p$ are generated on the basis of model (1) with $J_i=4$, for each $i=1,\ldots,n$. The visit time points t_{ij} are generated from Unif$(0,J_i)$, sorted in an increasing order, and then shifted and scaled to have mean 0 and unit variance. Given each set of data, MLE was used for estimating the noise variance $\sigma^2 \sum_U$ associated with X_i, assuming the "unstructured" correlation structure on $\mathbf{\Sigma}_U$ and that X_i are uncorrelated. We consider the following four approaches for fitting model (2).

- **Cond.(SCAD)**: The proposed conditional method that solves Equation (14) by Algorithm 1 (see Remark 1 for the notation) using the SCAD penalization.
- **Cond.(Full)**: The conditional method that solves Equation (11) by the Newton–Raphson iteration (13), without employing any penalization ("Full").
- **Naive(SCAD)**: A naive method ignoring the noise in X_i, where model (2) is estimated by maximizing the SCAD-penalized likelihood, given a set of X_i obtained from MLE of model (1). For implementation, Algorithm 1 with the single "outer" loop (i.e., $k=0$ only) is utilized, in which the vector $(1, X_i')'$ takes the part of $G_i^{(k)}$ throughout the whole algorithm (i.e., the update on $G_i^{(k)}$ is not performed).
- **Naive(Full)**: A naive method ignoring the noise in X_i, where model (2) is estimated by maximizing the likelihood, given a set of X_i obtained from MLE of model (1), implemented by the standard GLM routine implemented in R (R Core Team, 2019), without penalization ("Full").

First, in Figure 3, for the $2p \in \{30,40\}$ cases, we compare the results of the variable selection performance of the two approaches that incorporated a simultaneous covariate selection procedure: Cond.(SCAD) and Naive(SCAD). We omitted the results for $2p \in \{10,20\}$, as they were qualitatively similar to their higher dimensional counterparts $2p \in \{30,40\}$. In Figure 3, we report the proportions of *correctly selecting* the relevant random effects (i.e., "correct signals") out of the six true *nonzeros*, and those of *correctly unselecting* the irrelevant random effects (i.e., "correct zeros") out of the $2p-6$ true *zeros*. The naive estimators would be consistent to a vector that is different from the true $\boldsymbol{\beta}_1$ because of the noise in covariates (Liang & Li, 2009). Therefore, ignoring the noise in covariate would tend to falsely classify the irrelevant coefficients as significant ones. In Figure 3, the results exactly demonstrate this point, indicating that ignoring the covariate noise causes errors in correctly unselecting the true zeros, resulting in selection of spurious covariates and a more complex model. In terms correctly identifying the "signal" random effects, the proposed method also outperforms the naive method in all scenarios. Overall, the method incorporating the covariate errors outperforms the naive method.

Second, we compare the estimation accuracy of the four methods, on the basis of the estimation error defined as $\|\hat{\beta}_1 - \beta_1\|/\sqrt{2p}$, in which $\hat{\beta}_1$ denotes an estimate of $\boldsymbol{\beta}_1$. Figure 4 reports the averaged estimation error over the 200 simulation runs, for each combination of $2p \in \{10,20,30,40\}$ (displayed from top to bottom) and $n \in \{250,500\}$ (displayed from left to right), with a varying noise level in covariates, $\sigma^2 \in \{0.25,0.5,1,1.5\}$. The level of the noise in covariates certainly affects the estimation performance of all methods, as the performance generally deteriorates with an increasing covariate measurement noise σ^2. However, in all cases, the method Cond.(SCAD) outperforms both of the naive approaches. Although the unpenalized method (Cond.(Full)) performs at a similar level as the penalized one (Cond.(SCAD)) for $n=500$, it suffers from a substantial instability in larger $2p$ settings given a smaller sample size, for example, $2p \in \{30,40\}$ with $n=250$. In particular, the unpenalized method (Cond.(Full)) is often outperformed by the naive approaches (Naive.(SCAD) and Naive.(Full)). On the other hand, Cond.(SCAD) outperforms the naive approaches in all cases.

