Computational Biology and Chemistry 99 (2022) 107697

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/cbac

Bias correction for multiple covariate analysis using empirical bayesian estimation in mixed-effects models for longitudinal data

Yi Li [a], Yaning Yang [a], Xu Steven Xu [b], Min Yuan [c,*], for the Alzheimer's Disease Neuroimaging Initiative

[a] Department of Statistics and Finance, University of Science and Technology of China, China
[b] Genmab US, Inc, United States
[c] Anhui Medical University, China

ARTICLE INFO

Keywords:
Empirical Bayes estimates
Shrinkage factor
Multiple mixed-effect model
Longitudinal data
Maximum likelihood estimates

ABSTRACT

The naïve empirical Bayes method has been widely used as an ad hoc tool in fitting linear mixed-effect models, which is much computationally efficient than the maximum likelihood estimation method. However, the shrinkage effect of the empirical Bayes method causes bias in the estimates of the fixed effects. Bias-correction has been proposed for the mixed-effects model when only one covariate is present. In this paper, we derive the shrinkage factor of the empirical Bayes predictors of the random effects and the variance-covariance matrix of the corrected estimates when the model has more than one covariate. The empirical Bayes estimates and test statistics are then corrected using the derived factor. Theoretical derivations, simulation studies and a real data application demonstrate the validity of the proposed method in that the corrected estimates are unbiased and the corrected tests have correct p-values.

1. Introduction

Mixed-effects modeling provide a useful tool for the analysis of longitudinal data. Likelihood-based algorithms, including maximum likelihood (ML) or restricted maximum likelihood (REML), are often used for parameter estimation and hypothesis testing in mixed-effects model (Searle et al., 1992), including covariate analysis. To circumvent complex algorithms involved in ML or REML, Empirical Bayes Estimates (EBEs), derived from the base mixed-effects model without covariates has long been used as an ad hoc approach to facilitate variable selection (Davidian and Giltinan, 2003; Combes et al., 2014). For example, EBEs are extremely useful in population pharmacokinetic studies for investigating the influence of the individual's baseline characteristics on the individual parameters. EBEs based inferences help researchers to gain insight into within-subject pharmacokinetic processes of absorption, distribution, and elimination of drug concentration in human body (Pinheiro and Bates, 2000). Other examples can be found in HIV viral dynamic studies, tumor growth inhibition modeling and genome-wide association studies with longitudinal phenotypes (Maitre et al., 1991; Mandema et al., 1992; Wu and Ding, 1999; Lindbom et al., 2004; Savic and Karlsson, 2009; Londono et al., 2013; Meirelles et al., 2013; Combes et al., 2014; Sikorska et al., 2015; Barbolosi et al., 2016). Approaches based EBEs generally decompose the full mixed-effects model into a null mixed-effects model without covariates and a simple linear model, thus the fitting complexity of a full mixed-effects model with many covariates is substantially reduced. Despite its simplicity, it is well known that the EBEs are biased as they tend to be shrunk to the corresponding population mean estimate, and may not be suitable for identification of significant variables (Davidian and Giltinan, 2003; Savic and Karlsson, 2009).

In order to utilize the simplicity and effectiveness of empirical Bayes estimates, algorithms have been developed to correct the bias in estimation and variance of EBEs for univariate analysis in the situation that only one covariate is considered in the model (Yuan et al., 2019, 2020, 2021). Yuan proposed a quick and efficient bias correction method for modeling longitudinal data with the mixed-effects model (Yuan et al., 2019). They considered both linear and non-linear mixed effects model with one covariate having effect on one random-effects parameter, derived the expression of shrinkage factor and used it for bias correction. Yuan extended the approach to the situation when one covariate has effects on several random parameters (Yuan et al., 2020, 2021). These algorithms, particularly simultaneous correction methods, not only

* Correspondence to: School of Public Health Administration, Anhui Medical University, Hefei 230032 Anhui, China.
E-mail address: myuan@ustc.edu.cn (M. Yuan).

https://doi.org/10.1016/j.compbiolchem.2022.107697
Received 24 February 2022; Received in revised form 4 May 2022; Accepted 11 May 2022
Available online 23 May 2022
1476-9271/© 2022 Elsevier Ltd. All rights reserved.

correct the bias caused by shrinkage, and provides numerically identical estimation and p-values to those from the standard mixed-effects model, but also drastically improve the computational efficiency for high-dimension variable selection with linear and nonlinear longitudinal outcomes.

However, when multiple correlated covariates have joint effects on random-effect parameters, which is more common in practice, the correction algorithms for single variable analysis may not work properly. Therefore, it is necessary to develop bias correction methods for EBE-based approaches with more than one covariate. Here we extend the single-variable correction approach to a more general situation in which multiple covariates have effects on multiple random-effect parameters; the extended approach is named mSCEBE. We derived the correction matrix theoretically and correct biases for both estimation and variance. Extensive simulations and real application are implemented to illustrate the proposed method.

The rest of this paper is organized as follows. Section 2 introduces mixed-effects models and the theoretical results. Extensive simulation studies are performed to examine the performance of the proposed methods in Section 3. In Section 4, we apply the bias correction methods to a real data from a subsample of the Alzheimer's Disease Neuroimaging Initiative (ADNI), who had developed Alzheimer's disease from mild cognitive impairment (MCI) at the base line. Conclusions and discussions are given in the last section.

2. Methods

Suppose there are m individuals and the ith individual has n_i observations $y_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})'$ at time points $t_i = (t_{i1}, t_{i2}, \ldots, t_{in_i})'$. There are q candidate covariates considered to be associated with observations y_i. A typical linear mixed model with multiple covariates can be described as follows,

$$\begin{aligned} &y_i = Z_i\beta_i + \epsilon_i \\ &\beta_i = \alpha + \gamma x_i + \eta_i \; i = 1, 2, \ldots, m \\ &\epsilon_i \sim N(0, G_i) \; and \; \eta_i \sim N(0, R) \end{aligned} \tag{1}$$

where β_i is the $p \times 1$ random effect vector, Z_i is a $n_i \times p$ design matrix, $x_i = (x_{i1}, x_{i2}, \ldots, x_{iq})'$ is the observed values of the q covariates for the ith individual, α is the $p \times 1$ intercept parameter, and the slope parameter γ is a $p \times q$ matrix. The base model is constructed by omitting all covariates in the linear model for random effects,

$$\begin{aligned} &y_i = Z_i\beta_i^* + \epsilon_i^* \\ &\beta_i^* = \alpha^* + \eta_i^* \; i = 1, 2, \ldots, m \\ &\epsilon_i^* \sim N(0, G_i^*) \; and \; \eta_i^* \sim N(0, R^*) \end{aligned} \tag{2}$$

The best linear unbiased predictor (BLUPs) for β_i^*, defined as the posterior mean of β_i^* given data y_i and nuisance parameters α^*, G_i^*, R^*, equals to

$$BP(\beta_i^*) = (Z_i' G_i^{*-1} Z_i + R^{*-1})^{-1} (Z_i' G_i^{*-1} y_i + R^{*-1}\alpha^*)$$

The parametrical empirical Bayesian estimators (naive EBE) of β_i^*, denoted as $\widehat{\beta}_i^*$, is then obtained by plugging the MLEs of nuisance parameters such as α^*, G_i^*, R^*. In the second step, we regress the BLUPs on covariates via the following multiple and multivariate regression model,

$$\begin{pmatrix} \widehat{\beta}_1^{*T} \\ \widehat{\beta}_2^{*T} \\ \vdots \\ \widehat{\beta}_m^{*T} \end{pmatrix} = \begin{pmatrix} 1 & x_1^T \\ 1 & x_2^T \\ \vdots & \vdots \\ 1 & x_m^T \end{pmatrix} \begin{pmatrix} \alpha^{**T} \\ \gamma^{*T} \end{pmatrix} + \eta^{**} \tag{3}$$

Denote $X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_m^T \end{pmatrix}$, $\widehat{B}^* = \begin{pmatrix} \widehat{\beta}_1^{*T} \\ \widehat{\beta}_2^{*T} \\ \vdots \\ \widehat{\beta}_m^{*T} \end{pmatrix}$, and $X_c = \begin{pmatrix} (x_1 - \overline{x})^T \\ (x_2 - \overline{x})^T \\ \vdots \\ (x_m - \overline{x})^T \end{pmatrix}$ with $\overline{x} = \frac{1}{m}\sum_{i=1}^m x_i$. The least square estimator for parameter matrix γ^* is $\widehat{\gamma}^* = \widehat{B}^{*T} X_c (X_c^T X_c)^{-1}$ which is called NEBE in this article. We made vectorization of matrix $\widehat{\gamma}^*$ and calculated its expectation as $E[vec(\widehat{\gamma}^*)] = S_c vec(\gamma)$. The expression of S_c is given as follows and the details to derive this expression can be found in Appendix A1,

$$\begin{aligned} S_c = (X_c^T X_c)^{-1} \bigotimes I_p \Big(\sum_{i=1}^m \big((x_i - \overline{x}) x_i^T\big) \bigotimes (I_p - S_i^*) + \sum_{i=1}^m (x_i - \overline{x}) \bigotimes S_i^* \\ \times \sum_{j=1}^m (x_j^T \bigotimes W_j^*) \Big) \end{aligned}$$

where $S_i^* = \left(Z_i' G_i^{*-1} Z_i + R^{*-1}\right)^{-1} R^{*-1}, W_j^* = \left(\sum_{j=1}^m Z_j' \Sigma_j^{*-1} Z_j\right)^{-1} Z_j' \Sigma_j^{*-1} Z_j$, and $\bigotimes$ denotes the Kronecker product.

Then S_c^{-1} can be served as the simultaneous correction matrix and the simultaneously corrected estimator of γ, which is called mSCEBE, can be expressed as $vec(\widehat{\gamma}_{sim}) = S_c^{-1} vec(\widehat{\gamma}^*)$. $vec(\cdot)$ is the vectorization operation for a certain matrix and $\bigotimes$ stands for the kronecker product. In order to construct the testing statistics, we need to derive the covariance matrix of $\widehat{\gamma}_{sim}$. The covariance matrix of $vec(\widehat{\gamma}^*)$ under model (1) can be derived by calculating covariance matrix of y_i and formal delta method. Denote

$$A_i^* = \left(Z_i' G_i^{*-1} Z_i + R^{*-1}\right)^{-1} Z_i' G_i^{*-1}$$

$$B_i^* = \left(Z_i' G_i^{*-1} Z_i + R^{*-1}\right)^{-1} R^{*-1}$$

$$C_i^* = \left(\sum_{i=1}^m Z_i' \Sigma_i^{*-1} Z_i\right)^{-1} Z_i' \Sigma_i^{*-1}$$

$$D^* = blockdiag(A_i, i = 1, 2, \ldots, m) + (B_i^* C_j^*, i, j = 1, 2, \ldots, m)$$

'blockdiag' means block diagonal matrix with each block element A_i. The second term in D^* is a block matrix with the ijth block element $B_i^* C_j^*$. Then the covariance matrix of $vec(\widehat{\gamma}^*)$ can be calculated as

$$\begin{aligned} &\left\{ \left[(X_c^T X_c)^{-1} X_c^T \right] \bigotimes I_p \right\} D^* blockdiag(\Sigma_i, i \\ &= 1, 2, \ldots, m) D^{*T} \left\{ \left[(X_c^T X_c)^{-1} X_c^T \right] \bigotimes I_p \right\}^T \end{aligned}$$

The t-test for $H_{0ij} : \gamma_{ij} = \gamma_{ij0}$ can be constructed as follows

$$t_{ij} = \frac{\widehat{\gamma}_{ij} - \gamma_{ij0}}{\sqrt{\left[S_c^{-1} Var(vec(\widehat{\gamma}^*)) \left(S_c^{-1}\right)' \right]_{u,u}}}$$

$$u = i + p(j-1)$$

where $\widehat{\gamma}_{ij}$ is the (i,j) component of $\widehat{\gamma}_{sim}$, and the subscript (u,u) denotes the uth diagonal of the matrix, which is the variance estimation of $\widehat{\gamma}_{ij}$. Details to derive the covariance matrix of $vec(\widehat{\gamma}^*)$ are provided in Appendix A2.

It should be noted that the covariance matrix of random effects estimated based on the base model are inflated due to removed covariate effects. Thus it should be considered with caution when taking $\widehat{R}^*$ as an estimation of R in the calculation of $Var(\widehat{\gamma}_{sim})$. Since η_i^* in the base model is related to the covariates and error term in the full model by $\eta_i^* = \gamma x_i + \varepsilon_i$, which implies that $R^* = var(\eta_i^*) = var(\gamma x_i) + var(\varepsilon_i) = \gamma' cov(x_i)\gamma + R$. Therefore, an alternative estimate of R can be chosen to be $(\widehat{R}^* -$

Table 1
Number of measurements and sampling time points.

Number of measurements	Sampling time points
3	0.05,0.3,1
5	0.05,0.15,0.3,0.6,1
7	0.01,0.05,0.1,0.2,0.4,0.6,1
9	0.01,0.05,0.1,0.2,0.3,0.4,0.6,0.8,1

$\widehat{\gamma}_{sim} Var(x) \widehat{\gamma}_{sim}^{T}$), the derived method with this variance correction is named rmSCEBE. We will compare the performance of those two estimates of covariance matrix in the following contexts.

3. Simulation

3.1. Parameter settings

To investigate the accuracy of corrected estimators and tests by the multiple correction factor, we perform simulations with four approaches (LMER/NEBE/mSCEBE/rmSCEBE) based on a linear mixed-effects model with ten covariates, where LMER refers to the full model-based likelihood method, which is used in this paper as the gold standard for comparison.

$$y_{ij} = \beta_{i1} + \beta_{i2} t_{ij} + e_{ij},\ j = 1, 2, \ldots, n_i$$

$$\beta_{i1} = \alpha_1 + \gamma_{11} x_{i1} + \ldots + \gamma_{110} x_{i10} + b_{i1}$$

$$\beta_{i2} = \alpha_2 + \gamma_{21} x_{i1} + \ldots + \gamma_{210} x_{i10} + b_{i2},\ i = 1, 2, \ldots, m$$

where the random intercept effects b_{i1} and slope effects b_{i2} are independently generated from a normal distribution $N(0,\ \sigma_b^2)$, and the n_i-dimensional within-subject error vector $e_i = (e_{i1}, e_{i2}, \ldots, e_{in_i})^T$ are independently simulated from a multivariate normal distribution with mean 0 and a diagonal covariance matrix $\sigma^2 I$. In simulations, α_1 and α_2 are set to be zero, and the coefficient matrix$\gamma = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{110} \\ \gamma_{21} & \cdots & \gamma_{210} \end{pmatrix}$ is set to be one of the following four situations:

Situation 1. $: \gamma_{1k} = 0.2, \gamma_{2k} = 0, k = 1, \ldots, 5; \gamma_{1k} = 0, \gamma_{2k} = 0.2, k = 6, \ldots, 10;$.

Situation 2. $: \gamma_{1k} = 0.2, \gamma_{2k} = 0, k = 1, \ldots, 10;$.

Situation 3. $: \gamma_{1k} = 0, \gamma_{2k} = 0.2, k = 1, \ldots, 10;$.

Situation 4. $: \gamma_{1k} = \gamma_{2k} = 0.2, k = 1, \ldots, 10.$.