4 | APPLICATION

In this section, the method is applied to the ADNI data to identify biomarkers for dementia transition in MCI participants. The data were downloaded from the ADNI database (http://adni.loni.usc.edu). The initial phase (ADNI-I) recruited 800 participants, including approximately 200 healthy controls, 400 patients with late MCI, and 200 patients clinically diagnosed with probable AD over 50 sites across the United States and Canada and followed up at 6- to 12-month intervals for 2–3 years. ADNI has been followed by ADNI-GO and ADNI-2 for existing participants

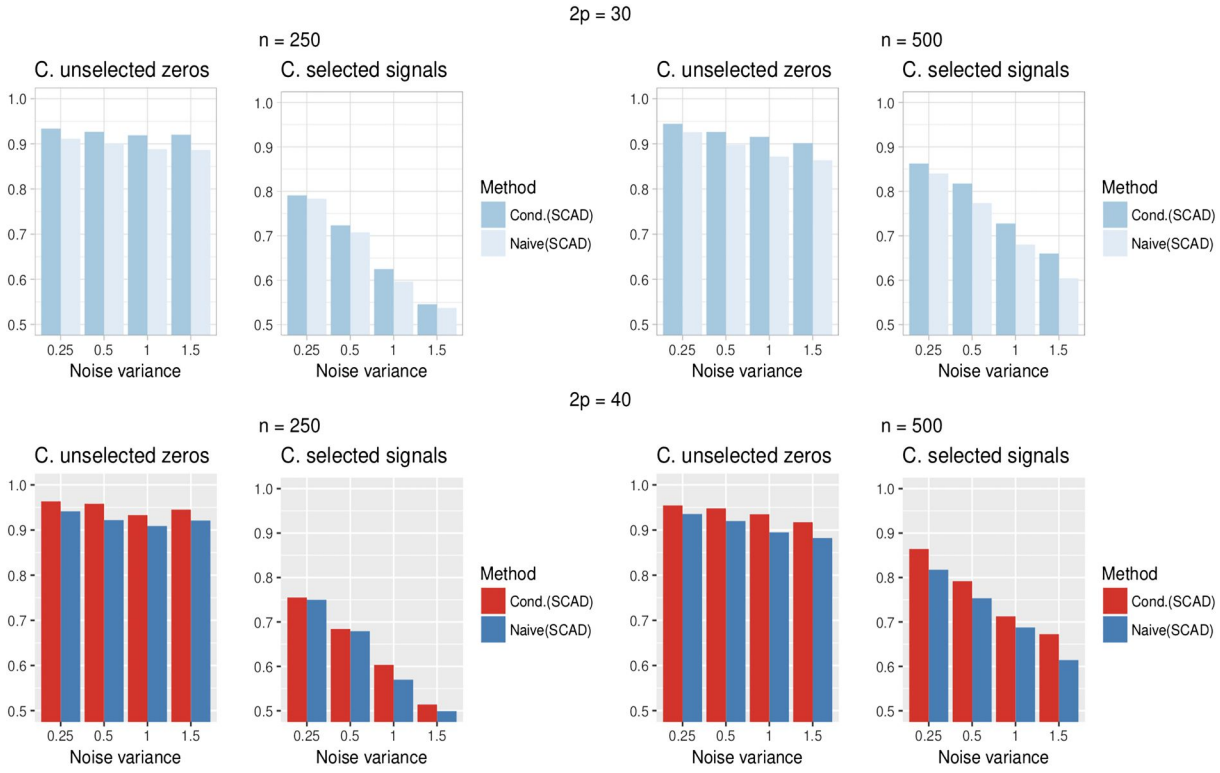

FIGURE 3 The proportions of correctly unselected (C. unselected) "zero" covariates and those of correctly selected (C.selected) "signal" covariates, averaged over 200 replications, comparing Cond.(SCAD) and Naive(SCAD), with a varying noise (in covariates) level σ^2 *in* {0.25,0.5,1.0,1.5}. The top (bottom) row corresponds to the $2p=30$($2p=40$) cases; the left (right) two columns corresponds to the $n=250$ ($n=500$) cases. The larger values indicate superior performance