We chose these scenarios for investigating both cross-sectional covariate effects and the interaction effects. In situation 1, there are only cross-sectional effects on random intercept parameter and only covariate-time interaction effects on the random slope parameter. In situation 2, there are both cross-sectional and interaction effects on random intercept, but no effects on the random slope. The third situation is the opposite of the second situation. It has effects on the random slope while no effect on the random intercept. The fourth situation simulates that cross-section and interaction effects are imposed on both random intercept and slope. These four situations of coefficient include all situations that may be encountered in practice. The between-subject error $\sigma_b = 0.3, 0.4, 0.5$ and the within-subject error $\sigma = 0.3, 0.6$. Simulations

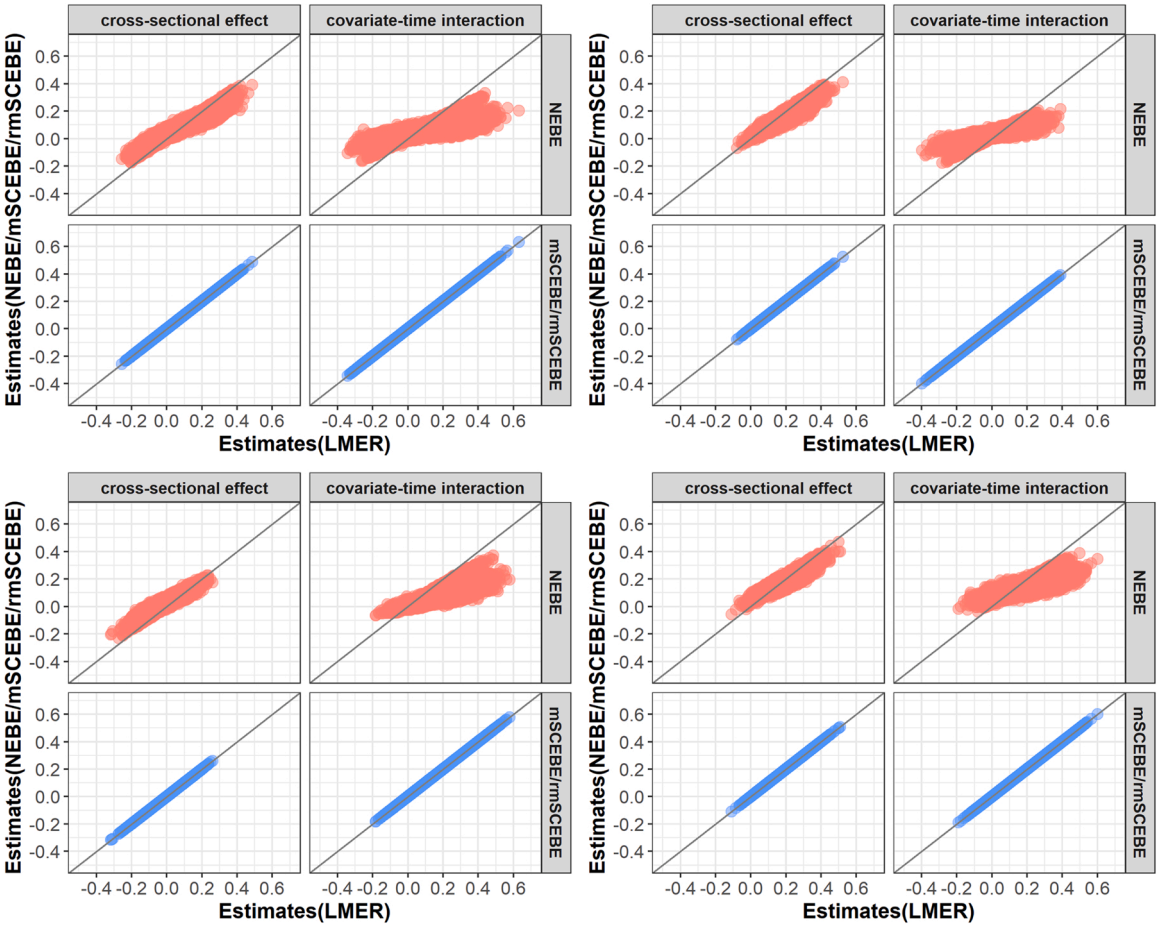

Fig. 1. Estimation comparisons for both cross-sectional effects and covariate-time interaction effects based on NEBE, mSCEBE, rmSCEBE and LMER methods. NEBE: EBEs-based method without bias correction; mSCEBE: EBEs-based method with bias correction; rmSCEBE: EBEs-based method with revised bias correction for covariance matrix; LMER: standard maximum estimates and likelihood ratio test produced by *lmer* package in R. Four figures correspond to situations 1–4 in the order from left to right and top to bottom.

Table 2
Familywise error rates for LMER, NEBE, mSCEBE and rmSCEBE under various scenarios.

N	nt	σ	σ_b	FWER			
				LMER	NEBE	mSCEBE	rmSCEBE
500	3	0.3	0.3	0.054	0.046	0.050	0.059
		0.3	0.4	0.046	0.043	0.041	0.051
		0.3	0.5	0.063	0.063	0.058	0.067
		0.6	0.3	0.052	0.045	0.048	0.056
		0.6	0.4	0.046	0.046	0.044	0.053
		0.6	0.5	0.053	0.051	0.051	0.061
500	7	0.3	0.3	0.053	0.052	0.050	0.061
		0.3	0.4	0.038	0.036	0.034	0.040
		0.3	0.5	0.054	0.052	0.050	0.061
		0.6	0.3	0.058	0.044	0.057	0.060
		0.6	0.4	0.056	0.060	0.051	0.062
		0.6	0.5	0.058	0.066	0.052	0.064
800	3	0.3	0.3	0.048	0.057	0.044	0.053
		0.3	0.4	0.057	0.051	0.053	0.058
		0.3	0.5	0.055	0.051	0.052	0.055
		0.6	0.3	0.054	0.047	0.051	0.058
		0.6	0.4	0.058	0.044	0.055	0.066
		0.6	0.5	0.037	0.049	0.035	0.041
800	7	0.3	0.3	0.055	0.053	0.055	0.058
		0.3	0.4	0.057	0.054	0.053	0.061
		0.3	0.5	0.070	0.059	0.068	0.072
		0.6	0.3	0.046	0.043	0.044	0.048
		0.6	0.4	0.059	0.053	0.056	0.065
		0.6	0.5	0.059	0.050	0.055	0.061

are replicated 1000 times for each scenario. Both balanced design and unbalanced design are investigated. The sample size $m = 500, 800$, and the number of measurements is fixed at 3 or 7 for the balanced design, and are randomly sampled from 3, 5, 7, 9 with replacement for the unbalanced design. The corresponding measuring time points are summarized in Table 1. Covariates $x_i = (x_{i1}, \ldots, x_{i10})^T$ are independently generated from a multivariate normal distribution with mean zero and a covariance matrix with 0.09 for all the diagonals and 0.009 for the other elements. There are 24 scenarios in total for both balanced and unbalanced design.

3.2. Estimation

Fig. 1 compares model estimates for both cross-sectional and covariate-time interaction effects based on NEBE, the derived methods (mSCEBE/rmSCEBE) and LMER (maximum likelihood estimates obtained by a standard approach implemented in *lmer* package in R software) with various simulation scenarios. As expected, compared to LMER, NEBE is biased for either cross-sectional effects or the covariate-time interaction (Fig. 1). However, after corrections, the estimates from mSCEBE are virtually identical to those based on the LMER approach as the data points perfectly aligned on the 1:1 identity line. rmSCEBE has the same correction factor as mSCEBE for effect estimation, thus provides the same results.

3.3. Association tests

For the association test, all of the four investigated approaches can well controlled the familywise error rates at the nominal level 0.05

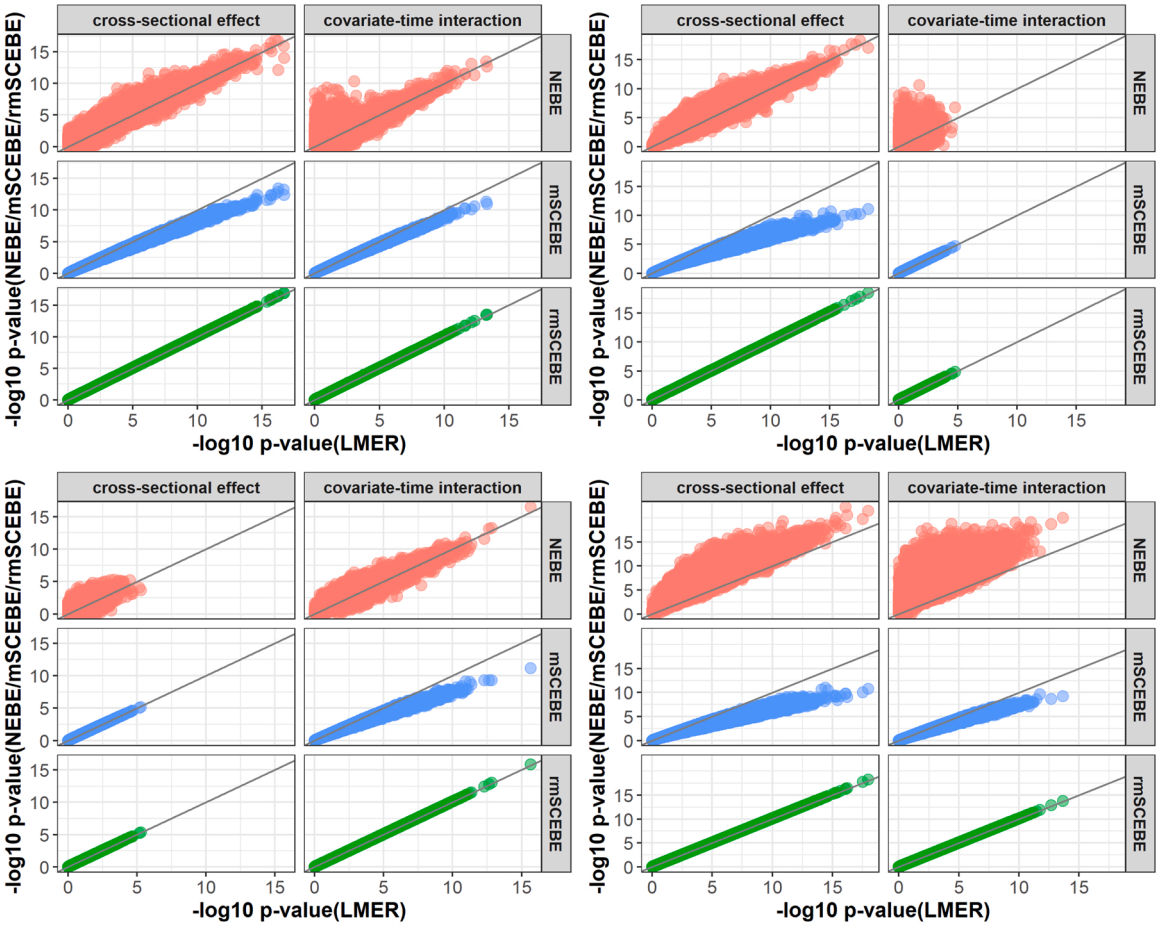

Fig. 2. P-values (on the –log10 scale) comparisons for both cross-sectional effects and covariate-time interaction effects based on NEBE, mSCEBE, rmSCEBE and LMER method in simulation situations 1–4. NEBE: EBEs-based method without bias correction; mSCEBE: EBEs-based method with bias correction; rmSCEBE: EBEs-based method with revised bias correction for covariance matrix; LMER: standard maximum estimates and likelihood ratio test produced by *lmer* package in R. Four figures correspond to situations 1–4 in the order from left to right and top to bottom.

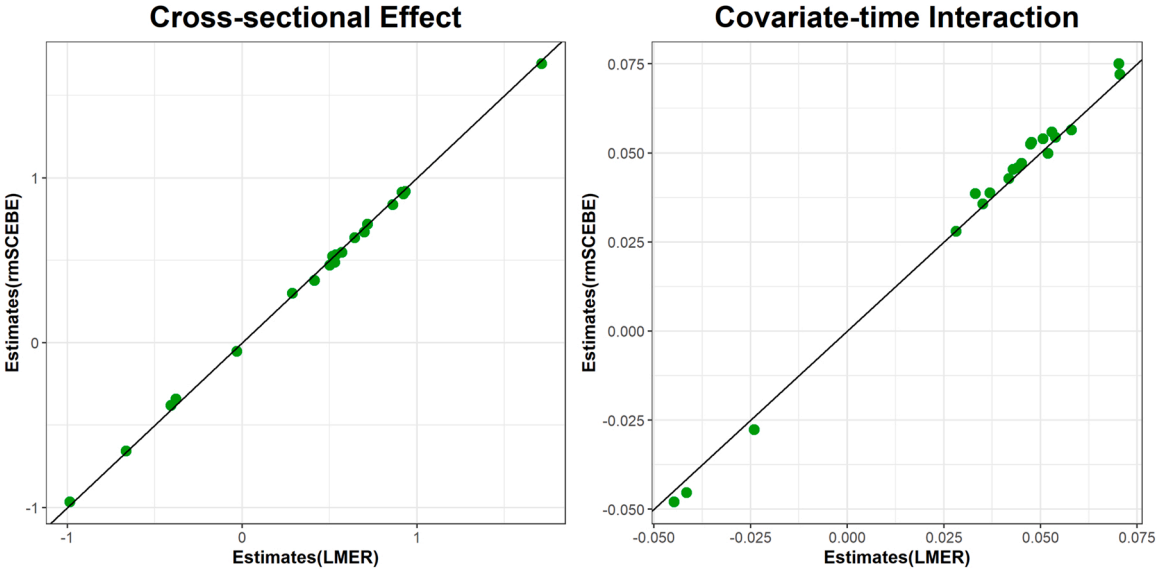

Fig. 3. Comparisons between estimates from LMER and rmSCEBE with the top 20 most significant SNPs selected by the EBE-APML0 method (Xu et al., 2020).

(Table 2). Although the p-values calculated based on NEBE appear to be trending the same way as those based on the LMER approach, the discrepancy in the p-values from these two approaches was obvious as the data points scatter around the 1:1 identity line (Fig. 2). On the contrast, mSCEBE provided more accurate p-values for the association test than the NEBE on both intercept and the slope of the model compared to the LMER approach regardless of the level of shrinkage. However, mSCEBE still tends to be biased when p-value is very small. The rmSCEBE can completely correct the bias in p-value and provide almost identical p-value to the standard LMER approach.

4. Application

We applied our method to a cohort of 785 individuals who had developed Alzheimer's disease(AD) from mild cognitive impairment (MCI) at the baseline, which is a section of the Alzheimer's Disease Neuroimaging Initiative (ADNI; ADNI-1, ADNI-2, and ADNI-GO) cohort study. In the current analysis, we worked on the top 20 most significant SNPs selected by the EBE-APML0 method (Xu et al., 2020). Both cross-sectional covariate effects and interactions were examined by the rmSCEBE method and the lmer function in R. Comparisons between estimates and p-values are shown in Figs. 3 and 4. The results show that rmSCEBE can correct the deviation of naïve EBE and obtain unbiased estimates and p-values compared to LMER from the standard likelihood approach.

We presented the results of rmSCEBE ordered by the ascending p-values of each effect in Table 3. SNP rs429358 has a relatively large effect (1.693) and p-value smaller than a suggestive significance level ($p < 0.05/20$) at both cross-sectional effect and interaction. This SNP is located in the fourth exon of the APOE gene, and its combination with rs7412 determines APOE isoforms ($\epsilon 2$, $\epsilon 3$, $\epsilon 4$). The allele APOE $\epsilon 4$ has been shown to be involved in the pathogenesis of both late-onset familial and sporadic AD (Saunders et al., 1993). Our results suggest that rs429358 is associated with both onset and development (time course) of AD.

5. Discussion

Longitudinal data arise in a wide variety of areas, including agriculture, biology, public health and biomedicine. Examples include wheat yields (Stroup et al., 1994), body weight growth (Hand and Crowder, 1996), tumor growth inhibition modeling (Barbolosi et al.,

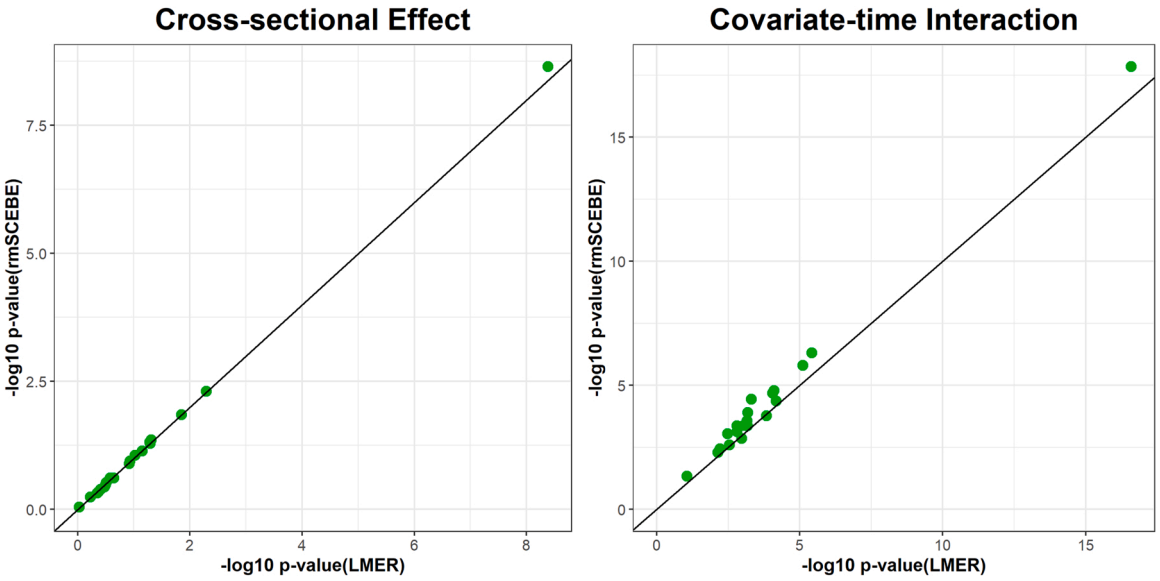

Fig. 4. Comparisons between –log 10 (p-values) from LMER and rmSCEBE with the top 20 most significant SNPs selected by the EBE-APML0 method (Xu et al., 2020).