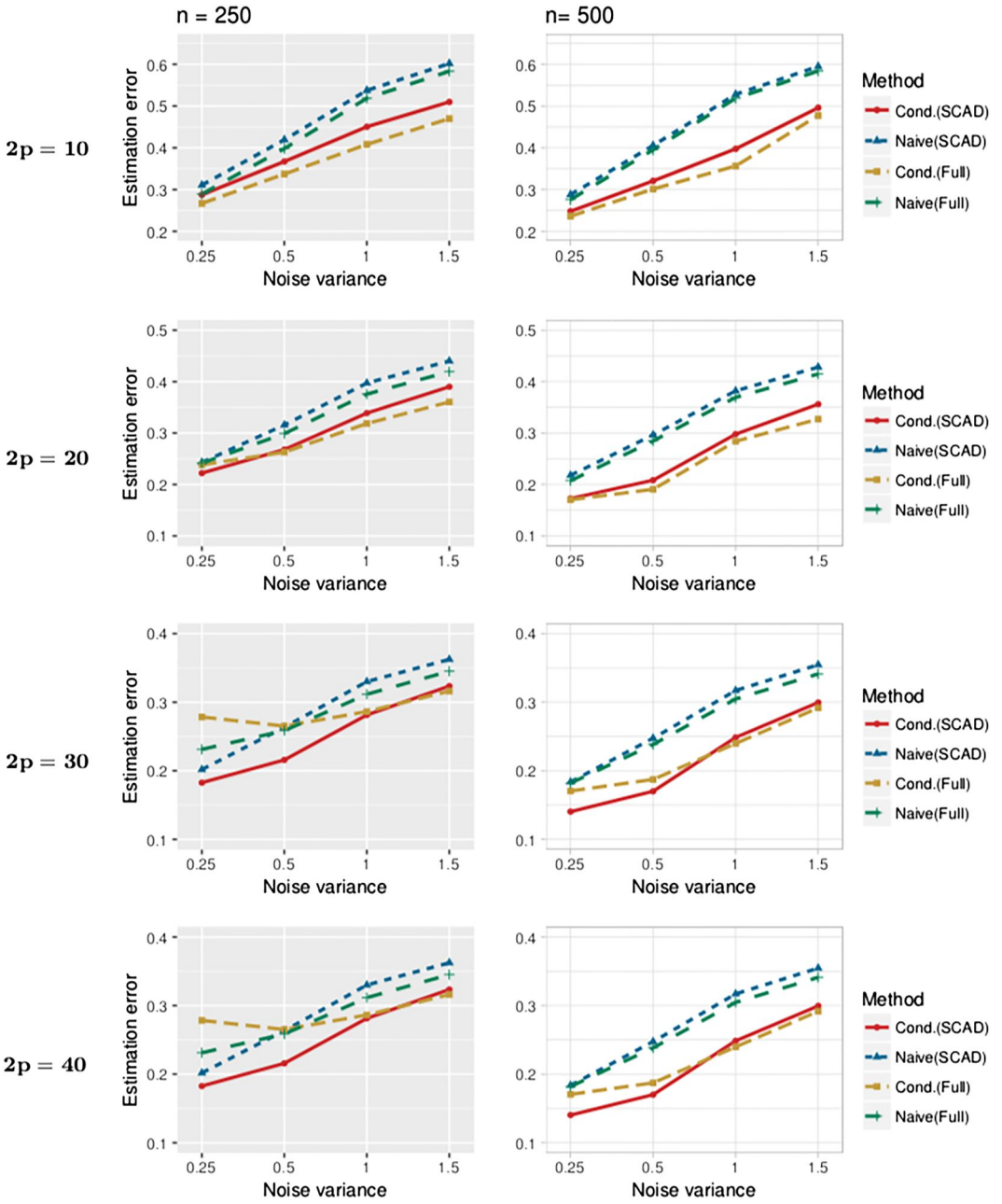

FIGURE 4 The estimation errors averaged over 200 simulation runs, comparing the methods Cond.(SCAD), Naive(SCAD), Cond.(Full), and Naive(Full), for each covariate dimension $2p$ *in* {10,20,30,40} (from top to bottom), with a varying noise (in covariates) level σ^2 *in* {0.25,0.5,1.0,1.5}; the left (right) panels correspond to n=250 (n=500) cases

and enrolled additional individuals, including early MCI. To be classified as MCI in ADNI, a subject needed an inclusive Mini-Mental State Examination score of between 24 and 30, subjective memory complaint, objective evidence of impaired memory calculated by scores of the Wechsler Memory Scale Logical Memory II adjusted for education, a score of 0.5 on the Global Clinical Dementia Rating, absence of significant confounding conditions such as current major depression, normal or near normal daily activities, and absence of clinical dementia.