Table 3
Top 20 significant SNPs and their corresponding genes for baseline disease status and disease progression.

Snp.name	Gene	Relationship	MAF[a]	Disease Status[a]		Disease Progression[a]	
				estimation	p-value	estimation	p-value
rs429358	APOE	within	0.2369	1.6933	0.0000	0.0720	0.0000
rs17836364	LILRA4	within	0.1726	0.9176	0.0050	0.0386	0.0000
rs157357	ZNF274	within	0.1102	-0.9660	0.0142	-0.0480	0.0000
rs62111293	NOVA2	within	0.1408	0.7191	0.0438	0.0357	0.0004
rs12973761	ZIM3	within	0.1580	-0.6569	0.0483	0.0280	0.0025
rs143988316	PBX4	nearby	0.0764	0.9018	0.0518	0.0525	0.0001
rs12979207	ZNF431	within	0.0720	0.8385	0.0728	0.0471	0.0004
rs62131315	ZNF805	nearby	0.0580	0.9137	0.0884	0.0428	0.0051
19_14572615	NDUFB7	nearby	0.1121	0.6367	0.1135	0.0559	0.0000
rs62109563	ERCC1	within	0.0796	0.6703	0.1288	0.0540	0.0000
rs73488486	ZNF358 & LOC105372261	within	0.1006	0.4712	0.2452	0.0387	0.0007
rs3816034	LINC01837 & LINC01533	within	0.0732	0.5254	0.2459	0.0544	0.0000
rs55869726	HCN2	nearby	0.0637	0.5330	0.3039	0.0564	0.0002
rs16969505	SCGB1B2P	within	0.0503	0.5474	0.3323	0.0457	0.0037
19_10096680	ANGPTL6	within	0.0580	0.4870	0.3647	0.0530	0.0009
rs147388909	LOC107987267	within	0.0752	-0.3808	0.4124	-0.0453	0.0004
rs62111468	ZNF493	nearby	0.0567	0.3781	0.4622	0.0751	0.0000
rs11878192	RPL7AP69	nearby	0.0688	-0.3418	0.4852	-0.0276	0.0464
rs35194062	RELB	within	0.0573	0.2989	0.5739	0.0499	0.0014
rs111677971	ERFL	within	0.0898	-0.0524	0.9010	0.0453	0.0003

[a] MAF: minor allele frequency; Disease Status: the effect of SNP on disease status (intercept in mixed-effects model); Disease Progression: the effect of SNP on disease progression (slope in mixed-effects model).

2016), viral dynamic modeling (Wu and Ding, 1999) and gene expression dynamic modeling (Marguet et al. 2019) etc. The mixed-effects models can distinguish between the within-group variabilities and between-group variabilities, thus can easy handle the unbalanced and missing data.

In this study, we investigated the EBE-based approaches of linear mixed-effects model with multiple correlated covariates, and presented a simultaneous correction method rmSCEBE. The classical likelihood theory based method (LMER) is computationally inefficient when multiple covariates are considered. In contrast, our proposed method is much more efficient, it only needs to fit a basic mixed-effects model without covariates and perform a simple regression of the predictions of random effects on the covariates. However, estimates obtained from EBEs are biased. We considered to estimate the bias caused by shrinkage effect of the empirical Bayes method and the overestimated variances on the base model, and use these quantities to correct for the biases of EBE. It was shown that our method could correct for biases of e EBEs effectively.

CRediT authorship contribution statement

Yi Li: Simulation, Data processing, Methodology. **Yaning Yang:** Conceptualization, Methodology, Writing – review & editing. **Xu Steven Xu:** Validation, Writing – review & editing. **MinYuan:** Conceptualization, Funding acquisition, Investigation, Methodology, Supervision, Writing – original draft.

Acknowledgements

The Authors declare that there is no conflict of interest. Yang is supported by National Science Foundation of China (NSFC), Grant No. 11671375. Yuan is supported by the Natural Science Foundation of Anhui Province (No. 2008085MA09) and the Doctoral Research Funding of Anhui Medical University (No. XJ201710).

Appendix

A1:Derivation of simultaneous correction matrix S_c

Denote the covariance matrix of y_i under model (2) to be $\Sigma_i^* = Z_i R^* Z_i' + G_i^*$, then the MLE of α^* under the base model (2) is $\left(\sum_{i=1}^m Z_i' \Sigma_i^{*-1} Z_i\right)^{-1} \sum_{i=1}^m Z_i' \Sigma_i^{*-1} y_i$ which can be regarded as the weighted average of y_i. The expectation of $\widehat{\beta}_i^*$ under the true model (1) is

$$E\widehat{\beta}_i^* = \alpha + (I - S_i^*)\gamma x_i + S_i^* \left(\sum_{i=1}^m Z_i' \Sigma_i^{*-1} Z_i \right)^{-1} \sum_{i=1}^m (Z_i' \Sigma_i^{*-1} Z_i \gamma x_i)$$

where $S_i^* = \left(Z_i' G_i^{*-1} Z_i + R^{*-1}\right)^{-1} R^{*-1}$. By plugging the expression of $E\widehat{\beta}_i^*$ and applying the vectorization of matrix, we obtain the expectation of the vectorization of $\widehat{\gamma}^*$ as follows,

$$vec(E\widehat{\gamma}^*) = \left(X_c^T X_c\right)^{-1} \bigotimes I_p \left(\sum_{i=1}^m \left((x_i - \bar{x}) x_i^T \right) \bigotimes (I_p - S_i^*) + \sum_{i=1}^m (x_i - \bar{x}) \bigotimes S_i^* \sum_{j=1}^m \left(x_j^T \bigotimes W_j^* \right) \right) vec(\gamma)$$

where $W_j^* = \left(\sum_{j=1}^m Z_j' \Sigma_j^{*-1} Z_j\right)^{-1} Z_j' \Sigma_j^{*-1} Z_j$ and $\bigotimes$ denotes the Kronecker product. Denote.

$S_c = \left(X_c^T X_c\right)^{-1} \bigotimes I_p \left(\sum_{i=1}^m \left((x_i - \bar{x}) x_i^T \right) \bigotimes (I_p - S_i^*) + \sum_{i=1}^m (x_i - \bar{x}) \bigotimes S_i^* \sum_{j=1}^m (x_j^T \bigotimes W_j^*) \right)$, then S_c can be served as the simultaneous correction matrix.

A2: Derivation of the covariance matrix of vec($\widehat{\gamma}^*$)

First, we calculate the covariance matrix of the vectorization of $\widehat{\gamma}^*$. The vectorization of $\widehat{\gamma}^*$ has the explicit form as follows

$$vec(\widehat{\gamma}^*) = \left[\left(X_c^T X_c\right)^{-1} X_c^T\right] \otimes I_p vec\left(\widehat{B}^{*T}\right) = \left[\left(X_c^T X_c\right)^{-1} X_c^T\right] \otimes I_p D^* (y_1, y_2, \ldots, y_m)^T$$

By applying the delta method and noticing that the covariance matrix of y_i is $\Sigma_i = Z_i R Z_i' + G_i$ under the true model. Therefore, the covariance matrix of $vec(\widehat{\gamma}^*)$ under model (1) can be expressed as follows.

$$Var(vec(\widehat{\gamma}^*)) = \left\{\left[\left(X_c^T X_c\right)^{-1} X_c^T\right] \otimes I_p\right\} D^* \begin{pmatrix} \Sigma_1 & & & \\ & \Sigma_2 & & \\ & & \ddots & \\ & & & \Sigma_m \end{pmatrix} D^{*T} \left\{\left[\left(X_c^T X_c\right)^{-1} X_c^T\right] \otimes I_p\right\}^T$$

where D^* is defined as:

$$D^* = \begin{pmatrix} A_1^* & & & \\ & A_2^* & & \\ & & \ddots & \\ & & & A_m^* \end{pmatrix} + \begin{pmatrix} B_1^* \\ B_2^* \\ \vdots \\ B_m^* \end{pmatrix} \begin{pmatrix} C_1^* & C_2^* & \ldots & C_m^* \end{pmatrix}$$

with $A_i^* = \left(Z_i' G_i^{*-1} Z_i + R^{*-1}\right)^{-1} Z_i' G_i^{*-1}$, $B_i^* = \left(Z_i' G_i^{*-1} Z_i + R^{*-1}\right)^{-1} R^{*-1}$ and $C_i^* = \left(\sum_{i=1}^m Z_i' \Sigma_i^{*-1} Z_i\right)^{-1} Z_i' \Sigma_i^{*-1}$.