ADNI used 1.5-T and 3.0 MP-RAGE T1-weighted MR images that were later preprocessed and corrected for non-linearity via "GradWarp." The scans were implemented using a standardized ADNI protocol adjusted for use at each specific collection site and then underwent scaling and vetting to meet quality control criteria. Cross-sectional image processing was performed using FreeSurfer Version 4.3 by researchers at UCSF group (http://adni.loni.usc.edu/methods/mri-analysis/). Region of interest (ROI)-specific cortical thickness measures were extracted from the automated FreeSurfer 5.1 anatomical parcellation using the Desikan–Killiany atlas (Desikan et al., 2006); there were 68 ROIs (34 each on the left and right hemispheres), in which the longitudinal cortical thickness measures were collected. For internal consistency, we focused on subjects with 1.5-T magnetic resonance imaging scans and diagnosed as MCI at the screening. We excluded participants who converted back to cognitive normal and who had images that did not pass the quality check, yielding 339 subjects. Out of the 339 subjects, we included n=221 patients who had

at least two visits during 1–4 years. The 118 subjects who were not followed up for at least 1 year were excluded from the analysis. Of the 221 subjects analysed, 101 subjects were translated to dementia before their last available records, whereas 120 subjects remained as MCI.

The dementia transition was the primary outcome Y_i of the study, coded as 1 for "demented" and 0 for "not-demented." Age and gender were considered as "error-free" covariates Z_i. The average age was 73.89 with *SD* of 7.12, and about 40% of the subjects were female. Under model (1), we estimate the subject-specific random intercepts $X_i^{(0)} \in \mathbb{R}^{68}$ (i.e., the baseline cortical thickness) and the slopes $X_i^{(1)} \in \mathbb{R}^{68}$ (i.e., the change rates) associated with the cortical thickness of 68 ROIs, and we collectively denoted them as $X_i = \left(X_i^{(0)'}, X_i^{(1)'}\right)' \in \mathbb{R}^{136}$. These subject-specific random effects X_i associated with the ROIs were considered as potential biomarkers for dementia transition, and model (2) was estimated. For estimation of joint models (1) and (2), we used three methods, that is, Cond.(LFPCA), Naive(LFPCA), and Naive(Variable-wise), described in Section 3.1. For Cond.(LFPCA) and Naive(LFPCA), the numbers of the LFPC components, N_X and N_U, of LFPC representation (3), were chosen to explain 75% of the total variability of the longitudinal data, resulting in N_X=12 for representing the random effect covariates X_i, and N_U=2 for representing the noise U_{ij}. Beyond the 75% cut-off, the estimated eigenvalues of the LFPC model were negligibly small.

Figures 5 and 6 display the estimated regression coefficients associated with the random intercepts $X_i^{(0)}$ and the random slopes $X_i^{(1)}$, respectively. To identify significant biomarkers for the dementia progression, we computed 95% bootstrap confidence intervals for the coefficient estimates on the basis of 500 bootstrap replications. For each bootstrap replication, we sampled with replacement n(=221) quadruplets $\{(Y_i, W_i, Z_i, \mathbf{t}_i), i=1,\ldots,221\}$, each time fitting regression model (2) on the basis of the resampled quadruplets; for Cond.(LFPCA) and Naive(LFPCA), we use LFPC dimension reduction model (3) (with a 75% variance cut-off for the values N_X and N_U) to represent covariate model (1); for

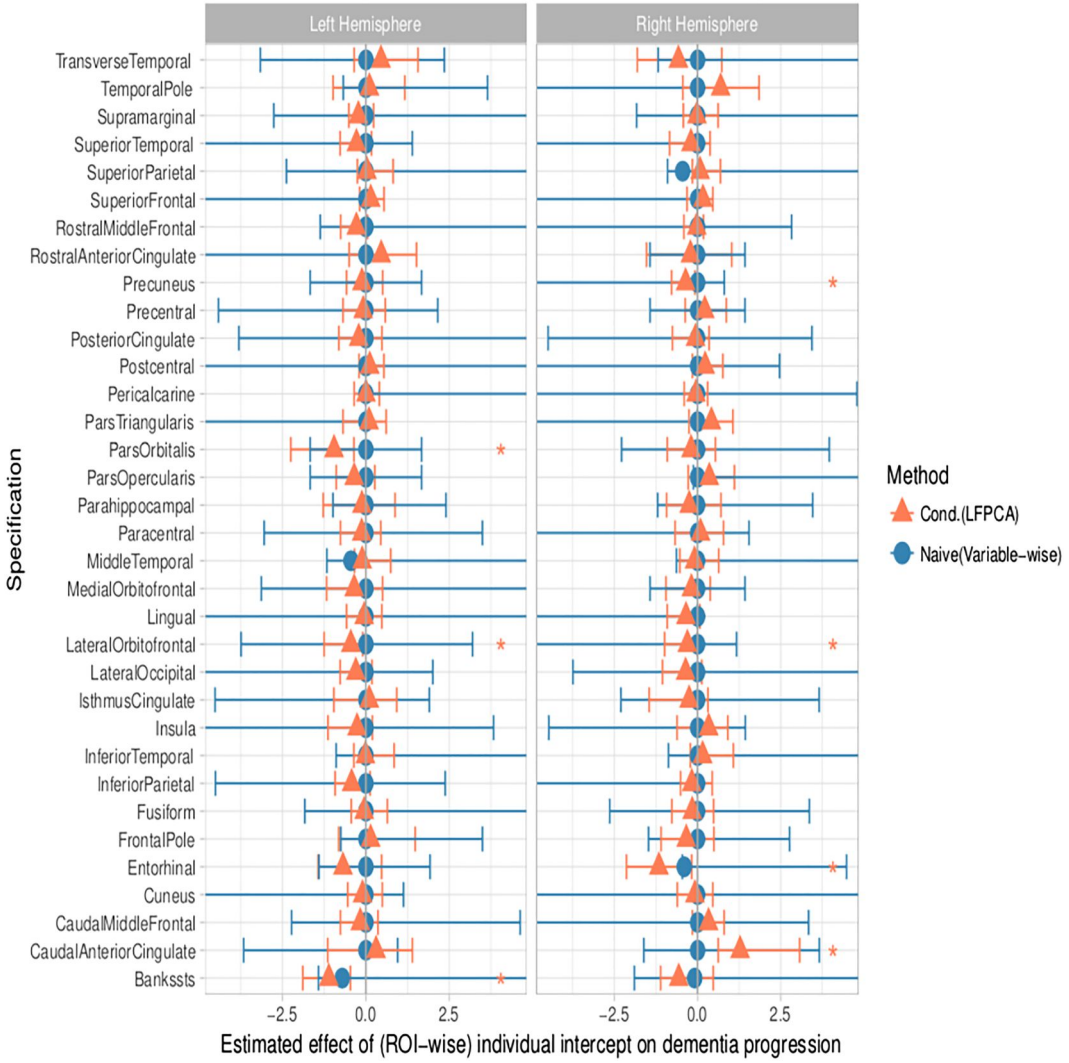

FIGURE 5 The coefficient estimates associated with the random intercepts that correspond to the 34 regions of interest (ROIs) of the left hemisphere (in the left panel) and of the right hemisphere (in the right panel), comparing Cond.(LFPCA) and Naive(Variable-wise), with 95% confidence intervals obtained from 500 bootstrap replications. Estimated significant ROIs are marked with an asterisk on the right