A3: Results of more simulation scenarios

See Figs. A1–A5.

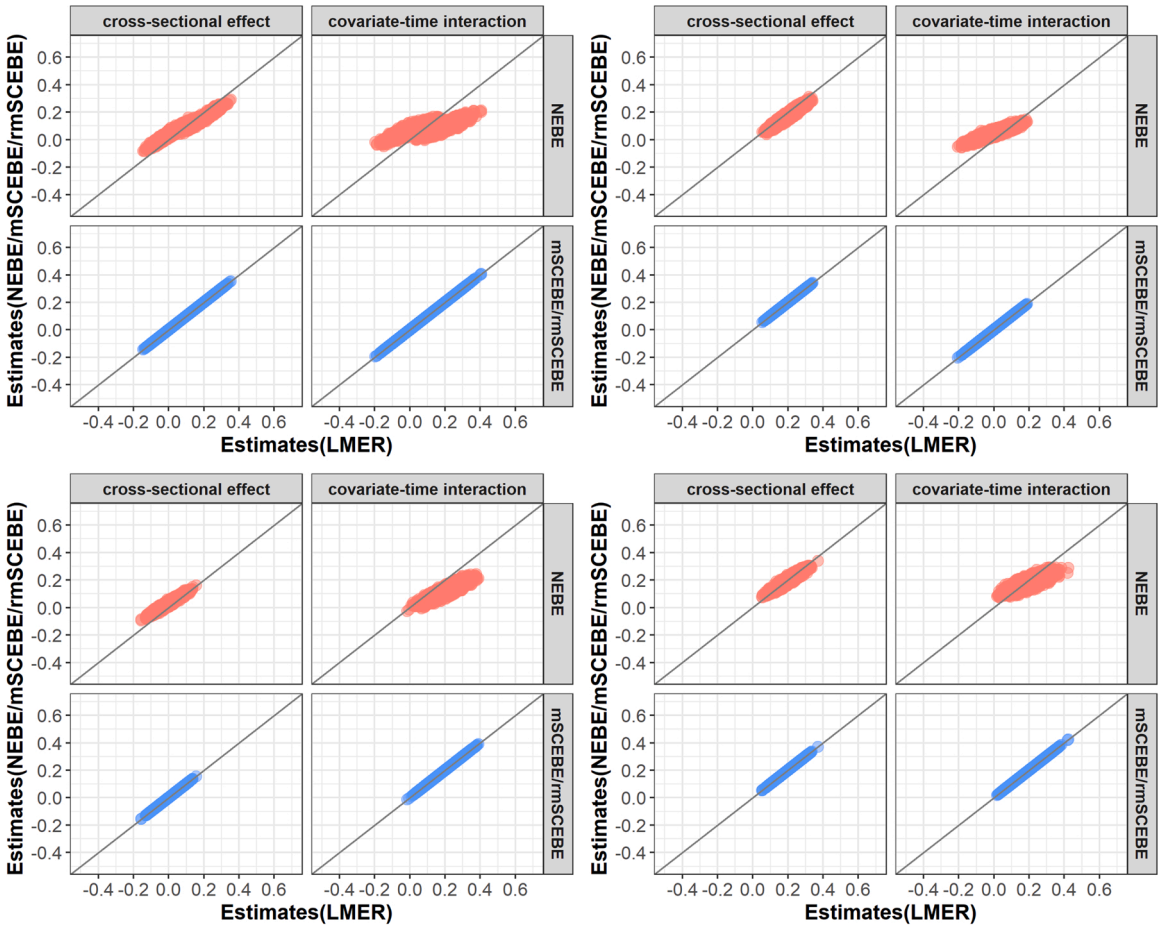

Fig. A1. Estimation comparisons for cross-sectional effects and covariate-time interaction effects based on NEBE, mSCEBE, rmSCEBE and LMER methods. Parameter γ is set according to the four scenarios in Section 3, but the random effects is distributed as multivariate normal with mean 0 and covariance matrix $\begin{pmatrix} 0.09 & 0.045 \\ 0.045 & 0.09 \end{pmatrix}$. Four figures are arranged from left to right and top to bottom in the order of scenario 1–4.

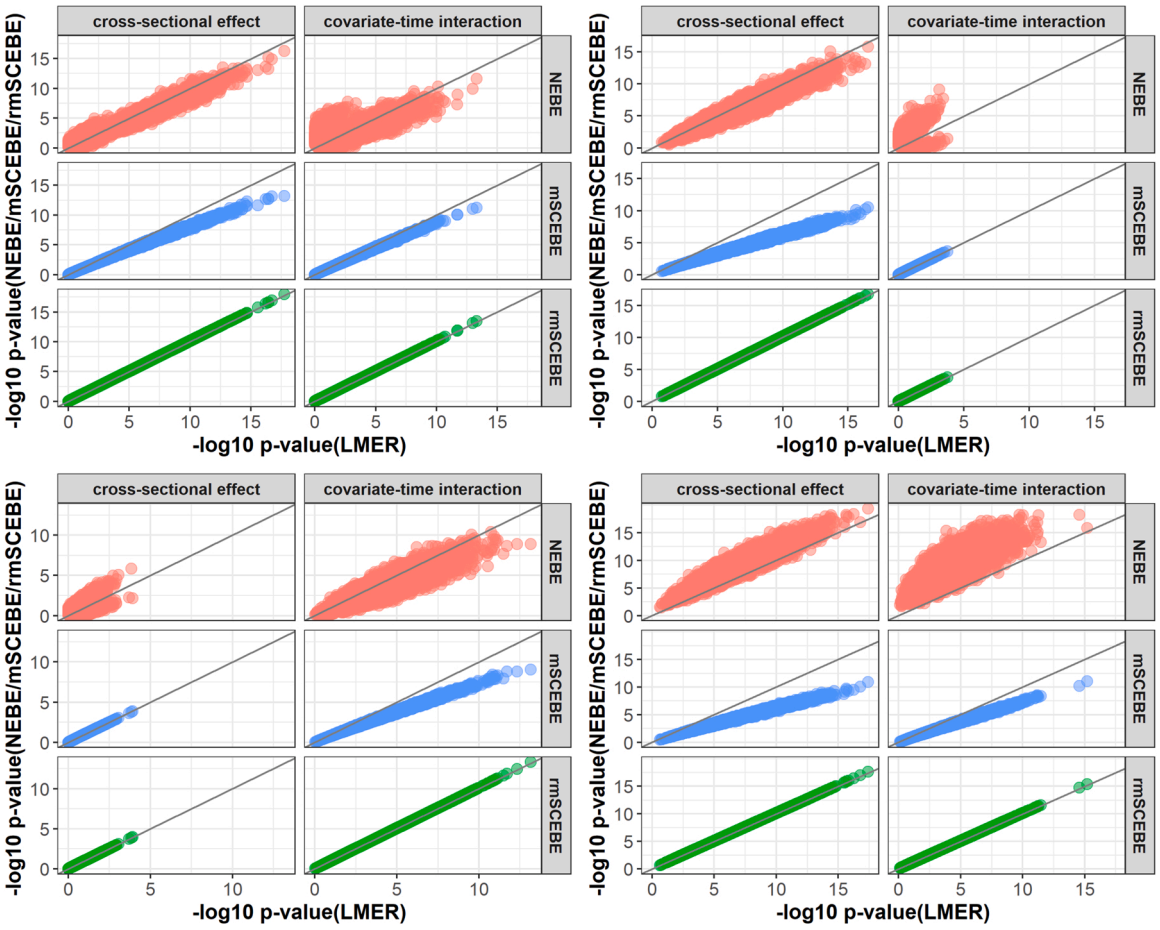

Fig. A2. P-values (on the –log10 scale) comparisons for both cross-sectional effects and covariate-time interaction effects based on NEBE, mSCEBE, rmSCEBE and LMER method in simulation situation 1–4. Parameter γ is set according to the four scenarios in Section 3, but the random effects is distributed as multivariate normal with mean 0 and covariance matrix $\begin{pmatrix} 0.09 & 0.045 \\ 0.045 & 0.09 \end{pmatrix}$. Four figures are arranged from left to right and top to bottom in the order scenario 1–4.