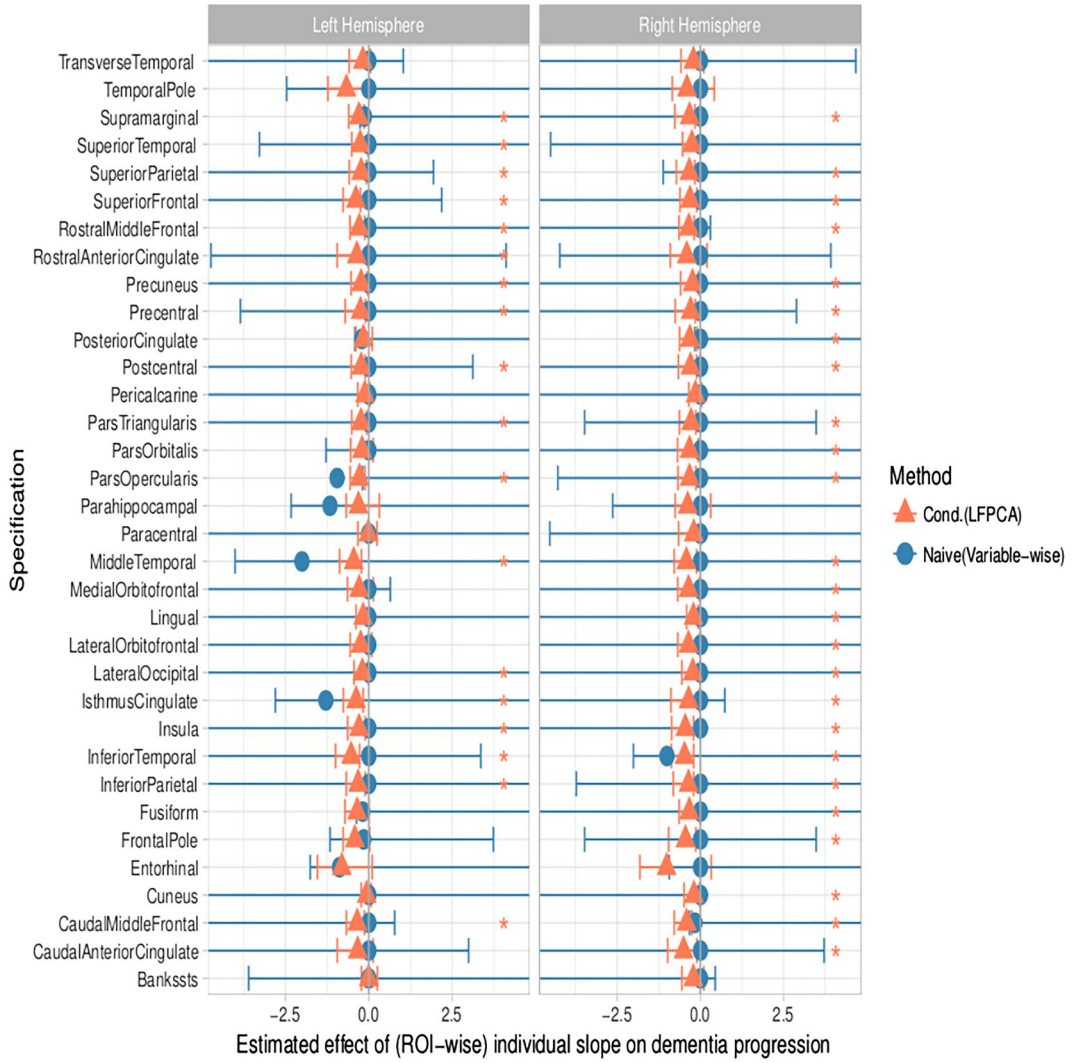

FIGURE 6 The coefficient estimates associated with the random slopes that correspond to the 34 regions of interest (ROIs) of the left hemisphere (in the left panel) and of the right hemisphere (in the right panel), comparing Cond.(LFPCA) and Naive(Variable-wise), with 95% confidence intervals obtained from 500 bootstrap replications. Estimated significant ROIs are marked with an asterisk on the right

Naive(Variable-wise), we estimate covariate model (1) ROI-wise, using an ROI-specific intercept and slope model without performing LFPC dimension reduction. We note that the estimated coefficients (and the associated confidence intervals) obtained from Naive(LFPCA) were relatively very similar to the estimates obtained from Cond.(LFPCA), in comparison with the difference in the estimates between Cond.(LFPCA) and Naive(Variable-wise). Hence, for clarity of presentation, we did not display the estimates from Naive(LFPCA) in Figures 5 and 6. The coefficient estimates from Cond.(LFPCA) are marked with the red triangles, and those obtained from Naive(Variable-wise) are marked with the blue circles.