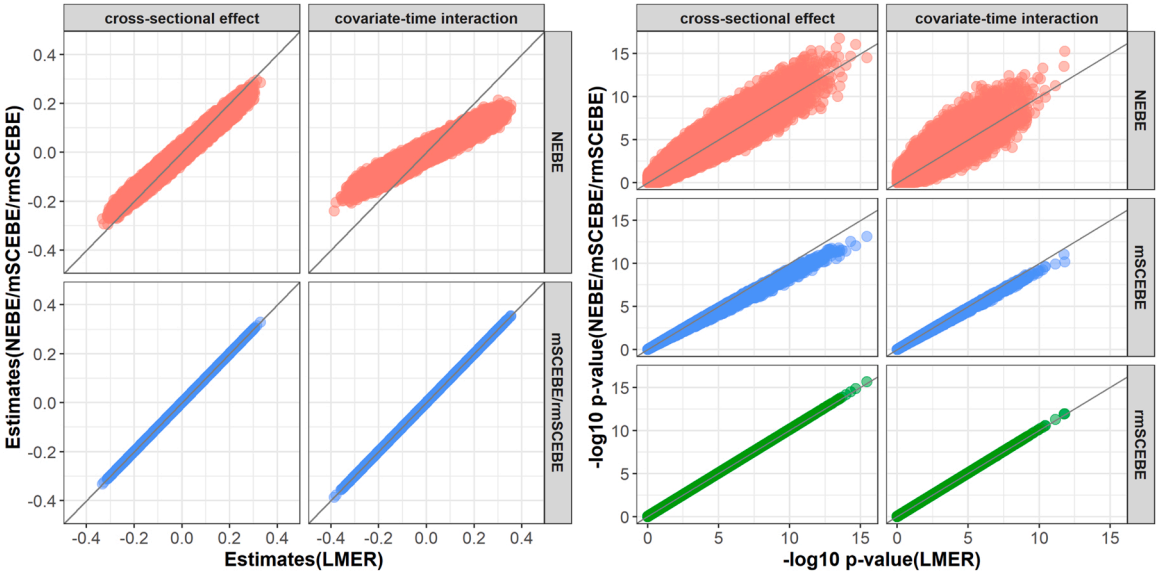

Fig. A3. Results for both cross-sectional effects and covariate-time interaction effects (left: estimation and right: -log10 (p-value)) based on NEBE, mSCEBE, rmSCEBE and LMER method in simulation when gamma is randomly sampled from uniform distribution U(−0.2,0.2).

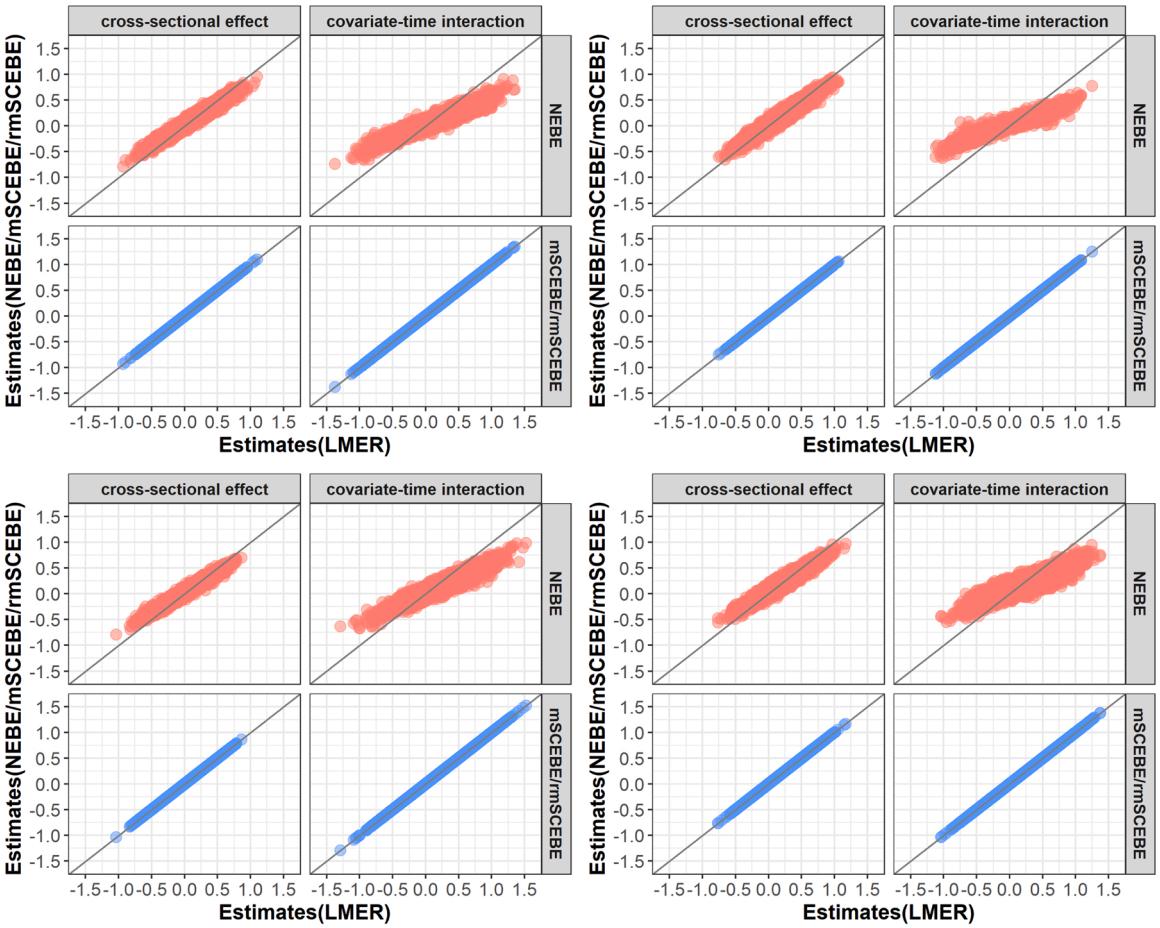

Fig. A4. Estimation comparisons for cross-sectional effects and covariate-time interaction effects based on NEBE, mSCEBE, rmSCEBE and LMER methods when sample size is 100 and covariates are correlated with a coefficient 0.5.