Cond.(LFPCA) and Naive(LFPCA) (although not displayed in the figures) identified 46 common random effects as significant covariates, in which the associated 95% bootstrap confidence intervals did not contain zeros. Cond.(LFPCA) that accounts for the errors in covariates additionally identified the baseline value (the intercept) of right lateral orbitofrontal ($\hat{\beta}_1 = -0.30$), the change rates (the slopes) of superior temporal (−0.25), right fusiform (−0.33), and right lateral occipital (−0.23), as significant covariates, giving a total of 50 significant covariates. Naive(LFPCA) additionally identified the change rate of left temporal pole (−0.67) as a significant covariate, giving a total of 47 significant covariates. On the other hand, Naive(Variable-wise) exhibited a large variability in the coefficient estimates, due to the high correlations among the random effects. The associated 95% bootstrap confidence intervals were much larger than those of Cond.(LFPCA) and Naive(LFPCA) and did not identify any covariates as significant.

Most of the significant coefficients associated with the intercepts were negative, which indicates that thinner cortical thickness in the identified areas is associated with a higher chance of dementia transition. Also, negative estimates associated with the slopes imply that faster cortical thinning in the identified ROIs is associated with dementia transition.

5 | DISCUSSION

In this paper, we focused on a special case where the covariates are subject-specific random intercepts and slopes from a longitudinal mixed effects model. However, the method can be extended to a more general random effects model with minor modification. Also, by utilizing the LFPCA and a regularized estimation via CD, the proposed method could handle high-dimensional correlated random effects effectively.

Simulations showed that the proposed approach outperforms the naive approaches that ignore the errors associated with the random effect covariates. A robust majorization employed in Algorithm 1 was critical in the implementation of the method, because directly utilizing a CD/Newton-Raphson-type algorithm (particularly for a high dimension) is prone to instability in the sequence of iterations (13) and (16) due to the substantial non-linearity (with respect to $\boldsymbol{\beta}$) present in conditional-score equation (11), as discussed in Section 2.3. We note that multiple solutions of the parameters might exist (Carroll et al., 2006), and in such cases, the estimates can be sensitive to the choice of initialization. The naive estimate Naive(LFPCA) of the model is often a reasonable initial estimate. A further research direction for the proposed method includes incorporation of multiple functional covariates and prediction based on the estimated models.

The application to ADNI data revealed that intercepts in the right entorhinal and bilateral lateral orbitofrontal thickness are highly associated with dementia transition. Particularly, many previous studies have reported that those with entorhinal cortical thickness (e.g., ; Devanand et al., 2012; Lee et al., 2016) and bilateral lateral orbitofrontal cortical thickness are highly associated with having MCI than are cognitively normal people (Zhao et al., 2015). In this ADNI sample, the slopes of most ROIs had negative coefficient estimates. This reflects that neurodegeneration occurs in many of cortical areas during the follow-up, and MCI participants with more neurodegeneration transitioned to dementia. The results suggest that taking account for measurement errors and high correlations between random effect estimates improved estimation and yielded meaningful results that are also consistent with literature.

ACKNOWLEDGEMENT

This work was supported by National Institutes of Health Grant K01AG051348.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon request.

ORCID

Hyung Park https://orcid.org/0000-0002-8994-9583

Seonjoo Lee https://orcid.org/0000-0003-3177-6357

REFERENCES

Bickel, P. J., & Ritov, Y. (1987). Efficient estimation in the errors in variables model. *The Annals of Statistics*, *15*(2), 513–540.

Cai, X. (2015). Methods for handling measurement error and sources of variation in functional data models. *Columbia University Academic Commons*. https://doi.org/10.7916/D8M907CJ

Carroll, R. J., Knickerbocker, R. K., & Wang, C. Y. (1995). Dimension reduction in a semiparametric regression model with errors in covariates. *The Annals of Statistics*, *23*(1), 161–181.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. (2006). *Measurement error in nonlinear models: A modern perspective, second edition*: Chapman and Hall/CRC.

Datta, A., & Zou, H. (2017). CoCoLasso for high-dimensional error-in-variables regression. *The Annals of Statistics*, *45*(6), 2400–2426.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T, Dickerson, B. C, Blacker, D., … Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*, *31*(3), 968–980.