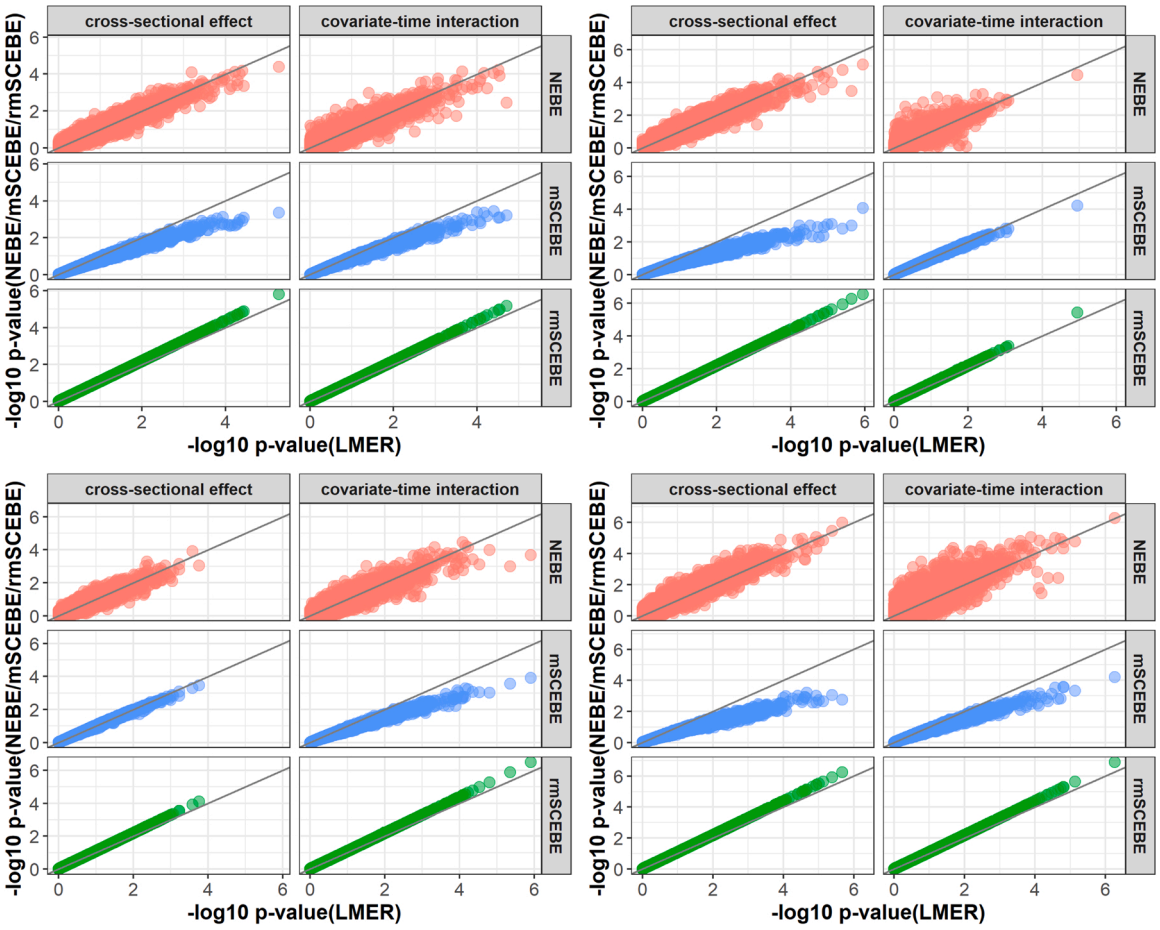

Fig. A5. P-value comparisons for cross-sectional effects and covariate-time interaction effects based on NEBE, mSCEBE, rmSCEBE and LMER methods when sample size is 100 and covariates are correlated with a coefficient 0.5.

References

Barbolosi, D., Ciccolini, J., Lacarelle, B., Barlési, F., André, N., 2016. Computational oncology — mathematical modelling of drug regimens for precision medicine. Nat. Rev. Clin. Oncol. 13 (4), 242–254. https://doi.org/10.1038/nrclinonc.2015.204.

Combes, F.P., Retout, S., Frey, N., Mentré, F., 2014. Powers of the likelihood ratio test and the correlation test using empirical Bayes estimates for various shrinkages in population pharmacokinetics. CPT: Pharmacomet. Syst. Pharmacol. 3 (4), 1–9. https://doi.org/10.1038/psp.2014.5.

Davidian, M., Giltinan, D.M., 2003. Nonlinear models for repeated measurement data: an overview and update. J. Agric. Biol. Environ. Stat. 8 (4), 387–419. https://doi.org/10.1198/1085711032697.

Hand, D.J., Crowder, M.J., 1996. Practical Longitudinal Data Analysis. Chapman and Hall/CRC, London.

Lindbom, L., Ribbing, J., Jonsson, E.N., 2004. Perl-speaks-NONMEM (PsN)–a Perl module for NONMEM related programming. Comput. Methods Prog. Biomed. 75 (2), 85–94. https://doi.org/10.1016/j.cmpb.2003.11.003.

Londono, D., Chen, K., Musolf, A., Wang, R., Shen, T., Brandon, J., Herring, J.A., et al., 2013. A novel method for analyzing genetic association with longitudinal phenotypes. Stat. Appl. Genet. Mol. Biol. 12 (2), 241–261. https://doi.org/10.1515/sagmb-2012-0070.

Maitre, P.O., Buhrer, M., Thomson, D., Stanski, D.R., 1991. A three-step approach combining Bayesian regression and NONMEM population analysis: application to midazolam. J. Pharm. Biopharm. 19 (4), 377–384. https://doi.org/10.1007/BF01061662.

Mandema, J.W., Verotta, D., Sheiner, L.B., 1992. Building population pharmacokinetic–pharmacodynamic models. I. Models for covariate effects. J. Pharm. Biopharm. 20 (5), 511–528. https://doi.org/10.1007/BF01061469.

Marguet, A., Lavielle, M., Cinquemani, E., 2019. Inheritance and variability of kinetic gene expression parameters in microbial cells: modeling and inference from lineage tree data. Bioinformatics 35 (14), 586–595. https://doi.org/10.1093/bioinformatics/btz378.

Meirelles, O., Ding, J., Tanaka, T., Sanna, S., Yang, H., Dudekula, D.B., Cucca, F., et al., 2013. SHAVE: shrinkage estimator measured for multiple visits increases power in GWAS of quantitative traits. Eur. J. Hum. Genet. 21 (6), 673–679. https://doi.org/10.1038/ejhg.2012.215.

Pinheiro, J.C., Bates, D.M., 2000. Mixed-effects Models in S and S-PLUS. Springer, New York.

Saunders, A.M., Strittmatter, W.J., Schmechel, D., George-Hyslop, P.H., St, Pericak-Vance, M.A., Joo, S.H., Rosi, B.L., et al., 1993. Association of apolipoprotein E allele ϵ4 with late-onset familial and sporadic Alzheimer's disease. Neurology 43 (8), 1467. https://doi.org/10.1212/wnl.43.8.1467.

Savic, R.M., Karlsson, M.O., 2009. Importance of shrinkage in empirical bayes estimates for diagnostics: problems and solutions. AAPS J. 11 (3), 558–569. https://doi.org/10.1208/s12248-009-9133-0.

Searle, S.R., Casella, G., McCulloch, C.E., 1992. Variance Components. Wiley, New York.

Sikorska, K., Montazeri, N.M., Uitterlinden, A., Rivadeneira, F., Eilers, P.H., Lesaffre, E., 2015. GWAS with longitudinal phenotypes: performance of approximate procedures. Eur. J. Hum. Genet. 23 (10), 1384–1391. https://doi.org/10.1038/ejhg.2015.1.

Stroup, W.W., Baenziger, P.S., Mulitze, D.K., 1994. Removing spatial variation from wheat yield trials: a comparison of methods. Crop Sci. 34 (1), 62–66. https://doi.org/10.2135/cropsci1994.0011183×003400010011x.

Wu, H., Ding, A.A., 1999. Population HIV-1 dynamics in vivo: applicable models and inferential tools for virological data from AIDS clinical trials. Biometrics 55 (2), 410–418. https://doi.org/10.1111/j.0006-341x.1999.00410.x.

Xu, H., Li, X., Yang, Y., Li, Y., Pinheiro, J., Sasser, K., Hamadeh, H., Xu, S., Yuan, M., for the Alzheimer's Disease Neuroimaging Initiative, 2020. High-throughput and efficient multilocus genome-wide association study on longitudinal outcomes. Bioinformatics 36 (10), 3004–3010. https://doi.org/10.1093/bioinformatics/btaa120.

Yuan, M., Li, Y., Yang, Y., Xu, J., Tao, F., Zhao, L., Zhou, H., Pinheiro, J., Xu, X.S., 2020. A novel quantification of information for longitudinal data analyzed by mixed-effects modeling. Pharm. Stat. 19 (4), 388–398. https://doi.org/10.1002/pst.1996.

Yuan, M., Xu, X.S., Yang, Y., Xu, J., Huang, X., Tao, F., Zhao, L., Zhang, L., Pinheiro, J., 2019. A quick and accurate method for the estimation of covariate effects based on empirical Bayes estimates in mixed-effects modeling: correction of bias due to shrinkage. Stat. Methods Med. Res. 28 (12), 3568–3578. https://doi.org/10.1177/0962280218812595.

Yuan, M., Xu, X.S., Yang, Y., Zhou, Y., Li, Y., Xu, J., Pinheiro, J., 2021. SCEBE: an efficient and scalable algorithm for genome-wide association studies on longitudinal outcomes with mixed-effects modeling. Brief. Bioinform. 22 (3), bbaa130. https://doi.org/10.1093/bib/bbaa130.