Devanand, D. P., Bansal, R., Liu, J., Hao, X., Pradhaban, G., & Peterson, B. S (2012). MRI hippocampal and entorhinal cortex mapping in predicting conversion to Alzheimer's disease. *Neuroimage*, *60*(3), 1622–1629.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360.

Yi, G. (2016). *Statistical analysis with measurement error or misclassification*. Springer.

Greven, S., Crainiceanu, C. M., Caffo, B. S., & Reich, D. (2010). Longitudinal functional principal component analysis. *Electronics Journal Statistical*, *4*, 1022–1054.

Huque, M. H., Bondell, H. D., Carroll, R. J., & Ryan, L. M. (2016). Spatial regression with covariate measurement error: A semiparametric approach. *Biometrics*, *72*, 678–686.

Iscan, Z., Jin, T. B., Kendrick, A., Szeglin, B., Lu, H., Trivedi, M., … DeLorenzo C. (2015). Test–retest reliability of freesurfer measurements within and between sites: Effects of visual approval process. *Human Brain Mapping*, *36*(9), 3472–3485.

Jiang, D., & Huang, J. (2014). Majorization minimization by coordinate descent for concave penalized generalized linear models. *Statistics and Computing*, *24*(5), 871–883.

Lee, S., Brickman, A. M., Andrews, H., Stern, Y., Schupf, N., Manly, J. J., … Devanand, D. P. (2016). Predictive utility of entorhinal cortex thinning and odor identification test for transition to dementia and cognitive decline in an urban community population. *Alzheimer's Dementia*, *12*(7), P316.

Li, L., Shao, J., & Palta, M. (2005). A longitudinal measurement error model with a semicontinuous covariate. *Biometrics*, *61*(3), 824–830.

Li, Y., Tang, H., & Lin, X. (2009). Spatial linear mixed models with covariate measurement errors. *Statistica Sinica*, *19*, 1077–1093.

Li, E., Zhang, D., & Davidian, M. (2004). Conditional estimation for generalized linear models when covariates are subject-specific parameters in a mixed model for longitudinal measurements. *Biometrics*, *60*, 1–7.

Liang, H., & Li, R. (2009). Variable selection for partially linear models with measurement errors. *Journal of the American Statistical Association*, *104*(485).

Ma, Y., & Li, R. (2010). Variable selection in measurement error models. *Bernoulli (Andover)*, *16*(1), 274–300.

Midthune, D., Carroll, R. J., Freedman, L. S., & Kipnis, V. (2016). Measurement error models with interactions. *Biostatistics*, *17*(2), 277–290.

R Core Team (2019). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org

Stefanski, L. A., & Carroll, R. J. (1985). Covariate measurement error in logistic regression. *The Annals of Statistics*, *13*(4), 1335–1351.

Stefanski, L. A., & Carroll, R. J. (1987). Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika*, *74*(4), 703–716.

Stefanski, L. A., Wu, Y., & White, K. (2014). Variable selection in nonparametric classification via measurement error model selection likelihoods. *Journal of the American StatisticalAssociation*, *109*(506), 574–589.

Zhang, X., Wang, H., Ma, Y., & Carroll, R. J. (2017). Linear model selection when covariates contain errors. *Journal of the American StatisticalAssociation*, *112*(520), 1553–1561.

Zhao, Hui, Li, Xiaoxi, Wu, Wenbo, Li, Zheng, Qian, Lai, Li, ShanShan, … Xu, Yun (2015). Atrophic patterns of the frontal-subcortical circuits in patients with mild cognitive impairment and Alzheimer's disease. *PloS one*, *10*(6), e0130017.

Zipunnikov, V., Greven, S., Shou, Hl., Caffo, B., Reich, D., & Crainiceanu, C. (2014). Longitudinal high-dimensional principal components analysis with application to diffusion tensor imaging of multiple sclerosis. *The Annals of Applied Statistics*, *8*(4), 2175–2202.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*, 301–320.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Park H, Lee S. Logistic regression error-in-covariate models for longitudinal high-dimensional covariates. *Stat*. 2019:8;e246. https://doi.org/10.1002/sta4.246